

CULTURAL POLITICS OF USER-GENERATED ENCYCLOPAEDIAS:
Comparing Chinese Wikipedia and Baidu Baike

A D.Phil. thesis

SUBMITTED TO THE EXAM SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

D.Phil

Program in Information, Communication, & Social Sciences

by

Han-Teng Liao

Keble College

University of Oxford

Trinity Term 2014

Abstract

Cultural politics of user-generated encyclopaedias:

Comparing Chinese Wikipedia and Baidu Baike

Han-Teng Liao, Keble College, Trinity Term 2014

The question of how the Internet affects existing geo-cultural or geo-linguistic communities in relation to nation-states has continued to receive attention among academics and policymakers alike. Language-based technologies and services that aggregate, index, and distribute materials online may reshape pre-existing boundaries of the relationship between users and content, for instance with different language versions of user-generated encyclopaedias or different local versions of search engines. By comparing two major Chinese online encyclopaedias, Baidu Baike and Chinese Wikipedia, this thesis investigates whether the Internet overcomes, shifts, or reinforces boundaries among Chinese language users. The Chinese language provides an excellent case for examining the boundary question. While the Internet can potentially connect the largest number of native speakers around the world, the majority (i.e. those from mainland China) face an Internet censorship and filtering regime that may limit this very potential. Modern Chinese history has also complicated the cultural-political boundaries among the regions of mainland China, Hong Kong, and Taiwan.

This thesis compares the conditions and outcomes of their respective editorial processes, content features, and users' reception. Multiple findings emerge from a combination of quantitative and qualitative methods, including content analysis, webometrics, and search engine result visibility tests. These methods show that boundaries are drawn in the process of creating, linking, and searching content on the Chinese Internet. Their geolinguistic extent differs, a phenomenon that reflects the cultural-political division between mainland China and the rest of Chinese-speaking world. Both the findings and methods of the thesis have important implications for research and policy for understanding the globalizing regionalization and nationalization effects of the Internet.

Keywords: online encyclopaedias; social media; user-generated content; China

Acknowledgements

This study of Chinese-language Internet would not have been possible without the financial, institutional, social and personal support that I have received. It is through this interrelated life support, I had some freedom to conduct a D.Phil. research project that is independent.

I would never have been able to begin this endeavour without the financial support by two institutions for the initial two years. First, I acknowledge the National Science Council (now Ministry of Science and Technology) of Taiwan, who found my proposal worthy of the support by Taiwan Merit Scholarships Program (NSC-095-SAF-I-564-028-TMS). Second, I thank Oxford Internet Institute (OII) for the OII and PGP Scholarships.

In the process of conducting and presenting the research, I have also received support from Academia Sinica (Institute of Sociology), City University of Hong Kong (Department of Chinese, Translation and Linguistics), Keble Association, Wikimedia UK, etc. For encouragement and aid in diverse ways beyond my D.Phil. research project, I also thank the committee of Chinese Internet Research Conference and Georgetown University (Institute for the Study of Diplomacy) for the opportunity to work as the Yahoo! Fellow and organise a conference, which in turn informed and shaped the research project. I express my gratitude to a great number of individuals involved who have supported me.

Over the past years, I have received encouragement and guidance from Dr. Ralph Schroeder, who has been a mentor and friend along this rewarding journey. I would like to thank the examiners Dr. William H. Dutton and Dr. Mike Thelwall for their valuable questions and comments. In addition, Dr. Bernie Hogan, Dr. Eric T. Meyer, and other researchers and colleagues at the OII have provided valuable methodological advice. I am also indebted to Dr. Jonathan Bright and Mr. Scott A. Hale for reading and listening for the mock viva.

I would like to thank the users who contributed to Baidu Baike and Chinese Wikipedia for sharing their time and ideas in compiling a knowledge resource for all Chinese-language users. I would also like to thank the users who are willing to share their experience with the two encyclopedia websites, by either talking to me face-to-face or voicing their thoughts publicly online.

Finally, I would like to express my deepest gratitude to my family, especially my parents, who originally had great expectations out of their scholarship boy before my D.Phil. project, though probably not in the domains of academia. Nevertheless, their unconditional support and love have been the safe harbour for me during this laborious journey. Thanks also to my one and only sister for her encouragement and honest reality checks.

Table of Contents

Table of Contents	i
List of Tables	iii
List of Figures	vi
Chapter 1 Introduction	1
1.1 Origins of the study	1
1.2 Background and research question.....	7
1.3 Boundaries: offline and online	15
1.4 Overview of subsequent chapters	24
Chapter 2 Literature review	28
2.1 National boundaries and communicative spaces.....	29
2.2 Chinese national boundaries and media technologies	38
2.3 Boundaries of media-language systems.....	51
2.4 Summary and implications.....	63
Chapter 3 Theoretical framework and methods used	65
3.1 Three main concepts	67
3.2 Mixed methods: geolinguistic analysis of cultural patterns	95
3.3 Overall expected outcomes.....	105
Chapter 4 Editorial processes, Internet control, and Internet diffusion	109
4.1 Editorial policies and processes: filtering user contribution.....	112
4.2 Gatekeeping mainland Chinese users from 2005–2008	139
4.3 Chapter conclusions.....	168
Chapter 5 Citation and content analysis	174
5.1 Geolinguistic patterns and preferences.....	177
5.2 Size and institutional considerations.....	193
5.3 Defining and negotiating Chineseness	210
5.4 Chapter conclusions.....	221
Chapter 6 Reception and use	229

6.1 Search engine result pages (SERPs)	232
6.2 Microblog posts.....	261
6.3 Chapter conclusions.....	293
Chapter 7 Conclusion	300
7.1 Understanding the Chinese-language Internet	300
7.2 Implications for research and policy	310
7.3 Understanding the geolinguistic dynamics of knowledge and information	316
References	324

List of Tables

Table 1-1	<i>Chinese-language support by Google Search, Google News, Wikipedia, Facebook and Microsoft Windows as of 2012</i>	16
Table 1-2	<i>Four Major Chinese-speaking Regions and Their Records in Political Internet filtering, Free Speech, Democracy and Human Rights.....</i>	18
Table 1-3	<i>Visible difference (double-underlined and in the colour red) of the same text written in simplified and traditional Chinese script.....</i>	20
Table 1-4	<i>The existing cultural and political boundaries among major Chinese-speaking regions, using mainland China as a reference point</i>	23
Table 2-1	<i>Chinese clusters: previous research.....</i>	49
Table 2-2	<i>Types of texts and technologies: a comparison</i>	55
Table 2-3	<i>Examples of the IETF language tags for Chinese</i>	59
Table 2-4	<i>Examples of major writing systems being digitized.....</i>	60
Table 3-1	<i>Different specifications of web spheres.....</i>	70
Table 3-2	<i>Search engine variants: Belgium</i>	72
Table 3-3	<i>Different keyword suggestions for the Chinese query of “Communist Party”</i>	81
Table 3-4	<i>Possible cultural thickening patterns.....</i>	106
Table 3-5	<i>Factors influencing cultural-political boundaries.....</i>	107
Table 4-1	<i>Basic information about Baidu Baike and Chinese Wikipedia</i>	111
Table 4-2	<i>Geographic distribution of the power users.....</i>	126
Table 4-3	<i>Wikipedia in mainland China: blocked periods before 2008.....</i>	140
Table 4-4	<i>Categorization of different levels of diffusion based on the 2010 data.....</i>	155
Table 4-5	<i>Comparing editorial development and patterns.....</i>	169
Table 5-1	<i>Examples of character encoding standards (character sets).....</i>	182
Table 5-2	<i>Measuring the deviance from simplified Chinese.....</i>	185
Table 5-3	<i>Numbers of collected article pages, external web links and pages</i>	187
Table 5-4	<i>Exclusion criteria for the external web links</i>	187

Table 5-5 <i>Skewed geographic distribution of links</i>	188
Table 5-6 <i>Comparing the geographic distribution of links</i>	189
Table 5-7 <i>Comparing distribution of links across selected regions</i>	196
Table 5-8 <i>Top-linked China-based websites</i>	198
Table 5-9 <i>Top-linked U.S.-hosted websites</i>	200
Table 5-10 <i>Top-linked Hong Kong-hosted websites</i>	203
Table 5-11 <i>Top-linked Taiwan-hosted websites</i>	205
Table 5-12 <i>Top-linked Macau-hosted websites</i>	206
Table 5-13 <i>Comparing distribution of citing and content patterns</i>	223
Table 6-1 <i>Search engine market shares: top 5 according to StatCounter</i>	237
Table 6-2 <i>Sources and numbers of search queries</i>	243
Table 6-3 <i>Visibility scores: the top five websites and their top five domains</i>	246
Table 6-4 <i>Percentage of visibility scores: encyclopedia sites among the top-10</i> ...	248
Table 6-5 <i>Ranking of visibility scores: encyclopedia sites among the top-10</i>	249
Table 6-6 <i>Clusters identified by blockmodeling</i>	252
Table 6-7 <i>Blockmodeling result matrix</i>	253
Table 6-8 <i>Wikipedia-related and Wiki-similar Chinese terms</i>	266
Table 6-9 <i>Posts collected: including the expected false positives</i>	268
Table 6-10 <i>Posts collected: false positives removed</i>	268
Table 6-11 <i>Posts that mention wenyunchao (the set "WYC")</i>	269
Table 6-12 <i>Most frequently-occurring words</i>	271
Table 6-13 <i>Most frequently-occurring words when mentioning</i>	273
Table 6-14 <i>Most frequently-appeared words when mentioning Baidu Baide or Chinese Wikipedia (one major incident removed)</i>	274
Table 6-15 <i>Selected posts that mentioned Chinese Wikipedia in relation to censorship/filtering</i>	276
Table 6-16 <i>Selected posts that mentioned both encyclopaedias (in relation to censorship/filtering)</i>	280

Table 6-17 <i>Selected posts that mentioned both (apart from those concerning censorship/filtering).....</i>	282
Table 6-18 <i>Baidu/Wiki a bit.....</i>	286
Table 6-19 <i>Comparing use and reception patterns.....</i>	295
Table 7-1 <i>Different gatekeeping lead to different cultural thickening patterns...</i>	305
Table 7-2 <i>Views on the impact of the Great Firewall (GFW) of China.....</i>	309

List of Figures

<i>Figure 1-1.</i> Screenshots of Typical Google's Search Engine Results Page (SERP) and Wikipedia's Entry Page. Screenshots reproduced with permission of the website owners Google (2012a) and Wikimedia.	9
<i>Figure 1-2.</i> The diagram of a typical Search Engine Results Page (SERP) and a User-Generated Encyclopaedia Entry Page (UGEPP).	10
<i>Figure 2-1</i> Growth of Unicode on the Web (M. Davis, 2012).....	58
<i>Figure 3-1.</i> Processability shapes web spheres	66
<i>Figure 3-2.</i> Cultural thickening patterns that overcome or reinforce boundaries..	66
<i>Figure 3-3.</i> Map showing Belgium's language areas (User:Stevenfruitsmaak, 2006)	71
<i>Figure 4-1.</i> Number of entries and external links, and their normalized frequency	119
<i>Figure 4-2.</i> Baidu Baike team help: Rules outlined for Baike Kedou members	122
<i>Figure 4-3.</i> Distribution of power users of Baidu Baike and Chinese Wikipedia in 2012.....	127
<i>Figure 4-4.</i> Distribution of power users of Baidu Baike(BB) and Chinese Wikipedia (CW) in 2012: East and South-East Asia	128
<i>Figure 4-5.</i> Screenshot of the entry Taxi without the popup box.	133
<i>Figure 4-6.</i> Timeline of major events for Baidu Baike and Chinese Wikipedia ...	141
<i>Figure 4-7.</i> Taneja & Wu's "Internet Access Blockage and User Behavior"	143
<i>Figure 4-8.</i> The bell-shape curve and the S-shape curve of innovation diffusions	150
<i>Figure 4-9.</i> Internet diffusion rates for 17 East Asian and 31 Chinese regions	153
<i>Figure 4-10.</i> Chinese regions and East Asian regions: Category I average	156
<i>Figure 4-11.</i> Chinese regions and East Asian regions: Category II average.....	156
<i>Figure 4-12.</i> Chinese regions and East Asian ones: category III average	157
<i>Figure 4-13.</i> Comparison of Chinese regions with East Asian regions: Category I	158
<i>Figure 4-14.</i> Comparison of Chinese regions with East Asian regions: Category II	158
<i>Figure 5-1.</i> Language scripts detected: all language scripts	190
<i>Figure 5-2.</i> Language scripts detected: unpacking the Unicode-encoded results ..	191

<i>Figure 5-3.</i> A treemap comparison of citations and external sources for “Hanzi”, “Hanzu” and “Tianxia”	212
<i>Figure 5-4.</i> A treemap comparison: “Hanzi”	215
<i>Figure 5-5.</i> A treemap comparison: “Hanzu”	217
<i>Figure 6-1.</i> Higher SERP rankings x produce higher click-through rates and thus higher visibility scores	236
<i>Figure 6-2.</i> Search engine market shares: from 2009 to 2014	238
<i>Figure 6-3.</i> A 2-mode network before and after blockmodelling.	240
<i>Figure 6-4.</i> A blockmodelling example showing unfit data points.....	242
<i>Figure 6-5.</i> Concentrated visibility scores	245
<i>Figure 6-6.</i> Blockmodelling result network visualization	254
<i>Figure 6-7.</i> Contrasting Baidu Baike’s and Chinese Wikipedia’s visibility	256
<i>Figure 6-8.</i> Contrast of top-level domain names (TLD).....	257
<i>Figure 6-9.</i> Contrast of geoIP locations	257
<i>Figure 6-10.</i> Contrast of scripts among the largely Chinese-language links	258
<i>Figure 6-11.</i> Venn diagram of the mentions of the three encyclopaedias	278

Chapter 1 Introduction

1.1 Origins of the study

During an Oxford Union Debate, Jimmy Wales, the co-founder of Wikipedia cited a Taiwanese Wikipedian as evidence for proving that “the Internet is the greatest force for Democratisation in the World” (OII, 2007):

What's gonna bring down the totalitarian government in China? Is it gonna be guns? Is it gonna be tanks? Is it gonna be missiles? No it's gonna be KJ. KJ is a 25-year-old girl who lives in Taiwan, ... She is talking to the mainland Wikipedians, and she's helping them to get through the firewalls. Every day with something she is doing, and every day she is bringing in a bit of democracy and information of freedom to China. ... This is going on all the time in hundreds and thousands of ways all over the world. All of these censorship regimes are porous. They are becoming more porous and they will be more porous [sic].

As a Taiwanese who experienced the early phases of democratization during my high school years and Internet adoption during my undergraduate years in Taiwan, I appreciate democratic values and the ways in which the Internet may reinforce these values.

While Jimmy Wales was using language in the debate for rhetorical effect, the actual impact of the Internet in democratizing a totalitarian government is likely to be very limited. Despite his rhetorical optimism, the possibilities are almost non-existent for a group of Wikipedians to bring down a totalitarian government somehow. Vivid as the image of collaboration that overcomes existing boundaries is, the Internet may also have other effects, in the opposite direction. When Jimmy Wales was optimistic that a Taiwanese Wikipedian had been “bringing a bit of democracy and information of freedom” from Taiwan (which has one of the freest media environments in Asia) to China (which has one of the world’s most restrictive ones), it is also possible that non-democratic influences and censorship can be brought to Taiwan from mainland China. After all, Wikipedians, or users who contribute to the largest online

encyclopaedia that “everyone can edit”, may not all agree on which government can be described as totalitarian, let alone convince other Wikipedians to seek regime change of a state. A pessimistic scenario is also likely if the impact of censorship regime is porous enough for a totalitarian state to influence a democratic state.

For example, in Taiwan, a series of acquisitions of several media companies including a newspaper, a TV station and a cable network by a successful Taiwanese billionaire has raised concerns regarding his pro-Beijing line and possible use of censorship (Higgins, 2012b). Therefore, more evidence is needed to discern whether an optimistic or pessimistic view is warranted. In other words, researchers must not take the Internet’s potential for democracy and freedom as a given just because of its openness and connectivity. We need to examine what is connected and for what purposes. Openness and connectivity do not automatically guarantee that existing boundaries will become porous and thus be overcome. They also do not guarantee the results will always favour freedom and democracy. For instance, between China and Taiwan, the features of the Internet (openness and connectivity), similar to their increasing economic and social ties after the Cold War, do not necessarily guarantee that both will have freer and more democratic societies.

During the Cold War, Mainland China, Hong Kong and Taiwan were ruled respectively and separately by socialist, colonial, and authoritarian states, each of which had different “nationalizing” or “de-nationalizing” projects to delineate the content and boundaries of their citizenships (So, 2004). The boundaries among them experienced a number of further developments after the Cold War, including China’s “opening-up” in the 1980s, Hong Kong’s return to China in 1997 after a transitional period of decolonization and partial democratization from 1984, (So, 2004) and Taiwan’s democratization in the 1990s (B. He, 2003). Arguing that Taiwan is part of China, Beijing’s “One China” policy considered any self-determination process unacceptable. In such context,

whether Internet can facilitate democratization of these Chinese governments remains to be seen.

This thesis begins, therefore, with a difficult yet more fundamental question, both from the literature and my own experience. How have Internet technologies overcome, shifted or reinforced the existing boundaries of communication patterns? This question can be asked in different ways. For example, some researchers raise the question of whether the Internet has worsened political and social divisions by increasing social fragmentation and group polarization (Sunstein, 2007; Pariser, 2011), which can be translated into newly created boundaries (i.e. fragmentation) or reinforcement of old boundaries (i.e. polarization). For another, some researchers have continued the discussion regarding how the Internet reshapes the dynamics of globalization and national sovereignty (Fidler, 1998; Sassen, 1998). Such discussion is directly related to the question of how the Internet changes boundaries of nation-states or jurisdictions.

These issues are relevant to me as a Taiwanese who can read Chinese-written materials from mainland China and Hong Kong. How has the Internet changed the group dynamics of Chinese-speaking Internet users? How has the Internet changed the existing boundaries between Chinese-speaking regions, especially those among mainland China, Hong Kong and Taiwan, which are culturally and politically ambiguous and complicated? With some basic training in computer science, information science, media studies, social science and linguistics, I have come to the conclusion that this fundamental question regarding boundaries of communication patterns cannot be understood without examining both its linguistic and technological aspects.

What is implicit in Jimmy Wales' statement is the mediating role of language across political boundaries: A 25-year-old Taiwanese girl is talking to the mainland Wikipedians *in Chinese language* to edit and improve the *Chinese-language version* of Wikipedia. As an instance of exploiting the technical

features of an open and interconnected Internet, Wikipedia must digitize, aggregate, exchange and prioritize linguistic materials among volunteer editors. What is also implicit in Wales' statement is the possibility of spreading the idea of information freedom and democracy in the process, including overcoming the techno-political barriers of Internet censorship and filtering. Wales' statement was too optimistic, as will be shown in this thesis. To gain a sense of the central mediating role of language, it will be useful to provide a brief historical and personal account of the boundaries of communication patterns within the Chinese-speaking world.

During the Cold War, the clear boundary between Taiwan and mainland China was an outcome of the Cold War: "Free China" on this side of the Taiwan Strait, and Communist China on the other side. No travel, trade or communication had been possible until China's era of reform and "opening-up" of the 1980s. When the Berlin Wall came down in 1989, the voices and images of the Tiananmen Square protestors (tiān'ānmén shìwēi zhě 天安门示威者) in mainland China, which were transmitted via international broadcasting, moved me deeply, a feeling that was shared among many in Taiwan and Hong Kong. I related to the desire for change because I had undergone the political education of a waning authoritarian regime in Taiwan and witnessed the ensuing political protests and social movements on the streets of Taiwan. Then came the Internet. I was introduced to the power of the Internet as a university student of electrical and electronics engineering, when the digital and computer environment was not yet ready to process the Chinese language. The development of the Internet coincided with the period when mainland China was not only opening up to the world after the Cold War, but also reconnecting Chinese-speaking regions such as Hong Kong and Taiwan after the Cold War.

In order to make sense of the power of sharing unleashed by Internet technologies and nascent civic societies in Taiwan, I began to read a body of research on free software and open content development that examined the

potential and potential pitfalls of open collaboration on the Internet. I even took the initiative to engage scholars from mainland China about the implications of open participation for creative and content industries. In the process, the question of what the open Internet means or can do for Chinese-speaking regions such as mainland China, Hong Kong and Taiwan took shape. More generally, I became curious about what the open Internet means for users across different regions, particularly those having different political and media environments. This was the research question that motivated my D.Phil. research project.

Using this thesis to examine the specific case of the Chinese-language Internet, I want to present a number of arguments about the effects of Internet connectivity and openness on political and communication boundaries. First, “the web is increasingly grounded with geographical and linguistic specificity by platform and space” (R. Rogers, 2013, p. 58). Taking this idea further, I argue that language is the first and primary boundary factor, whereas the geography factor may facilitate or hinder interaction within linguistically demarcated Webs. The information and communication spaces online are thus likely linguistically demarcated Webs, wherein users access information and interact among themselves. In other words, actual implementation of Internet connectivity for cultural-political purposes still requires a common language that enjoys support by information technologies and actual usage by its speakers. For example, a Chinese-language information or communication space needs to process Chinese-language texts and/or inputs, and its users must have the skills and literacy required to participate.

For the purpose of the thesis, I shall define the large-scale human- and computer-processing of online materials as the latest of embodiment of media-language “processability”, by which openly available information is processed, filtered and ordered for different groups of users, which are often geolinguistically defined. While it has some universal efficacy in aggregating,

categorizing, and prioritizing information, information has been, in practice, and will continue to be, configured, appropriated, and practised in various ways according to the cultural-political context of the users and websites involved. That is to say, all large-scale online interaction depends on the power of processing information, but its cultural and political ramifications depend on how the linguistic, geographical and platform specificities are built into system design and the social processes associated with it. Therefore, researchers can study the way processability is constituted and executed with cultural-political consequences. Sharing the same language, like sharing the Internet, does not automatically guarantee the overcoming of boundaries, not to mention the triumph of information freedom or democracy. Researchers must test and examine where and how boundaries are reconstituted or reintroduced, including showing where sharing the same language *does* and *does not connect* across *geographical* or *platform* differences - or not.

The above arguments thus engage several debates in research on the general topic of connectivity and openness. One of them concerns the relationship between human values (connectivity and openness) and Internet technologies. These debates have taken place not only in relation to China, but also, for example, with respect to what the Internet's connectivity and openness has meant in relation to the Arab Spring starting in late 2010. In this respect, I maintain that we must examine how websites such as major search engines and user-generated encyclopaedias aggregate, categorize, and prioritize information in such a way that has multiple cultural-political implications. For instance, both democratic values and authoritarian control can be embedded into such information-processing powers. Instead of attributing the potential of democratisation or control to just the power of people or to mere technologies, this thesis argues that linguistic "processability", a mixture of human- and machine-processing of open and interconnected linguistic materials, is critical. A key factor is what or who is connected, determining which patterns of

information and communicative effects result in certain content becoming prominent for certain group of users. The Chinese language web, the Arabic language web, etc., can thus be better understood as the basic grounds for cultural-political connectivity.

Another debate, mostly situated in Chinese Internet Research or broadly in modern Chinese Studies, concerns the cultural-political boundaries among mainland China, Hong Kong and Taiwan, particularly in the context of Chinese and US geopolitical interest in the region. The data presented in this thesis indicates that the difference between Baidu Baike and Chinese Wikipedia reflects the intricate dynamics of Chinese-language Internet users and information, and less so the geopolitical or political struggles between China and the US assumed by many observers and researchers.

After describing the background of this study, I will first introduce both user-generated encyclopaedias and search engines, arguably two of the most important website types for many users in the world since 2004. The next section will discuss how they achieve information connectivity in an open environment, thereby formulating a general research question. In the third section of this chapter, I will develop a set of questions by grounding the general research question in the specific setting of the Chinese-language Internet, with an emphasis on how online encyclopaedias and search engines shape and are shaped by Chinese (not just China's) cultural politics. Then I will briefly describe how the thesis can provide an account of how the open participatory Web operates, one which can inform and help researchers and policy makers to better shape the design of online content practices and policies, especially for cross-regional, trans-national and even cross-lingual communication. I will conclude this chapter with an overview of the chapters to come.

1.2 Background and research question

1.2.1 User-generated encyclopaedias and search engines. Wikipedia and the Google search engine are central to many Internet users across the world when it

comes to information seeking. In some parts of the world, other successful alternatives exist and even dominate in certain local markets, with the salient examples of Baidu Baike and the Baidu search engines in mainland China (Shim & Yang, 2009). In the following paragraphs, I will highlight the ways in which user-generated encyclopaedias and search engines embody the connectivity of an open information environment. What follows is a basic introduction of the two types of websites, followed by a brief summary of their significance in everyday online usage and in terms of Internet connectivity and globalization, so that the research question of the thesis can be formulated and presented.

Users encounter webpages of search engines and user-written encyclopaedias, as shown in Figure 1-1 (as screenshots) and Figure 1-2 (as a diagram). On the left, a typical search engine result page (SERP) is composed of the following elements (from top to bottom): search query, a list of search results (usually around ten items per page), and links to further results. Each result item contains a title, a link and a short summary. On the right, in addition to an entry title, a typical user-generated encyclopaedia's entry page (UGEEP) has sections of introductory text, body text, "see also", and references. The diagrams in Figure 1-2 indicate how both resemble each other: A SERP's search query, like a UGEEP's title, indicates a topic. A SERP leads to other additional SERPs; a UGEEP leads to other related UGEEPs (in the 'See Also' section). A result item in the SERPs links to information hosted elsewhere; an item listed in the References section points readers to different external information sources online and offline.

The figure consists of two side-by-side screenshots. The left screenshot is a Google search results page for the term 'Tianxia'. It shows the Google logo at the top, followed by search filters (Web, Images, Shopping, Videos, Maps, More) and a search bar. Below the search bar, it indicates 'About 2,350,000 results (0.30 seconds)'. The first result is from Wikipedia, titled 'Tianxia - Wikipedia, the free encyclopedia', which defines it as a phrase in Chinese language and an ancient Chinese cultural concept. Other results include a Kickstarter campaign for 'Tianxia: Blood, Silk & Jade by Vigilance Press', a blog post 'Rethinking Tianxia | China Heritage Quarterly', and a TV series 'Tian xia (TV Series 1988-1990) - IMDb'. The right screenshot is the Wikipedia entry for 'Tianxia'. It features a title 'Tianxia' with a subtitle 'From Wikipedia, the free encyclopedia'. The main text explains that 'Tianxia' (literally 'under Heaven') is a phrase in Chinese language and an ancient Chinese cultural concept. It also includes a 'Contents' table of contents, a 'Historical and Political Development' section with a map, a 'Sinoxenic Uses of Tianxia' section, and a 'References' section. The entry is well-structured with various sub-sections and a detailed introduction.

SERP

UGEEP

Figure 1-1. Screenshots of Typical Google's Search Engine Results Page (SERP) and Wikipedia's Entry Page. Screenshots reproduced with permission of the website owners Google (2012a) and Wikimedia.

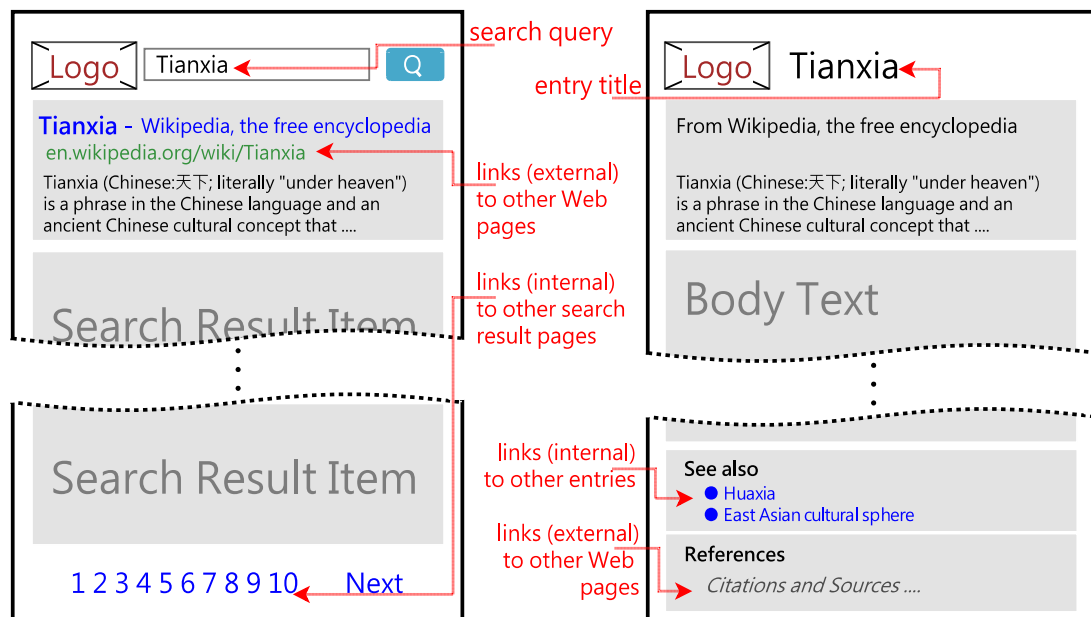
*SERP**UGEPP*

Figure 1-2. The diagram of a typical Search Engine Results Page (SERP) and a User-Generated Encyclopaedia Entry Page (UGEPP)

The SERPs and UGEPPs thus populate the Web with selections of links and information for almost any topics that can be represented by search queries or entry articles. Effectively, these patterns of links and keywords constitute the mechanisms by which information and meaning can be exchanged. The interdependence between links and keywords has implications for our understanding of one important aspect on how the Internet connects human activities: These activities are likely to be bounded by language boundaries because of the keywords used and the expected additional links occurring within rather than across respective languages. Thus, by providing relevant content and links, especially those hosted elsewhere, about certain topics, the SERPs and UGEPPs both provide a snapshot of materials deemed relevant by the respective websites. In addition, because their content and links may change according to the changing dynamics of the Web, both SERPs and UGEPPs essentially summarize the open Web materials by means of a specific phrase, allowing users to further explore and navigate the Web. Thus, the Internet's features of

openness and connectivity on the Web have been embedded and realized in the actual implementations of SERPs and UGEEPs.

Although it is commonly assumed that SERPs are mostly generated by computer algorithms and UGEEPs are mostly contributed by online users, the difference between human- versus computer-manipulation of the open linguistic materials on the Web is sometimes blurred. For SERPs, it has been reported that both conscious collective actions (Bar-Ilan, 2007a, 2007b; Bar-Ilan, 2006; Grimmelmann, 2009; McNichol, 2004; Tatum, 2005) and aggregated individual search trails (Bilenko & White, 2008; Pariser, 2011; Spink, Jansen, Wolfram, & Saracevic, 2002) are important. For UGEEPs, there has been research on the role of automatic computer programs in maintaining Wikipedia (Geiger, 2010; Geiger & Ribes, 2010; Halfaker & Riedl, 2012) and it has been shown that encyclopaedia articles can be automatically generated from Web search results (Sauper & Barzilay, 2009). Therefore, it is essential to examine both the human and computer side of SERPs and UGEEPs and how they shape the interconnectivity of the Web through linking and the sharing of linguistic materials.

1.2.2 Internet and globalization. If Wikipedia and Google Search (or other SERPs and UGEEPs providers) are central and pivotal to Internet users' everyday information-seeking experience (as will be documented further in the chapters to come), what are the implications for understanding the relationship between the Internet and globalization? I argue that connectivity is primarily cultural (as opposed to physical or technical) and based on the shared language of keywords/search queries and web links, thereby producing connections that are likely to be conditioned by largely linguistic and sometimes geographic factors. What follows is a brief elaboration of this point, using the Chinese language as an illustrative case, so as to lay out the research questions that follow.

Imagine a Chinese-speaking user from mainland China, Hong Kong or Taiwan who looks for some general information on a topic, say, the Second Sino-Japanese War, where the Republic of China fought the Empire of Japan mostly

in mainland China from 1937 to 1945. From the results of SERPs and UGEEPs, what will s/he see? Are the results for a user from mainland China the same, similar or completely different from those for another user from Hong Kong? What "editorial decisions" are behind the results of SERPs and UGEEPs, and to what effect? It should be noted here that editorial decisions are not just made by people, but also, as Cory Doctorow(2009, l. 1) puts it, "[s]earch algorithms are editorial decisions". This thesis will consider the language and region settings as part of the "editorial decisions" made by human- and computer- manipulation of linguistic materials. However, what kinds of materials are included in or excluded from these SERPs and UGEEPs through human and/or computer efforts? What are the cultural and political implications of the differences or similarities observed in the search and content results?

All the above questions both derive and depart from existing discussions on the relationship of the Internet and globalization. One of the central themes of this discussion is the impact of cross-border Chinese-language interactions on the Web. Recall the main premise of the thesis: The actual implementation of Internet connectivity requires language-based practices and technologies. The purpose of the thesis is to examine how culturally meaningful connections shape the boundaries among Chinese-speaking regions. How is a sense of online linguistic community, or the sense of fellow-users-of-the-same-language, supported and substantiated by such connectivity? In the case of the Chinese-speaking world, the old boundaries between print and broadcasting media across mainland China, Hong Kong and Taiwan, which were rigid and almost absolute during the Cold War, constitute fewer barriers for exchanges in an Internet environment that is open and interconnected. As will be discussed in Chapter 2, we need new theories and methods to observe how existing boundaries may be overcome or reintroduced into the Web. The research on how the feelings towards - and practices - of fellow-users-of-the-same-language on the Web can add to the on-going research on media and national or global identity.

1.2.3 Chinese-language Internet. As will be detailed in Chapter 2 and Chapter 4, the fast-growing number of Chinese-speaking Internet users not only raises questions about the sense of fellow-users-of-the-same-language across major Chinese-speaking regions, but also highlights the importance of the main players who maintain these websites.

As for search engines, the Baidu Search Engine, the national champion of mainland China listed in the US NASDAQ stock exchange, has dominated the market of mainland China, whereas Google and Yahoo! have been dominant in Hong Kong and Taiwan (CIC, 2009; CNNIC, 2009). With regard to user-generated encyclopaedias, both Baidu Baike and Chinese Wikipedia are available for those who can type and read Chinese characters. Do these websites promote or discourage exchange among the users across different regions? Does the nature of the organisations (i.e. the for-profit companies such as Baidu, Google and Yahoo! versus the international non-profit organisations such as the Wikimedia Foundation) matter? In other words, how does the openness and interactivity of the Web and the way they are implemented foster exchange across Chinese-speaking regions? How does the provision of UGEEPs and SERPs shape the sense of fellow users across major Chinese-speaking regions by promoting interconnectivity, in practice, via the Chinese language?

Although it is true that sharing the same language provides the possibility for more interaction across Chinese-speaking regions or by connecting existing online Chinese language areas, this is not guaranteed, and more research with concrete data is required before making such claims. Several caveats exist for interactions across regions. For example, users from mainland China, Singapore and Malaysia use simplified Chinese characters, while users from Hong Kong and Taiwan predominantly use traditional Chinese characters. This may constitute barriers to interconnection among Chinese language areas and further detract from, rather than create, a sense of fellow-readers and or fellow-editors. In addition, although Hong Kong is now ruled by Beijing and Taiwan is claimed

by Beijing to be part of China (when in fact Taiwan has an independent government with de facto sovereignty that has never been ruled by the People's Republic of China), it may also be that online exchanges and a sense of fellow-users may also differ between these countries. Last but not least, the filtering and censorship regimes implemented by Beijing have imposed a divide between mainland China on one side and Hong Kong, Taiwan and other Chinese-speaking regions on the other. Therefore, a central contribution of this research is to compare how the two major user-generated encyclopaedias, in the context of the major search engines in the region, are developed and how they manage these boundary issues to provide content for users across regions. This will also provide insights into the current status of the Chinese-language Internet. Moreover, again, although search engines are not part of the core research question of this thesis, they are included whenever they provide an essential context (e.g. in Chapter 4) and data (e.g. in Chapter 5) for comparisons across countries and regions, languages and uses.

1.2.4 Research question. It has been argued that it is of critical importance to examine whether and how user-generated encyclopaedias have realized the Web's potential for openness and interconnectivity, how these create a sense of fellow-users within and across boundaries, and the implications for the Chinese-language Internet. To understand these topics, the thesis focuses on the following research question:

How have the two major Chinese-written user-generated encyclopaedias, Baidu Baike and Chinese Wikipedia, overcome, reinforced or shifted the existing cultural-political boundaries among mainland China, Hong Kong and Taiwan?

Therefore, this is a study of how one type of influential website, user-generated encyclopaedias (or a compilation of general-purpose human

knowledge), written in Chinese, affects the existing boundaries among major Chinese-speaking regions.

Readers who are not familiar with modern Chinese written language may make the wrong assumption that all Han Chinese people in the world speak and write the same language. This is not the case. To avoid confusion, the Chinese language discussed here refers to the written Chinese expressions that may use variations of Chinese characters (or *hanzis*), thus excluding languages such as Mongolian and Tibetan, which authorities in Beijing consider minority languages inside the People's Republic of China. This selection fits with the online reality that simplified Chinese characters and traditional Chinese characters can usually be exchanged and converted to one another online, while this is not the case with Chinese and Mongolian or Chinese and Tibetan. In addition, only Mandarin Chinese is analysed here, excluding other popular languages (or dialects) such as Shanghainese and Cantonese.

Even within modern Mandarin Chinese in its written form, there are many existing boundaries among the Chinese-speaking populations in the world. There are at least two major language scripts (in terms of the choice of Chinese characters) and four regional differences that are migrated into the digital networked environment. Hence, the thesis focuses on the three regions (mainland China, Hong Kong and Taiwan) where Mandarin is prevalent, while sometimes including other regions such as Macau and Singapore as additional reference points. The next section summarizes the geographic and linguistic distinctions of Chinese language in the digital networked environment.

1.3 Boundaries: offline and online

This section will first introduce some major linguistic and geographic features used by information and communication technologies (ICTs) to serve different geolinguistically marked groups of users, and explain how such practices reintroduce offline boundaries online, which in turn may shape (or be shaped) by cultural-political expectations and decisions.

Variants of Chinese are summarized in Table 1-1, which shows different language support by major companies such as Google, Wikipedia, Facebook and Microsoft. Overall, three regions (mainland China, Hong Kong and Taiwan) enjoy dedicated support from most companies, each with its own Google Search, Google News, Facebook and Microsoft operating system arrangement. In contrast, Macau does not have as much support. Although specific support for Macau may be provided in the future, some companies seem to regard the difference between Hong Kong and Macau small enough to have the same support and features. Similarly, Malaysian Chinese does not have widely dedicated support.

Table 1-1

Chinese-language support by Google Search, Google News, Wikipedia, Facebook and Microsoft Windows as of 2012

Country Code (Name)	Google Search	Google News Editions	Chinese Wikipedia	Facebook Interface	Microsoft Windows 7
CN (Mainland China)	Dedicated Support	Dedicated Edition	Mainland Simplified	Chinese (Simplified)	Dedicated Support
HK (Hong Kong)	Dedicated Support	Dedicated Edition	Hong Kong & Macau Traditional	Chinese (Hong Kong)	Dedicated Support
MO (Macau)	None	None	Hong Kong & Macau Traditional	None	Dedicated Support
MY (Malaysia)	None	None	Singapore & Malaysia Simplified	None	None
SG (Singapore)	Dedicated Support	None	Singapore & Malaysia Simplified	None	Dedicated Support
TW (Taiwan)	Dedicated Support	Dedicated Edition	Taiwan Orthodox	Chinese (Taiwan)	Dedicated Support

The thesis thus focuses its analysis on the three regions that enjoy dedicated support: mainland China, Hong Kong and Taiwan. This focus serves the aim of the thesis in studying the boundary dynamics because of the intricate

cultural-political boundaries among the three regions (to be detailed in the following section). Other regions such as Macau and Singapore are included only sparingly in the data collection and analysis, depending on the availability of the data.

By selecting these regions as units of analysis, this thesis fills a gap by conducting a formal and thorough analysis regarding how such Chinese variants reintroduce and embody offline boundaries. To the best of my knowledge, this is among the first such attempts (the few others, including their limitations, will be reviewed in Chapter 2). In addition, the analysis avoids methodological nationalism, which will be further detailed in the literature review and methodology chapter. It should be mentioned at this point already that while this thesis focuses its discussion on the Chinese-language Internet, the ways in which it analyses the geographic and linguistic features of online activities can be more generally applied to other languages, especially those that have sizable populations across state boundaries, such as Hindi, Spanish, English, German and Arabic (Coulmas, 2000).

1.3.1 Baseline: existing boundaries. Table 1-2 summarizes some salient features of the political and media environment of each Chinese-speaking region to show the existing political boundaries among them. The first two columns show four major countries and their corresponding country codes. The other columns show five other indicators regarding the political and media environment of each country. The following paragraphs explain the rationales, with documentation, for using them as a baseline. In terms of Chinese characters or scripts, both Hong Kong and Taiwan have continued to use traditional Chinese characters (officially called “orthodox Chinese” in Taiwan), whereas mainland China and Singapore use simplified characters or scripts. The difference between traditional and simplified characters has cultural, political and thus technical implications, which will be unpacked in detail in the

following sub-section in the context of modernisation of the Chinese language and the Cold War divide.

Table 1-2

Four Major Chinese-speaking Regions and Their Records in Political Internet filtering, Free Speech, Democracy and Human Rights

Country Code (Name)	Political Internet Filtering ^a	Networked Readiness Index Ranking ^b	Democracy World Ranking ^c	Free Speech World Ranking ^d	Major Human Rights Violation ^e
CN (Mainland China)	Pervasive	59	138	163	Yes
HK (Hong Kong)	None	22	78	61	No
SG (Singapore)	None	5	84	141	No
TW (Taiwan)	None	13	32	32	No

^a Data from the OpenNet Initiative about Internet filtering of political contents.

^b Data from the Networked Readiness Index 2006-2007 Ranking, by World Economic Forum, among 122 countries surveyed.

^c Data from the Economist Intelligence Unit's index of democracy 2006, with 167 countries are ranked from the top democracies to authoritarian regimes.

^d Data from the Worldwide Press Freedom Index 2007 made by the Reporters without Borders, with 169 countries ranked.

^e Whether a state is mentioned in the Human Rights Watch World Report 2007.

In terms of membership of political institutions, both mainland China and Hong Kong now belong to the People's Republic of China, whereas Taiwan belongs to the Republic of China. Singapore belongs to the Republic of Singapore.

In terms of free speech and democratic development, the four regions have what historian Mitter (2008, p. 132) has succinctly described as “intriguing divisions between what is 'free' and 'democratic' ”:

China itself is neither fully free nor democratic. Taiwan, since the 1990s, has been both free and democratic. Singapore, a largely Chinese society, is democratic, in that it has regular elections which are nominally open to

opposition candidates (but at high cost to themselves), but is not free (the media and political activism are both heavily regulated). Most intriguing is Hong Kong, which is little more democratic than it was under the British. Yet it is a very free society: although there is political pressure and a certain level of self-censorship, it has a lively press, it is easy to publish books attacking the Chinese government, and it supports a variety of political parties (although the legislature is arranged to prevent any such party ever coming to power). There are few, if any, other such free, undemocratic societies.

What is summarized in Table 1-2 is consistent with the above description. Among the four, Taiwan is the best in all categories of democracy, free speech, and human rights; Hong Kong leads Singapore by a small margin; and the Mainland China is last. These offline differences have repercussions online as well. Mainland China has “pervasive” political filtering. Reporters without Borders (2007) claims that China is responsible for the 49 out of 64 imprisoned cyber-dissidents in the world.

Depending on their experience as citizens in their respective societies, Chinese-speaking Internet users may therefore also have different expectations of the Internet. In addition, since simplified Chinese characters are associated with less free and democratic societies, the question regarding the interplay between technologies and values is bound to arise in the design and exchange of Chinese-language spaces on the Web and in computer networks.

1.3.2 Chinese characters, modernity and the Cold War. The difference between traditional and simplified Chinese characters is not only visible but also significant for its association with Chinese modernity and the history of the Cold War. Somewhat akin to the spelling differences between American versus British English in words such as “color” versus “colour”, the visible difference in Chinese characters is shown by the characters that are double-underlined and in red in Table 1-3.

Table 1-3

Visible difference (double-underlined and in the colour red) of the same text written in simplified and traditional Chinese script

	Chinese, <u>Simplified</u>	Chinese, <u>Traditional</u>
UDHR*	<p>世界人权宣言</p> <p>联合国大会一九四八年十二月十日第217A(III)号决议通过并颁布</p> <p>1948年12月10日，联合国大会通过并颁布《世界人权宣言》。这一具有历史意义的《宣言》颁布后，大会要求所有会员国广为宣传，并且“不分国家或领土的政治地位，主要在各级学校和其他教育机构加以传播、展示、阅读和阐述。”《宣言》全文如下：</p>	<p>世界人權宣言</p> <p>聯合國大會一九四八年十二月十日第217A(III)號決議通過並宣布</p> <p>1948年12月10日，聯合國大會通過並頒布《世界人權宣言》。這一具有歷史意義的《宣言》頒布後，大會要求所有會員國廣為宣傳，並且「不分國家或領土的政治地位，主要在各級學校和其他教育機構加以傳播、展示、閱讀和闡述。」《宣言》全文如下：</p>
Preamble	<p>序言</p> <p>鉴于对人类家庭所有成员的固有尊严及其平等的和不移的权利的承认，乃是世界自由、正义与和平的基础，</p> <p>鉴于对人权的无视和侮蔑已发展为野蛮暴行，这些暴行玷污了人类的良心，而一个人人享有言论和信仰自由并免于恐惧和匮乏的世界的来临，已被宣布为普通人民的最高愿望，</p> <p>鉴于为使人类不致迫不得已铤而走险对暴政和压迫进行反叛，有必要使人权受法治的保护，</p> <p>鉴于有必要促进各国间友好关系的发展，</p> <p>鉴于各联合国国家的人民已在联合国宪章中重申他们对基本人权、人格尊严和价值以及男女平等权利的信念，并决心促成较大自由中的社会进步和生活水平的改善，</p> <p>鉴于各会员国业已誓愿同联合国合作以促进对人权和基本自由的普遍尊重和遵行，</p> <p>鉴于对这些权利和自由的普遍了解对于这个誓愿的充分实现具有很大的重要性，</p> <p>因此现在，</p> <p>大会，</p> <p>发布这一世界人权宣言，作为所有人民和所有国家努力实现共同标准，以期每一个人和社会机构经常铭记本宣言，努力通过教导和教育促进对权利和自由的尊重，并通过国家的和国际的渐进措施，使这些权利和自由在各会员国本身人民及在其管辖下领土的人民中得到普遍和有效的承认和遵行；</p>	<p>序言</p> <p>鑑於對人類家庭所有成員的固有尊嚴及其平等的和不移的權利的承認，乃是世界自由、正義與和平的基礎，</p> <p>鑑於對人權的無視和侮蔑已發展為野蠻暴行，這些暴行玷污了人類的良心，而一個人人享有言論和信仰自由並免于恐懼和匱乏的世界的來臨，已被宣布為普通人民的最高願望，</p> <p>鑑於為使人類不致迫不得已鋌而走險對暴政和壓迫進行反叛，有必要使人權受法治的保護，</p> <p>鑑於有必要促進各國間友好關係的發展，</p> <p>鑑於各聯合國國家的人民已在聯合國憲章中重申他們對基本人權、人格尊嚴和價值以及男女平等權利的信念，並決心促成較大自由中的社會進步和生活水平的改善，</p> <p>鑑於各會員國業已誓願同聯合國合作以促進對人權和基本自由的普遍尊重和遵行，</p> <p>鑑於對這些權利和自由的普遍了解對於這個誓願的充分實現具有很大的重要性，</p> <p>因此現在，</p> <p>大會，</p> <p>發布這一世界人權宣言，作為所有人民和所有國家努力实现共同標準，以期每一個人和社會機構經常銘念本宣言，努力通過教誨和教育促進對權利和自由的尊重，並通過國家的和國際的漸進措施，使這些權利和自由在各會員國本身人民及在其管轄下領土的人民中得到普遍和有效的承認和遵行；</p>
Article 1.	<p>第一条</p> <p>人人生而自由，在尊严和权利上一律平等。他们赋有理性和良心，并应以兄弟关系的精神相对待。</p>	<p>第一條</p> <p>人人生而自由，在尊嚴和權利上一律平等。他們賦有理性 and 良心，並應以兄弟關係的精神相對待。</p>
Article 2.	<p>第二条</p> <p>人人有资格享有本宣言所载的一切权利和自由，不分种族、肤色、性别、语言、宗教、政治或其他见解、国籍或社会出身、财产、出生或其他身分等任何区别。并且不得因一人所属的国家或领土的政治的、行政的或者国际的地位之不同而有所区别，无论该领土是独立领土、托管领土、非自治领土或者处于其他任何主权重受限制的情况之下。</p>	<p>第二條</p> <p>人人有資格享受本宣言所載的一切權利和自由，不分種族、膚色、性別、語言、宗教、政治或其他見解、國籍或社會出身、財產、出生或其他身分等任何區別。並且不得因一人所屬的國家或領土的政治的、行政的或者國際的地位之不同而有所區別，無論該領土是獨立領土、託管領土、非自治領土或者處於其他任何主權受限制的情況之下。</p>
Article 3.	<p>第三条</p> <p>人人有权享有生命、自由和人身安全。</p>	<p>第三條</p> <p>人人有權享有生命、自由和人身安全。</p>
Article 4.	<p>第四条</p> <p>任何人不得使为奴隶或奴役；一切形式的奴隶制度和奴隶买卖，均应予以禁止。</p>	<p>第四條</p> <p>任何人不得使為奴隸或奴役；一切形式的奴隸制度和奴隸買賣，均應予以禁止。</p>
Article 5.	<p>第五条</p> <p>任何人不得加以酷刑，或施以残忍的、不人道的或侮辱性的待遇或刑罚。</p>	<p>第五條</p> <p>任何人不得加以酷刑，或施以殘忍的、不人道的或侮辱性的待遇或刑罰。</p>
Article 6.	<p>第六条</p> <p>人人在任何地方有权被承认在法律前的人格。</p>	<p>第六條</p> <p>人人在任何地方有權被承認在法律前的人格。</p>

*Note. * The Universal Declaration of Human Rights*

Despite sharing some of the same linguistic heritage, a language gap between variants of written Chinese has developed because of more than six decades of separation and political tensions during the Cold War (Sui, 2011). After the Chinese Civil War, the Chinese Communist Party founded its regime in Beijing in 1949 as the sole legitimate leader of the People's Republic of China. The Chinese Nationalist Party (also known as Kuomintang) fled from Nanjing, the capital the Republic of China, to Taiwan, which had recently been reclaimed from Japan. As a result, language planning, mandated by Beijing to use simplified Chinese characters with the modern aim to raise general literacy, has been restricted to mainland China (Ji, 2004; Liao, 2009a; Shouhui Zhao & Baldauf, 2007a). Even after the transfer of sovereignty over Hong Kong from the United Kingdom to the People's Republic of China in 1997, Hong Kong has continued to use traditional Chinese characters. It should be noted that the language preferences across these regions include not only the choice of Chinese writing systems, but also the choice of words, phrases, pronunciation, and more (Sui, 2011). The difference between traditional and simplified Chinese characters is only one of the most visible and salient differences among these.

The difference in Chinese characters continues to have cultural and political implications in the greater context of balancing Chinese tradition and modernity. For example, Taiwan, under the presidency of Ma Ying-jeou since 2008, has been proactive in reclaiming the role of “orthodox Chinese forms” essential for “a strong grounding in traditional Chinese culture” (Y. Ma, 2009, para. 3), including holding “Chinese character festivals” that highlight the essence, virtue and beauty of orthodox Chinese characters and traditional Chinese culture (e.g. Department of Cultural Affairs, 2005, 2011). This can be interpreted as a reaffirmation of the Kuomintang's official “Chinese Cultural Renaissance Movement” in the 1970s, which was designed to oppose the communist “Cultural Revolution” in mainland China, which had rejected and

discarded traditional culture and Confucian classics (Ho, 2006; S. O. Lee, 2007). Since the Chinese Cultural Revolution of the 1960s and 70s, Communist leaders in China have gradually shifted cultural policy toward appropriating Chinese traditional culture for party propaganda purposes and for generating a sense of indigenous superiority and patriotism among its followers (Shouhui Zhao, 2008). However, China's language policy has not yet changed on the choice of simplified characters over traditional ones. Therefore, the simplified versus traditional difference is not only visible to all Chinese-language readers, but also points to historical, political and cultural tensions. As a result, a first-ever joint dictionary by China and Taiwan, announced in 2011, aims to close the language gaps by recording different ways of writing and speaking Chinese (Sui, 2011).

1.3.3 Chinese regions, media and politics. As previously noted, mainland China is neither free nor democratic; Taiwan has been both free and democratic since the 1990s; Singapore is nominally democratic but not free; Hong Kong is free and slightly more democratic than it was under British rule (Mitter, 2008, chap. 4). However, mainland China has begun exercising its cultural power through Mandarin-teaching Confucius Institutes around the globe (Ding & Saunders, 2006) and its "soft power" or other political influences effectively on Taiwan (DeLisle, 2010; Koike, 2012; Lowther, Shih, & Chao, 2011), and Hong Kong (Fung & Lee, 1994; Higgins, 2012a). Thus, with the increasing exchange of people, goods and ideas across these regions, the media environment in Hong Kong and Taiwan has seen the politics and media sometimes divided along pro-China versus anti-China lines (Higgins, 2012b; C. P.-Y. Lai, 2007; C. C. Lee, 2001). It remains an open question how the interconnectivity of Internet will shape the cultural-political dynamics of the three regions, which make up the vast majority of Chinese-language Internet users in the world.

The regional differences within the Chinese-language Internet are thus arguably more extensive and serious than those in other major languages in the world, such as Hindi, German, Spanish, English and Arabic. How will the

cultural-political boundaries be drawn across these Chinese-speaking regions? Table 1-4 has simplified the cultural and political differences described in the previous section, with mainland China as the reference point in order to show the difference from other regions. For instance, in terms of the forms of Chinese characters used (orthography), Singapore uses simplified Chinese as in mainland China while Hong Kong and Taiwan use traditional/orthodox Chinese. In terms of Internet filtering, only mainland China has extensive filtering for political content. In terms of political institutions, Hong Kong belongs to the People's Republic of China, as does mainland China.

Table 1-4

The existing cultural and political boundaries among major Chinese-speaking regions, using mainland China as a reference point

Country Code (Name)	Standard Chinese script	Republic	Internet Filtering	Media Systems	Political System
CN (Mainland China)	<u>simplified</u>	<u>PRC</u>	<u>Pervasive</u>	<u>Not Free</u>	<u>Not Democratic</u>
HK (Hong Kong)	traditional	<u>PRC</u>	None	Free	<u>Not Democratic</u>
SG (Singapore)	<u>simplified</u>	ROS	None	<u>Not Free</u>	Democratic
TW (Taiwan)	traditional	ROC	None	Free	Democratic

Based on Table 1-4, possible boundaries can be drawn as follows:

- If Chinese characters constitute barriers for online interactions, the boundary should reflect the difference between (1) simplified-character users in mainland China and Singapore and (2) traditional-character users in Hong Kong and Taiwan.
- If political institutions (belonging to the same Republic) are the main differentiating criterion, the boundary should differentiate between (1) citizens of the People's Republic of China (PRC) in mainland China and

Hong Kong, (2) citizens of the Republic of China (ROC) in Taiwan and (3) citizens of the Republic of Singapore (ROS) in Singapore.

- If political Internet filtering (imposed by authorities in Beijing on users within mainland China) is seen as the key distinguishing factor, the boundary should differentiate between (1) mainland China and (2) the other three regions including Hong Kong, Taiwan and Singapore.

How does the Web reintroduce the existing cultural and political boundaries in the Chinese context?

1.3.4 Theoretical question: online boundaries. Generally, how has the Internet reshaped geo-cultural or geo-linguistic communities in relations to nation-states?

With a common language, Chinese-language Internet users have the potential to share Chinese-written knowledge and exchange Chinese-speaking voices online, regardless of where they come from. However, with regional differences that encompass a range of historical, cultural and political tensions, Chinese-language Internet thus has both an integrating potential and a potential for creating divides. This thesis, using the two major Chinese-written user-generated encyclopaedias as major sites of observation, asks how Baidu Baike and Chinese Wikipedia have overcome, reinforced or ignored the existing boundaries among mainland China, Hong Kong and Taiwan. The answers are expected to provide insights into the cultural politics of the Chinese Internet. The methodological and theoretical insights are expected to be useful for the geolinguistic analysis of the Internet.

1.4 Overview of subsequent chapters

Chapter 1 has provided the background to the research, formulated the general research question, and outlined the basic arguments of the thesis.

Chapter 2 will introduce a number of observations and theories on the link between national boundaries and information/communication systems from

various disciplines. The literature review covers the role of language and media technologies and thus paves the way for a theoretical framework to be developed which accounts for the ways in which elements of cultural-political boundaries have been reintroduced into the design and processes of information and communication spaces. The aim is also to review various research approaches so as to develop a feasible research design and methods to identify the factors and workings of such reintroduction of cultural-political boundaries to the Web. Special attention will be paid to the ways in which the research bias of relying too much on a national framework and/or unit of analysis, or “methodological nationalism” (Wimmer & Schiller, 2002), can be alleviated. This general research concern is especially important for this specific research on Chinese boundaries online because of the ambiguous yet complicated cultural-political boundaries among Chinese-speaking regions such as mainland China, Hong Kong and Taiwan.

Chapter 3 will establish a theoretical framework based on three concepts: “web spheres”, “processability” and “cultural thickening”. Its fundamental claim is that we are witnessing a development of reintroduction of cultural-political boundaries (e.g. national boundaries) into dominant media-language systems, which reshape the group dynamics between information and users. Researchers must study the corresponding cultural patterns to see how cultural-political boundaries are reintroduced. As the proposed theoretical framework has both quantitative and qualitative dimensions in relation to cultural patterns, the thesis develops a multi-method of structured, focused comparison (George & Bennett, 2004; Drozdova & Gaubatz, 2009) to examine how the two encyclopaedia websites deal with the existing cultural-political boundaries. This chapter will thus present a comparative case study that uses mixed methods to gather evidence from three main areas of user-generated content projects: editorial development, content outcome, and user reception, each of which will be discussed in turn in Chapter 4, Chapter 5 and Chapter 6.

Chapter 4 will present the comparative analysis of the two user-generated encyclopaedias' editorial development and patterns. First, editorial policies and practices are analysed to see how they process information differently, and how such differences are associated with the website's relationship with the hosting organizations and volunteer contributors. Second, to consider the historical context in which the two websites are launched in relation to potential users (after all, they are user-generated encyclopaedias), the remaining part of the chapter will contextualize the development of Baidu Baike and Chinese Wikipedia in relation to the growth of mainland Chinese Internet users and Beijing's filtering and censorship regime.

Chapter 5 continues the comparative analysis on the content and citations inside encyclopaedia articles to show how differently they bring the world's information to their users. The first half of Chapter 5 will present mostly quantitative evidence, using novel Webometric techniques that consider the geographic and linguistic features of the external links of all the articles in both encyclopaedias, collected in 2010. Careful analysis with appropriate geographic normalization indicates that the main differences between the two encyclopaedias do not manifest themselves along the US versus China line, but rather along that between the mainland China on one side and Hong Kong, Macau and Taiwan on the other. The second half of the chapter will supplement the first half with mostly qualitative evidence on the selected articles that deal with the notion of "Chineseness", providing additional insights about the cultural politics of being "Chinese" in addition to confirming the findings of the first half.

Chapter 6 will systematically examine how user-readers perceive and use the two encyclopedias, as observed on major platforms of search engines and microblogs. Acknowledging the fact that many users come to user-generated encyclopaedias via search engines, the first half of Chapter 6 presents mostly quantitative findings based on visibility tests that indicate whether a website (for this research, which user-generated encyclopaedia website) is listed in the

outcome of the SERPs. The second half of Chapter 6 will present findings observed from major microblog platforms of Sina Weibo and Twitter, which supplement the overall findings of the first half with comments expressed and experiences documented therein.

Finally, I will articulate the theoretical and policy implications of the findings in Chapter 7. The findings fill various gaps in our understanding of the Chinese-language Internet. In particular, they show where and how the cultural-political boundaries manifest themselves. It discusses how geolinguistic arrangements are critical for our understanding of how information is ordered and for its boundaries. Geolinguistic processability, or the overall socio-technical processes and outcomes that process linguistic and geographic specificity of information, matters for defining the centre and boundaries of information-rich societies.

Chapter 2 Literature review

How has the Internet reshaped geo-cultural or geo-linguistic communities in relations to nation-states? During the 1990s, when the Internet was in its early world development, it was treated as if technologies know no national boundaries, being described by one author as part of the “death of distance” phenomenon (Cairncross, 2001). In the 2010s, researchers have noticed the rise of localization, as illustrated by the demise of the universal search engine and the rise of local Google domains tailored to different groups of local users (Liao, 2011; Pariser, 2011; R. Rogers, 2013). Localization seems to reinstate geographic, linguistic and national boundaries on the Web. Thus, this literature review must discern the relationship between media technologies and national boundaries.

This study aims to compare two Chinese-language user-generated encyclopaedias (though the focus of this literature review is not on user-generated encyclopaedias per se, but rather on the dynamics of cultural-political boundaries). A systematic review of scholarly research on Wikipedia has been conducted (Okoli, Mehdi, Mesgari, Nielsen, & Arto, 2012) and the Wikimedia Research Newsletter¹ has been documenting and reviewing the latest research on Wikipedia and its sister projects since 2001. While some research uses Wikipedia merely as a huge dataset for information retrieval or data mining, a major research area is to understand the “collaborative ecosystem” of participation, content and readership. However, almost no research in this review is comparative, and none examines cultural politics across language regions (Okoli et al., 2012).

The following literature review will first examine various claims concerning the relationship between national boundaries and communicative spaces from various disciplines. Then, it will narrow its focus to Chinese

¹ See <http://meta.wikimedia.org/wiki/Research:Newsletter>

boundaries. Finally, a review on the recent development of information and language processing leads to a modified version of medium theory and an approach focusing on media-language systems. In terms of mediatization and how it fosters (or not) globalization, we shall see that encyclopaedias constantly face the realities of linguistic and national differences. Altogether, this review sets the stage for the theoretical framework and methods in the next chapter by seeing national boundaries as the cultural-political construction of information and communication systems.

2.1 National boundaries and communicative spaces

National boundaries tend to become more permeable and flexible, especially in the economic and information spheres (Albert, Jacobson, & Lapid, 2001; Brunn, 1998; Morley & Robins, 1995). International relation scholars discuss two competing hypotheses: one is the nation-state thesis, where national identities, borders and orders are mutually defined and reinforcing, while the other is the end-of-boundaries thesis, where globalization is expected to render national boundaries ineffective barriers to information and economic flows (Albert et al., 2001).

The transborder flow of information via satellite and the Internet has tested the willingness and capacity of states to open and regulate their communicative boundaries, raising questions regarding the new territorial dimensions of modern statehood, or “a treaty of Silicon for the treaty of Westphalia” (Brunn, 1998). Information and communication technologies are expected to reduce costs for political and economic transactions, which should in turn reshape power in a time- and space-compressed world (Brunn, 1998). While some once imagined a borderless world where the factor of distance becomes increasingly irrelevant (Cairncross, 2001), others have pointed to the tensions between the bordered world of national governance and the borderless world of the Internet (Banerjee, 2007; Goldsmith & Wu, 2008; Mueller, 2010).

Furthermore, Internet and communications scholars have been studying the link between technologies and national boundaries, conducting research on the connectivity effects of the Internet (Haythornthwaite, 2005). Making new connections beyond existing social and political boundaries is central to the Internet's potentials for enhancing, redefining and reshaping civic engagement and civic culture (Benkler, 2006; Rheingold, 2008; Scammell, 2000; Shirky, 2010), with Wikipedia being cited as one of the major examples (Leung, 2009; Lih, 2009; Reagle, 2008). Additionally, the question of whether and how the Internet has shaped transnational activism and diaspora communities has received significant attention (e.g. Cammaerts & Van Audenhove, 2005; G. Yang, 2003), stimulating discussions on the transnational public sphere (e.g. Fraser, 2007; Wessler, Peters, Brüggemann, Kleinen-von KönigsLöw, & Sifft, 2008).

There is no clear answer when it comes to the relationship between national boundaries and Internet connectivity, and the connectivity effects are not evenly distributed. With the aim to identify the major disciplinary views and gaps on the subject of national boundaries, the first part of this chapter will summarize the major work from systems theory, communications theories of nationalism, comparative media research and national web studies.

2.1.1 National boundaries and systems theory. Boundaries of systems can be nationally defined, such as national systems of innovation (Högselius, 2006), or national media systems and national information systems (United States Congress, 1981). The concept of system boundaries provides one approach to national boundaries.

The concept of boundary is central to systems theory: "boundary maintenance is system maintenance" (Luhmann, 1995, p. 17). Based on the work of system theorists Karl Wolfgang Deutsch and Niklas Luhmann, Högselius (2006) defines two types of system boundaries by comparing domestic and foreign interactions. First, a Deutsch-type boundary emerges when the intensity of domestic interactions is significantly greater than the intensity of cross-border

interactions. Second, a Luhmann-type boundary marks a nation-specific structure and style. The two thus capture, respectively, the quantitative and the qualitative aspects of the ways in which systems regulate activities. Although the definitions by Högselius (2006) were meant to test the existence of national innovation systems, they offer more general analytical directions. Researchers can observe system activities (including the intensity and style of interactions) to assess whether system boundaries are nationally aligned.

Similarly, activities relating to the Internet can be examined to evaluate the Internet's connectivity effects on national boundaries. Different intensities of interactions indicate Deutsch-type boundaries, while distinct structures or styles suggest Luhmann-type boundaries. In addition to observing and analysing both quantitative and qualitative aspects of interactions, researchers can further ask whether changes in quantity lead to changes in quality. For example, does a quantitative increase of activities (Deutsch-type boundary) lead to a qualitative distinction (Luhmann-type boundary)? Conversely, does a distinction in the structure and style of activities contribute to different levels of interaction intensity *within* versus *across* national borders?

2.1.2 Communications theories of nationalism. Providing a more material foundation than systems theory, communications theories of nationalism highlight the mediating role of media in marking the boundaries of nation-states through national communicative spaces (Schlesinger, 2000, 2001).

Inspired by the study of information, communication and control in machines and biological organisms (Wiener, 1948), Karl W. Deutsch (1951) conducted inquiries based on cybernetics into human societies, including the issues of nationalism and nationality, which he called "some of the most baffling problems of modern science" (1951, p. 234). By recognizing that "[c]ommunication was social before it became elaborately technological" (p. 239), he studied the role of communication and control in human societies by defining "a *people* as a large group of persons linked by complementary habits and

facilities of communication, or more briefly, as a community of both internal and social communications equipment" (p.234). Social communication among individuals can thus enable national cohesion and unity "on a larger scale" than unscalable "mutual rapport" (Deutsch, 1966a, p. 188). Such scalability constitutes "communicative barriers" and "marked gaps" in the efficiency of communication relative to other groups (Deutsch, 1966a, p. 100). Simply put, nation-states must have scalable "facilities for storing, recalling, and recombining information, channels for its dissemination and interaction, facilities for deriving further information" (Deutsch, 1966a, p. 75).

Similarly, social anthropologist Ernest Gellner (1983) examined the role of mass education as the basis for industrialization and the emergence of nationalism. He argued that because industrial societies require educated workers and thus a national education system, "the global norm is a set of discontinuous breathing chambers or aquaria", and each national chamber has "relatively superficial, but deliberately stressed, brand-differentiating characteristics" (Gellner, 1983, p. 50). Thus, while the content of national education may differ from one nation-state to another, the very idea of delivering a universal, standardized system of education has become the global norm of developing nation-states.

In addition, nationalism researcher Benedict Anderson has drawn attention to how print capitalism fosters national consciousness based on the routine sense of "fellow-readers" (1983). Although his print-centric theory was recently updated by taking into account electronic communications and transnational migration (B. Anderson, 1992, 2001), it rightly highlights the historical role of print capitalism and languages in building national consciousness. Different print languages thus constitute boundaries for print markets and political communities.

The sense of cultural-political enclosure is best described by the social communications tradition, as argued by Schlesinger (2001, p. 28): "Media are

boundary markers, intimately related to the ‘political roof’ that caps a culture and makes it into a nation-state.” Thus, Habermasian public spheres have boundaries that coincide with quintessentially national boundaries: nation-states “ha[ve] provided the retaining wall, not only for the all-inclusive Deutsch-Gellner-Anderson conception of social communication as national culture but also for the more narrowly conceived Habermasian domain of political communication” (Schlesinger, 2001, p. 28). Communications theories of nationalism therefore postulate that communicative boundaries have become the *most* significant factor for the historical development of nation-states, including information/control systems, education systems, media systems and political communication systems. Nation-states and their boundaries are thus socially and historically constructed by those systems.

Communications theories of nationalism suggest that nations gain cohesion and solidarity by sharing a common communicative space that is traditionally “capped” by national boundaries. For the purposes of this thesis, these theories are useful in theorizing the existing boundaries across Chinese-speaking regions to account for the complexity of Chinese modern societies with varied government, education, media and language systems. However, these theories do not focus on the specificity of media and languages in a globalizing world, so we now turn to the literature of comparative media systems

2.1.3 Comparative media systems research and geolinguistic regions.

Complementing and sometimes challenging the communications theories of nationalism, transnational and national media systems research (e.g. Kraidy, 2011; Volkmer, 2011; Y. Zhao, 2011) constitutes another body of work that takes up the boundary questions of nation-states in relation to the increasingly global media systems.

In the field of international comparative media systems, Hallin & Mancini’s two edited volumes (2004, 2011) signified a shift from normative evaluations influenced by the Cold War (e.g. Siebert, Peterson, & Schramm, 1963)

to systematic inquiries for comparing media systems under a unifying conceptual framework. By conceptualizing the functions and structures of media systems, this approach has been used to compare and contrast the specific characteristics of each single system. The initial framework, designed mostly for Western democracies, had four dimensions: structure of media markets, political parallelism, professionalization of journalism and the role of the state concerning media systems. Such a comparative approach bridges nation-centred studies of media systems and globalization research, as exemplified by the convergence or homogenization thesis, mostly observed in Europe, where European countries might be pushed toward the North Atlantic or Liberal Model. While this body of work provides political analysis of the structural and functional differences of media systems across countries, it nonetheless assumes the national boundaries of media systems. While the nation-state unit of analysis helps structure the comparative approach in a systematic fashion, some scholars argue that such “methodological nationalism” (Wimmer & Schiller, 2002) is inadequate for understanding media systems that are no longer exclusively bounded to single political systems (Jakubowicz, 2010). Other researchers also question Hallin & Mancini’s (2004, 2011) case for the growing influence of the North Atlantic Liberal model throughout the world, citing cases such as the Chinese media system (Y. Zhao, 2011) and Pan-Arab media (Kraidy, 2011). Thus, although the theory of international comparative media systems concerns media systems per se – which differs from the aforementioned communication theories of nationalism – both theories suffer potential issues of methodological nationalism.

On media globalization, Sinclair, Jacka & Cunningham (1996) proposed the concept of “geolinguistic regions” to examine the marketing strategies and flow patterns of satellite televisions. Explicit in the term “geolinguistic regions” is the role of language and cultures. The linguistic choice made by distributors and viewers on TV programming influenced audience market segmentations,

resulting in geolinguistic or geocultural regions that may include diasporas. These regions include Greater China, Latin America, the Middle East, India and, in the English-speaking world, Canada and Australia. As opposed to Hallin & Mancini's (2004, 2011) focus on Europe and North America, Sinclair, Jacka & Cunningham's selection of these "peripheral" regions, along with the geolinguistic approach, challenges the "cultural imperialism" view (which focuses on the increasing dominance of certain cultures) of globalization that overlooks linguistic and cultural factors. Thus, as a particular form of media technologies that can transmit and distribute information over distance (potentially across national boundaries), global TV systems effectively group countries according to their linguistic, cultural and historical connections. The geolinguistic approach, as defined by this body of work on global TV, suggests emerging geolinguistic spheres with respective production centres, such as "Mexico and Brazil for Latin America, Hong Kong and Taiwan for the Chinese-speaking populations of Asia, Egypt for the Arab world, India for the Indian populations of Africa and Asia" (p.8).

The kind of regionalization inherent to the concepts of geolinguistic and geocultural regions makes the following assumptions and observations: (1) Content consumption is a dynamic process governed by the cultural identities of audiences and the sedimentation of social practices; (2) Content production depends on export viability and scalability, and the regionalization of cultural and media markets often depends on language (and thus geo-cultural) barriers as natural protective barriers.

Thus, the global flow of cultural products is no longer a simple picture of media imperialism with the West at the centre dominating the rest at the periphery, but rather a number of geolinguistic and geocultural regions having their internal dynamics and external contact points. Common cultural, linguistic and historical connections thus delineate the extent of geolinguistic or

geocultural regions that may transcend physical distances and include diaspora communities.

2.1.4 National web studies. Informetrics, the study of quantitative aspects of information, has witnessed significant growth in webometrics and visualization (Bar-Ilan, 2008), with some studies having attempted to analyse “national domains” or other national dimension of Web systems (Baeza-Yates, Castillo, Telefónica, & Efthimiadis, 2007; Graells & Baeza-Yates, 2008; Thelwall & Wilkinson, 2003; Tolosa, Bordignon, Baeza-Yates, & Castillo, 2007; Wilkinson & Thelwall, 2012). Noting the shift in the scholarly terms from “the Web” to “national webs”, digital method research scholar Richard Rogers (2013) called this emerging field of research “national web studies”. Similar to a Deutsch-type boundary, a boundary in the context of national web studies is drawn by the quantitative contrast of high versus low intensity of web links. Boundaries can also be demarcated by web domains, languages, institutions, website platforms, etc., on which a Luhmann-type boundary may be identified for cultural analysis. For example, in his study of the Iranian national Web, Rogers (2013) extended quantitative concerns to the cultural analysis of the information output of dominant devices, calling for methods that consider the particular device cultures that users belong to, especially the leading sites of a country and/or language. National device cultures are organized by the ways in which different websites are linked, valued and ranked. Studying device cultures involved the organization of interactions, which is related to the debate on the algorithmic structuring of online public opinion and whether it amounts to a public sphere with socially integrating effects (R. Stuart Geiger, 2009).

Complicating the issue even further, recent studies on the assumed time-compressed and distance-reduced connectivity effects of the Internet have produced mixed findings (Barnett, Chung, & Park, 2011; Leamer & Storper, 2001; Mok & Wellman, 2007; Mok, Wellman, & Carrasco, 2010; Oh, Curley, & Subramani, 2008; Takhteyev, Gruzd, & Wellman, 2012). Some researchers have

argued that, despite the technology of the Internet and low-cost telephony, face-to-face socially close interaction has hardly changed between the 1970s and the 2000s (Mok & Wellman, 2007; Mok et al., 2010). Some geographers have argued that the Internet will produce similar forces for deagglomeration within geographically limited ‘neighbourhoods’ (Leamer & Storper, 2001). One study showed that national borders and language differences, along with other geographic factors, predict social ties in the social media platform, Twitter (Takhteyev et al., 2012). Still other researchers have found a linking pattern among countries that is consistent with the world-systems theory of core and peripheral structure (Barnett et al., 2011). One way to explain all these diverse implications of the connectivity effects of the Internet is to consider the perceptions of distance influenced by computer-mediated communications (Oh et al., 2008). Thus, when examining the relationship between national boundaries and Internet connectivity effects, it is viable to study the information systems or their outcomes, thereby assessing whether national boundaries are overcome or reinforced.

Unlike the aforementioned research on national boundaries, national web studies (Baeza-Yates et al., 2007; Graells & Baeza-Yates, 2008; Thelwall & Wilkinson, 2003; Tolosa et al., 2007; Wilkinson & Thelwall, 2012) often requires quantitative analysis of communication and content data used in information or computer-mediated communication systems. In many ways, such a data-intensive approach may provide new opportunities to move beyond the conventional research pitfalls of “methodological nationalism” (Wimmer & Schiller, 2002) or “implicit state-centrism” (Derudder & Witlox, 2005). For instance, instead of assuming that national boundaries are engrained automatically in Twitter networks, researchers can examine whether and how the presence of national boundaries inhibit certain social ties (Takhteyev et al., 2012). Some research has demonstrated that nations can no longer be taken for granted as a unit of analysis, but rather should be considered as a research object to be examined.

Instead of assuming the existence of particular national boundaries, in web studies national expressions are expected to be captured by quantitative and qualitative analysis of data.

The above literature review has summarized some major claims regarding the connections between national boundaries and communicative spaces. National boundaries are not physical national borders, but rather the construction of communicative processes. The next section will provide a similar literature review, but with a focus on Chinese national boundaries.

2.2 Chinese national boundaries and media technologies

To build a modern nation-state for the world's most populous country is no easy task. Modern Chinese history can be seen as a combination of efforts to build social and technological systems that can be called “Chinese”. The following review aims to assess the existing boundaries across Chinese-speaking regions while addressing the role of Chinese language in defining a civilization, developing nation-states and building information systems.

I will first contextualize selected research on Chinese Wikipedia within the cultural-political background where Chinese intellectuals have sought solutions for its political systems and information systems since the late 1980s, when the so-called “three theories” (information theory, cybernetics, and systems theory) were introduced. Then, based on the communicative theory of nationalism, I will review how Chinese writing systems underwent nationalization processes in the region and the implications for delineating modern Chinese and East Asian cultural spheres. The review will then discuss how global TV research, using concepts such as geolinguistic regions or geocultural regions, has produced important comparative findings for the boundary dynamics across Chinese-speaking regions since the 1980s, which sets the historical context for Internet diffusion in the region. The last section will then summarize the problems and challenges of current Chinese web studies as an instance of national web studies.

2.2.1 Systems, order and chaos. Although “boundary maintenance is system maintenance” (Luhmann, 1995, p. 17), one major cultural-political discussion in China has revolved around the maintenance of political systems.

Systems theory and cybernetics, along with the theory of communication – together called the “three theories” – were at the centre of intellectual debate in China in the 1980s (Keane, 2007), leading to the development of finding administrative and management solutions to social problems, a process that has been described as the “mandarinization of systems theory” (Woodside, 2006). The discussions and applications at the time were part of the efforts to go beyond orthodox Marxism. One notable application of systems theory and cybernetics to the understanding of Chinese societies is Jin Guantao’s controversial thesis on the ultra-stable system of Chinese feudalism (J. Wang, 1996). By “ultrastability”, he meant repeating cycles of disruption-restoration, or order-chaos (zhì-luàn 治-亂). His thesis had influences on the highly popular TV documentary series “River Elegy” (hé shāng 河殤), which sought a radical break from the Chinese cyclic past by criticizing the problems of Chinese civilization. The documentary was aired in June 1988 and inspired the 1989 pro-democracy movement (Yangwen Zheng, 2011). Also, in terms of political geography, it implicitly praised the rise of South China and the coastal cities under reform (Cartier, 2011).

The metaphors used in “River Elegy” call for a cultural-political change in China. In “River Elegy,” the symbol of Chinese civilization, the Yellow River, is sand-clogged with the backward, authoritarian features of earthbound civilization. China must resort to the open blue sea for new inspiration, which represents progress, democracy and internationalization. Seen from the perspective of Luhmann’s notion that boundary maintenance is system maintenance, the pro-reform message thus involves shifting the boundaries and styles of China’s cultural-political systems. According to the TV series, these boundaries should be more open, progressive and willing to embrace external

and new maritime culture, thereby breaking away from the stagnant historical cycles of order-chaos.

After China's suppression of the 1989 pro-democracy movement (1989 民主运动), the dichotomy of order and chaos became part of the official and increasingly popular discourse to maintain the stability and order of China's party-state political system. The democratization of Taiwan and South Korea has since been described as "chaotic copies of the West", praising the order and discipline in Singapore (van Kemenade, 2010). The official discourse has developed the "harmonious society" rhetoric as a defense against chaos and the "stability maintenance" (维稳 *wéiwěn*) policies that mix "Confucian hierarchism, legalist authoritarianism, and Communist dictatorship" (J. T.-H. Lee, Nedilsky, & Cheung, 2012, p. 20). Meanwhile, in popular discourse, a notable illustration came from Hong Kong-born Hollywood star Jackie Chan, who publicly singled out Hong Kong and Taiwan as being too free and thus chaotic (Cole, 2009). Therefore, while the cultural-political discussions of China's political systems appear to have shifted towards a certain form of system maintenance, they also implicitly involve boundary maintenance by distancing China from other political systems in Hong Kong and Taiwan in relation to democracy and freedom.

It is in this cultural-political context that a body of work on information ordering process (*xìnxī yǒuxùhuà* 信息有序化) based on Wikipedia (Ma & Xia, 2009; Ma, 2008) and other Web 2.0 platforms (Ma, 2012) is reviewed here. Using the conceptual frameworks of self-organization (Prigogine & Stengers, 1984) and complex adaptive systems (Holland, 1995), Ma (2012) compared the user-contributing policies of Chinese Wikipedia, Baidu Baike and Tianya Forum (which can arguably be seen as complex adaptive systems). He found that, in contrast to Baidu Baike's top-down, centralized policies, Chinese Wikipedia's policies had evolved into a complex yet self-adaptive bottom-up, decentralized editorial environment. He thereby argued that the ordering and organization

patterns of Chinese Wikipedia are closer to what Austrian thinker Friedrich A. Hayek (1967, 1984) conceptualized as “spontaneous order”. Here, Ma explicitly pointed out Hayek’s emphasis on free competition in market economics theory. Ma’ ideas are in line with my previous work (Liao, 2009c), though I used Michael Polanyi’s polycentric conceptualization of “spontaneous order”. Differences in interpretation aside, the concept of spontaneous order, or the notion of order-out-of-chaos, breaks with the dichotomy between order versus chaos.

The above discussion provides the relevant cultural-political context for viewing the two information systems – Chinese Wikipedia and Beijing’s internet censorship/filtering regime – as important information ordering processes that involve different Chinese cultural-political systems. As briefly discussed in Chapter 1, the boundary dynamics across Chinese-speaking regions involve political struggles in relation to the experiences with and attitudes towards political systems across these regions. Beijing’s internet censorship/filtering regime therefore constitutes a major factor. Luo & Fu (2008) assessed the impact of the Great Firewall of China on Chinese Wikipedia’s information ordering process. Their research shows that it has negative impacts on its growth rates of new users, new edits and rankings among other language versions of Wikipedia. They argue that it will reinforce the cultural and communicative barriers among Chinese-language users and damage the content neutrality of Chinese Wikipedia, on the grounds that the absence of mainland Chinese participation entails the absence of mainland Chinese viewpoints. It is questionable, however, whether mainland Chinese users and their viewpoints are absent from Chinese Wikipedia. For instance, another study with more sophisticated data analysis (X. Zhang & Zhu, 2011) showed that blocking the mainland Chinese access to Chinese Wikipedia not only curbed participation among mainland Chinese users, but also triggered discouraging social effects for those who are not blocked (typically traditional Chinese users who reside outside mainland China): their contributions decreased by 42.8%. Nonetheless, both studies underline the role of

Beijing's censorship/filtering regime in shaping the information order and boundaries.

The discussions reviewed above thus provide an antithesis to what Zhou (2005) criticized as the “monster complex”: Many observers of China tend to see the Internet as a benign monster that will break through China's authoritarian political system. (Recall Jimmy Wales' statement in the beginning of this thesis.) Contrary to the notion that a free information system will change the existing political system of China, different notions of order and chaos have been shaping the ways in which political systems and information systems are managed. It is no surprise then that China's censorship/filtering regime has become an integral part of national efforts in maintaining stability (G. Yang, 2012).

2.2.2 Chinese writing systems, nationalism and civilization.

Communication theories of nationalism, as reviewed in section 2.1.2, suggest that media are boundary markers that cap a culture under a political roof, including the use of print languages and national education systems. This section focuses on the pre-modern and modern role of Chinese writing systems in East Asia.

Commenting on the nationalization of writing, linguist Coulmas (2000) argues that Chinese language provides the paradigm case for the nationalization of a script that was once universal for the pre-modern societies of East Asia. Coulmas identified five major writing systems that altogether cover the entire world: Roman, Cyrillic, Arabic, Chinese, and Brahmi-derived Indic.² Chinese characters were instrumental for the states to recruit Mandarin scholar-bureaucrats through civil service examination systems. They also constituted a code for interregional communication among the elites of China and other East Asian countries. Coulmas (2000) argues that modern China, Japan, Korea and

² Children's history books, for example, often describe a division of the world into five cultural spheres: Western (Latin), East Asia (Chinese script), Islamic (Arab), India (South Asia) and Eastern Europe (Slavic), which correspond to the five major writing systems.

Taiwan have inherited the statist attitude on written languages, and each has developed increasingly particular vernacular literacies, strengthened by modern compulsory education. Coulmas (2000) rightly describes the shift from the universal (often classic) Chinese writing as the shared written medium of an area of civilization towards European linguistic nationalism that promotes authenticity of languages as cultural assets for modernization and national ethnicity. The historical shift was particularly marked by modern Japan's colonization of Taiwan, Korea and Manchuria. Writing systems became the focus of political controversy as political symbols and identities in the region. North Korea abolished Chinese characters from its Korean. The Republic of China in Taiwan did not follow the People's Republic of China's reform to simplify Chinese characters, but Taiwan still achieved higher literacy rates faster. Divergent written forms of North and South Korean languages, as well as those of Beijing's and Taiwan's Chinese languages, became associated strongly with the identities of divided nations. An important element of national education programs and social communications, modern national languages have thus become a central boundary marker in East Asia, despite the fact that the classic Chinese writing system used to delineate borders of a civilization.

Contrary to Coulmas's (2000) "nationalization of writing" thesis, Bosworth's (2004) "globalization of writing" thesis posits that civilizations are longer-lived than the ephemeral nations they contain. Taking a long-term evolutionary perspective on civilizations (instead of the perspective which focuses on the construction of modern nation-states), Bosworth predicts that Chinese and Arabic writing systems will allow for emerging blocs that will challenge the existing world system, which is Western-oriented and based on Roman alphabets. Nonetheless, both scholars recognize the potential of new media and information technologies for these non-Western languages, and both acknowledge the importance of Chinese language for their theses. Therefore, more research is needed to see how the role of Chinese writing systems, in their

digitized forms, relates to the boundary dynamics among East Asian and Chinese-speaking regions. As discussed in section 1.3, with regard to the offline and online boundaries of Chinese-speaking regions, researchers need to show how the boundaries of cultural-political entries (e.g. sub-nation states, nation-states or civilizations) have been digitally constructed. While Bosworth (2004) claims that boundaries among Chinese-speaking regions would be overcome, Coulmas (2000) suggests that they will mostly be reinforced.

Goggin & McLelland (2009) recognize the digital difficulties that East Asian societies face during the internationalization of the Internet. More complex machines and systems are needed to support Japanese scripts and Chinese characters. Because of the orthographic difficulties, societies in the East Asian cultural sphere have encountered a slow start when it comes to office automation and computerization as compared to European languages that use alphabets. Nevertheless, Goggin & McLelland argue that character-based users are not always disadvantaged; the East Asian text might take longer to type, but they take less screen space than alphabets, which explains the more sophisticated mobile and short-messaging culture in East Asia. Empirical research has shown that, with the 140-character limit, East Asian languages such as Chinese and Japanese are more “expressive” (Dugan, 2011; Liao, 2013b; Neubig & Duh, 2013). Chinese characters become arguably more modern and sophisticated because the digital capacity has solved many processing issues that used to prevent them from being technologically ready, as in the cases of telegram and print technologies (Liao, 2009a). Digital support for the main East Asian languages (Chinese, Japanese and Korean) becomes an important milestone for the Internet’s internationalization. As put by a major East Asian Science Technology and Society (STS) scholar, Nakayama Shigeru (2001, p. 12):

East Asians are accustomed to dealing with a multibyte system, in contrast to Western monobyte reductionist culture. It may be that in the future our multibyte culture will prove advantageous for dealing with complex systems.

Thus, as digital capacity can turn orthographic difficulties into advantages in dealing with complex systems, more research is needed to examine whether digital Chinese writing systems provide a paradigm case for the nationalization or globalization of a script in both Chinese and East Asian contexts.

2.2.3 Chinese geolinguistic regions and Internet diffusion. This section now turns to the more narrowed time period of the 1980s, when transborder TV became the main media technology capable of overcoming the communicative boundaries among Chinese-speaking regions. The historical context is relevant for Internet diffusion since the late 1990s.

The regionalization of transborder TV markets in both Great China and East Asia has been consistently found in J. M. Chan's earlier (1996) and later work (2009). These TV markets have been largely influenced by common cultural tastes. The popular cultures of Taiwan and Hong Kong in general were "very influential in China" (Chan, 2009, p. 33) during the historical spillover of foreign TV programs in the 1980s, especially in the Pearl River Delta, where Hong Kong was not yet returned to China. Indeed, the wider impact of "Kong-Tai" (Hong Kong-Taiwan) style or culture, including film and music, has been instrumental in mainland China's development of its consumer society and commercialized mass culture (E. L. Davis, 2009). Kong-Tai culture is also part of the cultural-political debate on Chinese modernization during the so-called "culture fever" of the 1980s (J. Wang, 1996). Still, such regionalization is bound to encounter the political boundaries of nation-states. For instance, the aims of the governments of Beijing and Taipei are often to promote barriers imposed by nation-states. China has been "afraid of external ideological influence that may threaten the status quo, corrupt the public mind, and cause social instability" and

thus “still wary of alien forces coming from Hong Kong, Taiwan and the West” (Chan, 2009, p. 33). For example, acting on the principle of reciprocity, Taiwan attempted to prevent its domestic TV market from being penetrated by Chinese TV.

To discuss Chinese geolinguistic or geocultural regions is therefore anything but straightforward. Chan (1996) warned against the oversimplification of assuming a high level of cultural similarity among three constituent regions of Greater China, describing Hong Kong as a cosmopolitan city with a rich East-West interaction as a legacy of a British colonial city in Asia, Taiwan as a much less Westernized society with a continued emphasis on familial and traditional Confucian ethics, and mainland China as bearing the legacies of a socialist society that values “collectivism, patriotism and nationalism” (Chan, 1996, p. 150). Chan also pointed out language differences; Hong Kong residents speak Cantonese whereas mainland China and Taiwan residents speak Mandarin. Overall, Chan concluded that the three regions are distinct enough to have different TV products of their own, and yet similar enough without cultural barriers for TV import and export. Chan further argued that Greater China should be treated as a geocultural region rather than a single geolinguistic region. Jirik (2004) further argued that Chinese geocultural commonality and geolinguistic diversity underline the centripetal and centrifugal dynamics that may lead to different homogenization or heterogenization outcomes.

Indeed, the progression of Chan’s research in 1996 and 2009 suggests a historical shift that is relevant for the discussion of Chinese boundaries and the timeline of Chinese Internet diffusion. In 1996, Chan described Hong Kong as “a regional broadcasting centre” and “political periphery as cultural centre”, saying that Taiwan had “awakened to new opportunities” and China had a “cultural deficit”. In 2009, Chan began his chapter with a discussion on “the centrality of the China market” because of its market size and growing TV market potential in the large, mostly coastal, affluent cities. He then concluded that the West and

Hong Kong remain a common source of inspiration for China's TV to create a hybrid culture that leads to some level of cultural convergence, though not to a total homogenization of Chinese TV culture. Thus, for the TV systems from 1996 to 2009 (also noting the fact that Hong Kong was returned to China politically in 1997), the growing markets and institutions of mainland China are expected to have substantial impacts on the boundaries, the centre of gravity, and the cultural-political norms for each region. It is also instructive to note that the Internet diffusion process among Chinese-language users also happened during the same period of time and followed a similar pattern of diffusion from Hong Kong, Taiwan and overseas Chinese cultures to mainland China (Qiu, 2005).

2.2.4 Chinese web studies. Little research directly addresses Chinese boundaries on the Web, but some research has addressed how and why Chinese websites are clustered, with findings that bear upon the boundary dynamics within the Chinese geocultural region.

Based on inter-links and content data among Chinese-language blog websites, Etling, Kelly & Faris (2009) visualized how these sites are interconnected, using network analysis and visualization methods similar to the ones they used to map Iranian and Arabic blogospheres. Parallel to their findings regarding Iran's Persian blogosphere, Etling, Kelly & Faris (2009) found a wide range of opinions and diversity in the categories of Chinese blog sites (culture, Hong Kong/Taiwan, patriotism, business, etc.). With further qualitative content analysis, for these categories they found different overall preferences in their word choice and external out-links (e.g. linking to YouTube video versus Sina's video, or linking to blocked websites), suggesting two sides of the "Great Firewall". They conclude by identifying five "attentive spaces" roughly corresponding to five clusters/zones in the network map: on one side, two clusters of "Pro-state" and "Business" bloggers, and on the other, two clusters of "Overseas" bloggers (including Hong Kong and Taiwan) and "Culture". Situated between the three clusters of "Pro-state", "Overseas" and "Culture" (and thus at

the centre of the network map) is the remaining cluster, which they call the “critical discourse” cluster, which is at the intersection of the two sides (albeit more on the “blocked” side of the Great Firewall). Their choice of word for the centre cluster reflects their research motivation as searching for democratic potentials. Their study on the Chinese blogosphere thus shows the intricate role of Hong Kong, Taiwan and other overseas Chinese websites in interacting with the mainland Chinese Web on the other side of the Great Firewall.

Taneja & Wu (2013), based on a sample of the aggregated traffic data of the world’s top 1000 most visited websites, visualized the Chinese-language cluster of users in comparison with 17 other language-bound clusters. Unlike Etling, Kelly & Faris’ (2009) approach and findings, Taneja & Wu argue for a single “Chinese” cluster as an instance of a “culturally-defined market”; that is, as a whole, not isolated from the world because of bridging websites such as alibaba.com (a China-based B2B website), Chinese Wikipedia and several Hong Kong and Taiwanese websites. Taneja & Wu further argue that the Great Firewall does little to isolate mainland Chinese users from the world.

To reconcile the apparently contradictory findings between Etling, Kelly & Faris (2009) and Taneja & Wu (2013), Table 2-1 compares the identified Chinese clusters by putting them into three categories: Mainland, bridging, and other clusters. It becomes clear that Etling, Kelly & Faris focus on the dynamics *within* the Chinese blogosphere, while Taneja & Wu focus on the dynamics *among* different language clusters (e.g. English, Japanese, English, etc.). Thus, what bridges the different clusters varies: For Etling, Kelly & Faris, a critical discourse cluster mediates among pro-state and business bloggers on one side and, on the second and third sides, among cultures and overseas regions. For the other authors, Chinese Wikipedia, several Hong Kong/Taiwan websites and Alibaba.com bridge the mainland Chinese cluster to the world. Especially noteworthy is the role of Hong Kong and Taiwan in both typologies. They are considered by Etling, Kelly & Faris to be part of the overseas category whose

connection to the mainland cluster requires a bridge in the middle. In Taneja & Wu’s work, however, they become one of the main bridges and connect the mainland to the world. Given the fact that Hong Kong and Taiwan are outside of Beijing’s Internet censorship and filtering regime, it is better to synthesize the two ideas: the regime *does* have an impact on the overall network typology, especially within the Chinese-language online space, but it does not isolate mainland China completely, since intermediary bridges do exist.

Table 2-1

Chinese clusters: previous research

Findings	Mainland Clusters	Bridging Clusters	Other Clusters
EKF*	Pro-state bloggers Business bloggers	Critical discourse	Culture bloggers Overseas bloggers (including Hong Kong and Taiwan)
TW**	Mostly allowed (unblocked) websites	Chinese Wikipedia Several Hong Kong and Taiwan websites Alibaba.com	Other language clusters as part of the global dataset

* Etling, Kelly & Faris (2009)

**Taneja & Wu (2013)

Furthermore, if one disaggregates the findings by differentiating business data points from cultural-political ones, both findings can be consistently interpreted as the desired outcome of the filtering and censorship regime that targets cultural-political content. The regime filters cultural-political interactions while allowing business exchanges. Taneja & Wu’s (2013) findings should thus be interpreted as proof for selective isolation: There is no isolation for mainland China’s business connections to the world, with the major example of the mostly English-written B2B website Alibaba.com. However, there are apparent signs of isolation for mainland China’s cultural-political connections to the outside world,

as can be seen when we eliminate in Table 2-1 the Hong Kong and Taiwan data points along with Chinese Wikipedia. Thus, while both studies provide valuable insights on the clustering outcome of the Chinese Web, researchers should be cautious about finding patterns based on aggregated data and should seek a more nuanced understanding through data disaggregation.

Researchers also need to be critical with regard to the concepts and substance regarding Chinese geolinguistic regions, Chinese geocultural regions, or what Taneja & Wu (2013) describe as a “Chinese culturally-defined market”. The term “Chinese culturally-defined market” is as problematic as the term “Greater China” for research: here it is worthwhile to turn again to global TV scholar J. M. Chan’s (1996) advice to treat Greater China as a larger geocultural region, while considering Hong Kong and Taiwan as individual regions apart from mainland China. Further, the political and ideological barriers among these regions, which were central to Chan’s (1996, 2009) discussion, were largely missing in Taneja & Wu’s (2013). From this perspective, Taneja & Wu (2013) might have aggregated the Hong Kong and Taiwan data points into those of mainland China too hastily, thereby conflating the findings as evidence for a Chinese culturally-defined market that is unaffected by Beijing’s Internet filtering and censorship regime. In particular, it is undeniable, even for Taneja & Wu (2013), that the Great Firewall has kept Chinese Wikipedia from occupying the central network position that other language versions of Wikipedia have achieved in their respective language clusters.

A synthesis of these perspectives would be to see the filtering and censorship regime as a mechanism of boundary and system maintenance. The censorship/filtering regime can be seen as maintaining a Chinese national information system within the worldwide environment of the Web, thus producing a specific kind of cultural-political boundary. Thus we must ask, not so much whether the regime isolates mainland Chinese users completely from the world, but rather whether it creates or reinforces cultural-political

boundaries between mainland Chinese users and the rest of Chinese-speaking users in the world. It is likely that what keeps the mainland Chinese cluster culturally and politically connected to the world are the bridging websites of Chinese Wikipedia and several Hong Kong and Taiwanese websites that have been the targets of Beijing's Internet filtering. As mentioned in Chapter 1, Wikipedia was expected by its founder to embody the “the possibility of spreading the idea of information freedom and democracy”, a possibility that is among the main targets of Beijing’s censorship/filtering regime. Researchers thus require a clearer framework for understanding the boundary dynamics on the Chinese Web, one that can account for both the censorship/filtering regime and geolinguistic differences.

2.3 Boundaries of media-language systems

As languages and media technologies emerge as significant factors in shaping national communicative spaces and boundaries, this chapter’s final section will review the recent development of information and language processing and discuss how they can be theorized as part of the historical process of mediatization of human societies at a larger scale. In communication studies, medium theory, or the mediatization approach, has its long-term perspective on the relationship between media evolution and cultural change (Hepp, 2013). The scalability of information and communicative spaces has been shown to depend on the advancement of writing systems for language processing. Encyclopaedias can be seen as an instance of media becoming more global with larger scales and writing systems, though they also face the realities of linguistic and national differences.

2.3.1 From manuscripts, print and telegrams to digital texts. The following sections summarize what medium theorists have posited with regard to text technologies, focusing on its scalability across geography in relation to languages.

Considering writing itself as a media technology, the intellectual tradition of medium theory has accumulated arguments about the alphabet effects (Innis,

2007; Levinson, 1997; Logan, 1986, 2000; McLuhan, 1964; McLuhan & Logan, 1977; Meyrowitz, 2010; W. J. Ong, 1982, 1992; Postman, 2006). The underlying argument about alphabet effects (also known as the alphabetic literacy theory) is that alphabetic scripts have contributed to human cognitive skills such as abstract science, codified law, deductive logic, and others (Logan, 2007). However, the alphabet effect theory is rightly criticized by Grosswiler (2004) for having an ethnocentric bias (i.e. that Western alphabets are superior to non-alphabetic writing systems). In particular, the theory's understanding of Chinese writing systems, including the argument that Chinese characters are primitive or inferior to alphabets, has been challenged by various Chinese studies and media studies scholars (DeFrancis, 1989; Grosswiler, 2004; Liao, 2009a). Nevertheless, medium theory has its merits in showing the historical significance of writing systems for media and cultural change, which proves useful for this thesis's focus on media-language evolution and cultural-political change.

Manuscripts, literally meaning “handwritten”, were the earliest forms of written communication. Major language scripts included Latin, Sanskrit, Arabic and Chinese scripts, with historical examples of Catholic Europe, the Muslim Middle East, and Confucian East Asia (Howland, 1996; Innis, 2007; McLuhan, 1964). Still, processing manuscripts (copying, transporting and organizing) was labour-intensive and limited in scale and scope when compared to modern text technologies. Thus, their use was confined to elites for various administrative, literary, scholarly or ritualistic functions. Sharing a language script often entailed a sense of belonging to the same civilization, religion or cultural sphere. For example, the very term “civilization” (wénmíng 文明) in both Chinese and Japanese starts with the Chinese character “written texts” (wén 文); hence the notion that sharing the same written forms of language (tóngwén 同文) has been associated with the notion of sharing the same civilization (Howland, 1996). Past East Asian scholars, mandarins and diplomats were able to participate in “brush talks”, or voice-less “conversations” written in the literary Chinese

language for various purposes, including the exchange of poetry and civilized sociability. Thus, while geographical reach could be expansive for manuscript-based civilizations, social participation was historically confined to elite circles with clear social boundaries.

Modern print technologies overcame the barriers of production to make efficient and consistent copies using machines. As agents of change (Eisenstein, 1979), their development also coincided with the European Enlightenment (Darnton, 1979; Israel, 2001; Roche, 2006), print capitalism and nation-states (B. Anderson, 1983; Schlesinger, 2001). The cultural implications of sharing a language script were also shifted. First, during the early period of Enlightenment, new print languages such as French, German, Dutch and English began to challenge the monopoly of the Latin language, which had served as the basis for long-distance communication among the intellectual community often referred to as the “Republic of Letters” (Israel, 2001). Other European vernacular languages also benefited from the modular and malleable movable type technologies, repurposing Latin alphabet-based print technologies. Second, mass production encouraged consumption of print products, ranging from newspapers to encyclopaedias. Different “monoglot mass reading publics” for vernacular markets (B. Anderson, 1983, p. 45) thus fostered various standard languages for particular nation-states and provided the basis for mass literacy programs that have since expanded the scope and scale of participation in social and political communications (Deutsch, 1966b; Gellner, 1983; Schlesinger, 2001).

Later, telegraphy further overcame the problem of physically transmitting texts, using wired or wireless long-distance connections. To reduce cost, human-readable messages are often reduced and then encoded using codebooks. For example, a commercial code book named “Unicode: The Universal Telegraphic Phrase-book” (not to be confused with digital Unicode standard) used many Latin phrases as shorthand messages for communication across various languages (*“Unicode”. The Universal Telegraphic Phrase-Book. A code of cypher*

words for commercial, domestic and familiar phrases in ordinary use in inland and foreign telegrams., 1886). Because of its speed and world coverage, the use of telegrams has been associated with contributing to Western colonization (Baark, 1997), fostering multinational corporations (Carey, 1988; Winseck, 2007), centralizing diplomatic efforts (Nickles, 2003), remediating domestic and international political communication (Lubrano, 1997; Zhou, 2005) and creating detached, removed, impersonal news narratives (Carey, 1988). Although telegraphy alone is too costly to maintain communal solidarity for a large number of people, it had been used to complement other media such as newspapers and news agencies (Blondheim, 1994; Carey, 1988).

Since the late twentieth century, computers and networks have increased both storage and transmission capacity. As will be further detailed later, the development and application of natural language processing (NLP), information systems (IS), and computer-mediated communication (CMC) have made information and communicative spaces more scalable and malleable. Providing an overview, print summarizes a comparison of these text technologies, ranging from (1) labour-intensive, transportation-limited and socially confined manuscripts, (2) mass-produced and -consumed modern print, (3) fast but expensive telegrams and finally to (4) all-inclusive digital texts.

Table 2-2

Types of texts and technologies: a comparison

Features	Manuscript	Modern Print	Telegram	Digital Text
Technological innovation	Writing	Movable types	Code books	NLP, IS and CMC technologies
Main writing systems	Latin, Arabic, and Chinese characters (Hanzi)	Alphabetic scripts	Commercial, diplomatic, military, etc. codes	All-inclusive Unicode
Geolinguistic reach	Catholic Europe, Islamic Middle East, and Confucian East Asia	Nation-states and Republic of letters	Globalization, nation-states, and global corporations	State-less nations, nation-states, and global corporations
Communal solidarity	Scholar-clergy or scholar-gentry class	Fellow readers or citizens	Low	Indefinite
Overall target user size	Small (literati)	Mass audience	Often complementary to other media	Potentially everyone in the world

The review above indicates how the scope of information and communicative spaces may be bound or divided by language factors. Table 2-2 shows how the major languages or coding standards (the second row) correspond to different geolinguistic reach (the third row), communal solidarity (the fourth row) and overall target user size (the final row). The correspondence suggests that media technologies have conditioned the support of languages, groups of users and major institutions of the time. For instance, modern print empowered national vernaculars (mostly those written in alphabetic scripts), enabled readership of a certain size (mostly created by mass literacy programs) and fostered national consciousness. The next section will focus specifically on the features of digital writing systems. Table 2-2 shows that these systems have evolved to be more inclusive and universal by default, but their impact on communal solidarity remains indefinite.

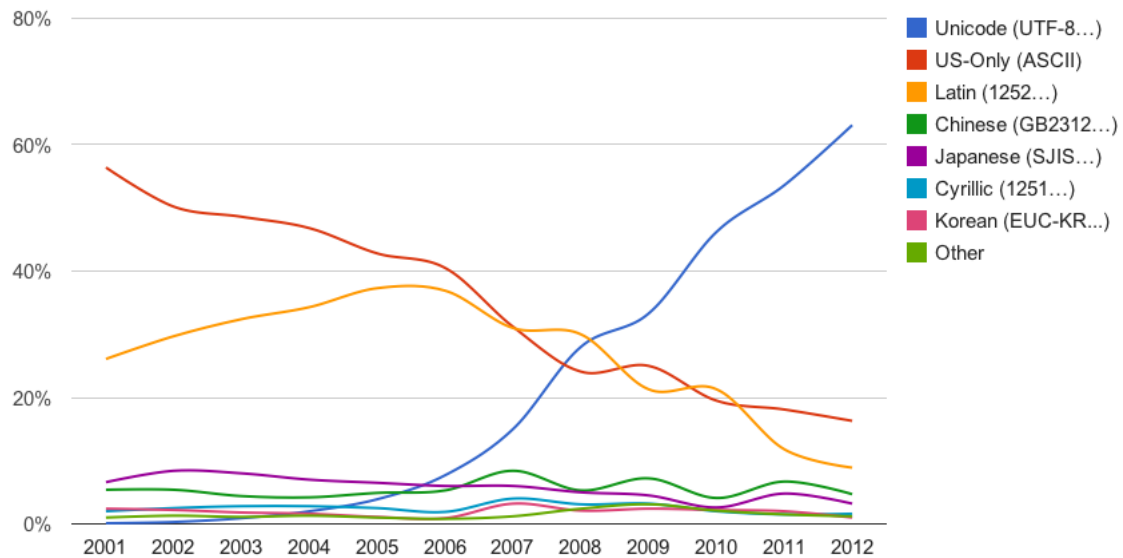
2.3.2 The role of digital writing systems. After presenting a historical account of major media-language systems in human history based on previous work by medium theorists, this section will first survey relevant technical developments in language and locale support for digital networked environments, and then it will discuss the implications of digital text-processing powers for national boundary dynamics.

"Will the Internet always speak English?" This was the question asked by an American linguist in 2002, recognizing the fact that the "Internet was basically an American development" and thus spreading faster across the English-speaking worlds, generating nearly 80 percent of Internet traffic (Nunberg, 2002). Nunberg also documented the fear, voiced by the French President and one Russian CEO, that the Internet's development would eventually lead to linguistic and cultural uniformity. Even the idea of an "open" Web was deemed to be "the ultimate act of intellectual colonialism" because, according to one author, it was not yet ready for Russian language and content (Specter, 1996). These responses reflected some of the early concerns surrounding the Internet as a conduit for only certain languages or cultures.

Since these early pronouncements, the linguistic infrastructure has advanced from monolingual to multilingual, as evidenced by the emergence and technologies of natural language processing, information systems and computer-mediated communication. In particular, the industry standard of Unicode is arguably the most evident indicator of such a multilingual shift. First, natural language processing, a subfield of artificial intelligence, computer science and linguistics that focuses on enabling computers to process human "natural" languages (as opposed to computer "programming" languages that instruct computing operations), have benefited from the standard of Unicode, including non-Western language scripts such as Chinese and Arabic (Habash, 2010; Wong, Li, & Xu, 2009). Second, in relation to library and information science, information systems (a subfield of management science that concerns building

systems for operations, management and decision-making) has been using Unicode standards to integrate and manage global and multilingual data for various information systems such as databases and browsers (Chebotarev & Ingersoll, 2004; Singh, Lehal, Sengupta, Sharma, & Goyal, 2011). Third, computer-mediated communication, a subfield of media studies that focuses on communication in digital networked environments, has shown the enabling role of Unicode for various non-Anglophone users and linguistic minorities such as Kurdish (Sheyholislami, 2011), Greek (Koutsogiannis & Mitsikopoulou, 2006), Hebrew (John, 2010), Chinese (Keniston, 2001; Liao, 2009b) and others.

The Unicode standard marked a milestone for multilingual support at the infrastructural level of digital texts, changing the once American English-only infrastructure. Recognizing the issue of multilingual support as early as the late 1990s, the Internet Engineering Task Force (IETF), the engineering community behind various Internet architecture and standards, declared that the "Internet is international" and thus it is "an absolute requirement to interchange data in a multiplicity of languages, which in turn utilize a bewildering number of characters" (Alvestrand, 1998). Coordinating the efforts among major computing and Internet companies, the Unicode Consortium has developed the Unicode Standard for the universal encoding of the world languages (Unicode Inc, 2011b, 2011c). Thanks to consortium and various experts around the globe, its wide adoption is evident in Google's statistics: texts on the Web that have adopted Unicode had reached 50% of the total Web in 2010 and 60% in 2012 (M. Davis, 2010, 2012), as shown in Figure 2-1. A clear multilingual shift thus suggests that the Web can now support a variety of languages, a progress that exceeded the IETF's original expectation of at least 50 years for widespread adoption.



Note: The decline in non-Unicode encoding standards does not reflect a decline in language usage.

Figure 2-1 Growth of Unicode on the Web (M. Davis, 2012)

In addition to language-script support, the Unicode Consortium also maintains a repository of "locales" which stores sets of parameters that define the region and language settings of users (Unicode Inc, 2011a). Locales thus represent different cultural preferences designed into digital systems. In terms of the institutions supporting foundation of language and locale, the software and Internet industry has contributed to internationalization (i18n) and Localization (L10n) of information systems (DePalma, 2002; Hussain & Mohan, 2008; Kim, 2005; Leong, Liu, & Wu, 1998; Sorasak & Kosona, 2010; Yunker, 2002). Within this framework, technologies can be tailored to provide adequate language support for users.

To identify or select a locale (and thus language content) preferences, the Internet Engineering Task Force's (IETF) language tags are used (IANA, 2011; Phillips & Davis, 2009). A language tag, or language code, is composed of a main tag of a language identifier (e.g. "zh" for Chinese), with or without a subtag that further specifies preferred scripts and/or regions. Table 2-3 lists some Chinese language tags, which correspond to variants of Chinese languages using subtags such as "hans", "hant", "CN", "SG", "HK", "MO" and "TW".

Table 2-3

Examples of the IETF language tags for Chinese

Code	Subtags	Language
zh	None	Chinese
zh-hans	script	Chinese written in Simplified script
zh-hant	script	Chinese written in Traditional script
zh-CN	country code	Chinese used in the People's Republic of China (PRC)
zh-SG	country code	Chinese used in the Republic of Singapore
zh-HK	country code	Chinese used in Hong Kong
zh-TW	country code	Chinese used in Taiwan, or the Republic of China (ROC)

Language tags indicate the digital existence of (a) different writing systems and (b) information processing outcomes. Because of this, languages remain potent boundary markers. The practice of localization, including information-processing mechanisms in identifying specific groups of users and content, thus contributes to the mechanisms of information gatekeeping on the side of information systems (Barzilai-Nahon, 2008). On the side of users, digital literacy skills with proper language support are one of the prerequisites for having a voice online (Liao, 2011). Table 2-4 lists the digitized status of major writing systems in the world based on the reviewed literature, showing a rough correspondence between civilization language-scripts and modern language tags.

Table 2-4

Examples of major writing systems being digitized

Major writing systems	Geolinguistic regions	Non-Unicode Encoding*	Major IETF language tags	
Roman	English	US-Only (ASCII)	en-US	American English
	Latin World	Latin (1252...)	pt-BR	Brazilian Portuguese
			es-MX	Mexican Spanish
Cyrillic	Slavic World***	1251, etc.	ru-RU	Russian
Arabic	Arab World		ar-EG	Egyptian Arabic
Chinese	•Greater China	Chinese (GB2312)	zh-CN	Mainland Chinese
			zh-HK	Hong Kong Zhongwen
			zh-TW	Taiwan Mandarin
Japanese**	•East Asian cultural sphere	Japanese (SJIS ...)	ja-JP	Japanese
Korean**		Korean (EUC-KR...)	ko-KR	Korean
Brahmi-derived Indic		India		hi-IN
			gu-IN	Gujarati

*Listed in the figure of a Google official blog post (Davis, 2012)

**Listed in Bosworth's (2004) 26 world civilizations but not in Coulmas(2000)'s major five

***Not listed in Sinclair, Jacka, & Cunningham (1996), but added by the author of this thesis

Effectively, language tags on the Web provide finer specificity for the concept of geolinguistic regions as used in global TV research. For example, within the geolinguistic region of the Latin World, information systems can exploit more specific language tags that include country codes such as “pt-BR” or “es-MX”. Thus, a more precise sense of geolinguistic regions can be digitally specified. To account for such localization practice, I therefore argue that the concept of geolinguistic regions should incorporate at least two levels of meaning. The first level refers to the geolinguistic regions in global TV research that are similar among civilizations; the other refers to the particular differentiations specified by country codes in language tags. Thus, while Hong Kong and Taiwan can be discussed as part of the more general sense of one Chinese geolinguistic or geocultural region, both can also be examined as specific Chinese geolinguistic regions for their respective language tags of “zh-HK” and “zh-TW”.

To conclude, despite its promise to accommodate all languages (Alvestrand, 1998) and its potential for cultural convergence (Dwyer, 2009; Jenkins, 2006), the Internet may very well still be divided along language and country lines, as shown by the different writing systems, geolinguistic regions, language codes and character-encoding standards displayed in Table 2-4. For instance, as previous research on global TV programming has suggested, do Brazilian Portuguese and Mexican Spanish remain the regional centre of the Latin world on the Web? The same question can be asked for Egyptian Arabic's role in the Arab world, for the Hong Kong-Taiwan style in the Chinese-speaking world, and so on. Whether and how such language-dependent digital-text processing may contribute to communal solidarity remains under-researched.

2.3.3 Encyclopaedias: knowledge, language and boundaries. The above discussion provides a historical context where encyclopaedias are viewed as an instance of media-language systems that may overcome existing human boundaries. Consider the following quotes:

...[O]nce the Imperial Institute of France and the Royal Society of London begin to work together on a new encyclopaedia, it will take less than a year to achieve a lasting peace between France and England.

French philosopher Henri Saint-Simon (1810/1975, p. 105)

A common ideology based on this Permanent World Encyclopaedia is a possible means, to some it seems the only means, of dissolving human conflict into unity.

English science fiction writer H.G. Wells (1937)

The vision statement of Wikipedia is very simple: a world in which every human being can freely share in the sum of all knowledge. My thesis for you is that by combining international, interagency, private-public, strategic communication, together, in this 21st century, we can create the sum of all security.

NATO Supreme Allied Commander James Stavridis (2012)

These quotes, made in 1810, 1937, and 2012 respectively, share a common theme: encyclopaedias may dissolve that which has divided humanity. French philosopher Henri Saint-Simon proposed that then warring France and England could achieve peace if they collaborate on a new encyclopaedia. After the First World War, writer H.G. Wells speculated that building a “World Encyclopaedia” could lead to the creation of a common ideology and thus a new world organization. What materialised since was Wikipedia, which in 2012 inspired the NATO Supreme Allied Commander to propose a thesis of “the sum of all security” based on Wikipedia’s vision statement on sharing “the sum of all knowledge” freely.

Will global encyclopaedia projects such as Wikipedia ever dissolve boundaries and resolve conflicts? One must consider the tensions between universal knowledge and national education on the one hand and the communication systems that deliver such knowledge on the other hand. Since the Enlightenment, tensions have existed between the ideal of the Republic of Letters and the idea of nation-states. Encyclopaedias of the Enlightenment were expected, along with similar genres such as dictionaries and lexicons, to make knowledge “known to his countrymen in their mother tongue” (Israel, 2001, p. 198). Similar to the internationalization/localization we see today in the digital environment, universal knowledge needed to be “localized” in print vernacular languages. Here, the ideal of universal knowledge for everyone began to face the realities and boundaries of media-language systems. In the European context, the solution largely involved national mass literacy and education projects, often using standardized national languages. Boundaries were thus made, mostly between nation-states. Encyclopaedias, along with other print products such as newspapers and novels, were thus not immune to the boundary dynamics built inside media-languages systems. Since most modern wars are fought between nation-states (and more recent within nation-states in the form of civil wars), the

ideal of universal knowledge to prevent human conflicts seems to be undercut by the national boundaries of education and communication systems that distribute human knowledge. Thus, while encyclopaedias may embody the enlightenment ideal of freedom, civility and empowerment of knowledge (Darnton, 1979), the boundary dynamics of media-language systems remains central.

Encyclopaedias can thus be studied as a boundary-sensitive case of media-language systems. In particular, online encyclopaedias can be examined for the scalability of digital texts.

2.4 Summary and implications

Each of the disciplinary approaches that has been reviewed has its merits when considering how information and communication systems shape national boundaries, the central concern of this thesis. From systems theory, the notion of system boundaries provides abstract but generic concepts to identify two types of boundaries: different intensities of internal versus external interactions (Deutsch-type) and distinct structures and styles of systems (Luhmann-type). From communications theories of nationalism, the concept of communicative space captures the modern historical construction of social communication systems that use media as boundary markers to cap a culture under a political roof. From comparative media systems research, media globalization and regionalization can be studied in terms of the dynamics among different media, market and political systems. From informetrics and digital methods research, national web studies have the quantitative potential to avoid the conventional research pitfalls of “methodological nationalism”. Altogether, this review suggests that because communicative spaces, as a modern social construction, have been shaping and shaped by salient historical, linguistic and geographic factors, researchers studying the relationship between communicative boundaries and national boundaries should focus on the formation and scalability of communicative spaces.

Understanding the influence of Chinese national boundaries thus requires examining the modern formation of Chinese communicative space in relation to its political and writing systems. Different notions of systems being orderly or chaotic have been put forward since the 1980s, reflecting varying opinions regarding Chinese political systems and information systems (Beijing's censorship/filtering regime included). Different East Asian and Chinese writing systems indicate the historical complexity of a modern nationalism rooted in Chinese or Sinocentric civilization with Chinese characters as its medium. For the three main constituent regions of Greater China, both centripetal and centrifugal dynamics in media globalization and regionalization have been observed in transborder TV since the 1980s and on the Internet since the late 1990s. The concepts of Chinese geolinguistic or geocultural regions, while useful, should be more vigorously applied when interpreting the Chinese Web data. In particular, the factor of Beijing's censorship/filtering regime must be considered along with the geolinguistic differences across Chinese-speaking regions. Chinese national boundary maintenance is thus maintenance of several Chinese political, information, communication and even writing systems.

To account for the latest development of information and language processing in the scalability of information and communicative spaces, the final part of this chapter historicized the boundary dynamics and the reach of several media-language systems. Encyclopaedias can then be viewed as a leading indicator of mediatization and globalization and how these face the realities of linguistic and national differences. Hence, following how media-language systems such as online encyclopaedias process information, researchers can examine system boundaries as well as the limits of such scaling up effects. The next chapter will introduce a set of working concepts to analyse the boundary-marking dynamics of modern information and communication systems.

Chapter 3 Theoretical framework and methods used

This chapter proposes a theoretical framework and a set of methods to achieve the following aims: (1) they should substantiate the research findings about Internet connectivity and national boundaries; (2) they should address cultural-political questions in relation to online spaces; and (3) they should have analytical power to organize data and explain how information is processed on the Web.

Before going into details, I will first describe the three main concepts of the thesis by way of an overview. First, “web spheres” are the online spaces of a collection or selected parts of websites and their content and links, which should be empirically specified and analytically defined. Second, “processability” refers to how the Internet enables the sharing of cultural resources and communicative spaces, thereby creating boundaries of web spheres (that may also reintroduce existing offline boundaries), especially through geolinguistic factors. Third, I define “cultural thickening” as the intensified and/or routinized processes and patterns of communicative and symbolic ties¹. Together, I argue that these three main concepts—“web spheres”, “geolinguistic processability” and “cultural thickening”—are central to examining cultural-political interaction online. For the purpose of this thesis then, the cultural-political patterns that are produced by the two encyclopaedia websites can thus be analysed as prime examples of boundary-making dynamics across Chinese-speaking regions.

Two figures illustrate these the relationship between these concepts. Figure 3-1 shows the basic ideas behind web spheres: they filter information from the World Wide Web for their users, and processability shapes the way information is filtered and thus shapes user experiences and choices. These user

¹ “Cultural thickening” denotes how cultural patterns overcome, reinforce or ignore the boundaries of mediatized processes. My definition draws on mediatization theories (e.g. Couldry & Hepp, 2013; Löfgren, 1997, as discussed in Chapter 2 and Chapter 3).

activities can in turn also shape the web spheres, forming two-way interactions. Web spheres usually have at least three constituting factors—platform, language and geography—that shape its processability, resulting in varied configurations of web spheres such as the Twitter universe, the Arabic online world, and the like. For each major Chinese-speaking region, a geolinguistic web sphere exists, which can be identified by their respective geolinguistic codes: zh-CN for mainland China, zh-HK for Hong Kong and zh-TW for Taiwan. These web spheres are expected to be more or less congruent with their own media and state systems, as visualized in Figure 3-2. Both Chinese Wikipedia and Baidu Baike are expected to provide “cultural thickening” patterns within and/or across these geolinguistic web spheres. These “cultural thickening” patterns may overcome existing boundaries (such as pattern B shown in Figure 3-2) or reinforce them (such as pattern A).

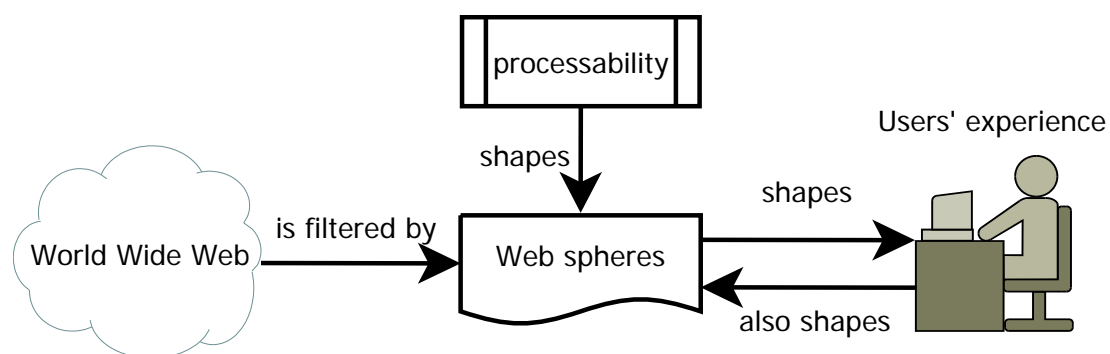


Figure 3-1. Processability shapes web spheres

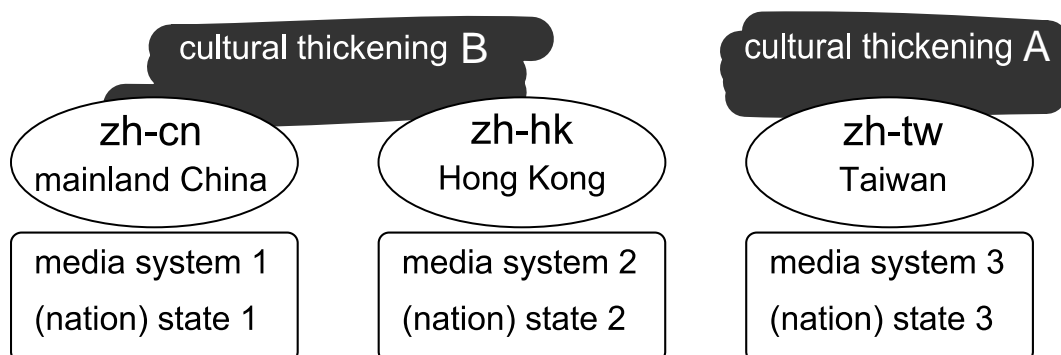


Figure 3-2. Cultural thickening patterns that overcome or reinforce boundaries

The graphs and paragraphs above provide an overview of the theoretical framework. This must be now elaborated in more detail.

3.1 Three main concepts

The three main concepts of web spheres, processability and cultural thickening come from different bodies of knowledge, some of which have already been briefly introduced in Chapter 2. First, “web spheres”, a term used by digital method researchers (e.g., Rogers, 2013), describes how information or cultural resources relate to online platforms. Second, I coin and use the term “(geolinguistic) processability” to describe how combinations of technologies - including natural language processing, information system localization and other computer mediated communication applications - process data. This explains how the factors of platform, language and geography are central to the construction of web spheres, potentially reintroducing national boundaries onto the Web. Third, as proposed by some comparative media researchers (e.g., Löfgren, 1997; Couldry & Hepp, 2013), and in order to avoid methodological nationalism (see Chapter 2), the notion of “cultural thickening” provides a way of understanding how content is patterned within and across boundaries. These concepts will be used to frame the findings gathered from Baidu Baike and Chinese Wikipedia.

3.1.1 Web spheres. In relation to online information or communication spaces, a major debate exists about whether these live up to the ideal notion of “public sphere”. This debate continues among various schools of thought about web spaces (e.g., Dean, 2003; Benkler, 2006; Castells, 2008), while various new terms have also been used describe different “spheres” of the Web. For example, the “blogosphere” can describe the (potentially rational) arguments among blog writers or simply the mechanical aggregation of all blogs and the links among them. Seeing the Web as a new forum or ‘great conversation’, some researchers have tried to find, map and/or foster online conversations that are pluralistic and/or deliberately democratic (Dahlberg, 2001; Papacharissi, 2002;

Cammaerts & Van Audenhove, 2005; Benkler, 2006; Burgess, Foth, & Klaebe, 2006; Fossum & Schlesinger, 2007; Geiger, 2009; Bruns, Burgess, Highfield, Kirchhoff, & Nicolai, 2011). Some use normative concepts such as “networked public spheres” (Benkler, 2006; Bruns, 2008; Bruns et al., 2011; Etling, Kelly, Faris, & Palfrey, 2010) or “online public spheres” (Dahlberg, 2001; Dahlgren, 2005). Yet others, like Foot & Schneider (2002; 2004, 2005), have proposed using ‘Websphere’ analysis to examine a set of thematically related websites.

Nevertheless, although some research has addressed how web spaces may be demarcated into different spheres, no consistent terminology has been arrived at to date.

Some use the terms “space” or “sphere” in a general sense to describe modern web spaces that serve information or communication purposes. For example, “Wikipedia space” may generally refer to the whole site (Biuk-Aghai, 2006) or more specifically only its user and discussion pages excluding the article pages (Derthick et al., 2011; Mac an Airchinnigh, 2012). Terms such as “twitter universe” or “Twitterverse” are used to refer to the community-at-large of a social media site (Fitton, 2012; Rubin, 2014; Thomases, 2009). Likewise, the term “Facebook nation” has been used (Dunay & Krueger, 2011; Lall, 2014; N. Lee, 2012).

Yet another view takes a comparative approach to web spaces, moving away from normative debates regarding public spheres to empirical analysis. Richard A. Rogers (2012, 2013), a researcher in digital methods, has noticed the emergence of sub-spaces of the Web. He analysed and compared different search results generated by search engines for the web sphere (for its general search function), the blogosphere (for its blog search), and the news sphere (for its news interface); thus the different cultural and political implications of search engines and how they placed website owners, bloggers, news and media organizations among their search engine results pages (SERPs). Rogers called this a “cross-spherical analysis” (2013, p. 206). Pushing the analytical use of the term “sphere”

in another dimension, Sanchez and Mesa (2011) compared regional variations of “Spanish-speaking spheres” and “French-speaking spheres”, using major Spanish-speaking and French-speaking countries as the basic unit of analysis. Thus, like Rogers’ cross-spherical analysis (2012, 2013), this research examines the variation in the shaping of web spaces and how it can be used for cultural-political analysis. A comparative approach therefore has several analytical benefits. First, it can bring down the unit of analysis from Google—the website as a whole—to its varied yet specific functions. Second, the comparative approach is useful for cultural-political analysis, as argued by Rogers: “[the] study of the politics of web space becomes cross-spherical” (R. Rogers, 2012, p. 15). Third, the approach demands a systematic analysis of “web spheres”, clarifying the unit of analysis, the scope of the research, and the like.

Therefore I define web spheres as online spaces consisting of a collection, or selected parts, of certain website(s) and their content (including hyperlinks). For this concept to be useful for research, social scientists must specify and justify the significance of communicative and/or information spaces as research objects. An example of this specification is shown in Table 3-1, where web spheres are specified according to three factors and combinations of factors: languages, regions and platforms.

Table 3-1

Different specifications of web spheres

Language only	Region only	Platform only	Language and platform	Region and platform
• Spanish-speaking web-sphere	• U.S. cyber domain	• Wikipedia	• Chinese Wikipedia • Spanish Wikipedia • Arabic Wikipedia	• Encyclopedia of Taiwan • The Encyclopedia of Hong Kong Virtual Communities
• Chinese-language Internet	• China's cyberspace	• Baidu Search • Yandex • Google Search • Google News • Google Books	• Chinese Weibo	• Baidu Japan • Yandex Russia • Google Hong Kong • Google News China • Hong Kong Weibo
• Arabic online world		• Twitter universe • Facebook nation • Craigslist	• Arabic Twitter	• Craigslist Hong Kong

Although this list is not exhaustive, Table 3-1 suggests that these specifications provide a useful aid to framing research questions. In addition, these perspectives highlight the fact that both researchers and designers can construct varied notions of web spheres as research objects (and, for designers, objects to be constructed). As indicated by the first and last two columns of Table 3-1, factors of language and region are central in the construction of web spheres because they often serve as the baselines for sorting users into identifiable groups.

The language perspective has been used to describe different language segments of the Web or the Internet, such as the “Chinese-language Internet” (e.g. Liao, 2011, 2013a; Maynard & Tian, 2004; Wu, 2007), the “Arabic online world” (Wahba, Taha, & England, 2013) or “Arabic-language Internet” (Lutz, 2009; Martin, & El-Toukhy, 2011; N.-C. Schneider & Gräf, 2011). Similarly, the region perspective has also been used to describe what are frequently country-specific segments, such as the “U.S. cyber domain” (R. Lai & Rahman, 2012), and “China’s cyberspace” (Xiao Qiang & Link, 2013). Nonetheless, languages and regions need to be specified in an analytical rather than a merely descriptive way.

To illustrate how languages and regions may shape web spheres, consider the example of Belgium, a country with several language areas, as shown by the Dutch language area (north), the French language area (south), the bilingual Brussels-Capital Region (near the centre) and the German language area (along Belgium’s eastern border), illustrated in Figure 3-3. The German language area is part of Wallonia, one of the three political regions of Belgium.

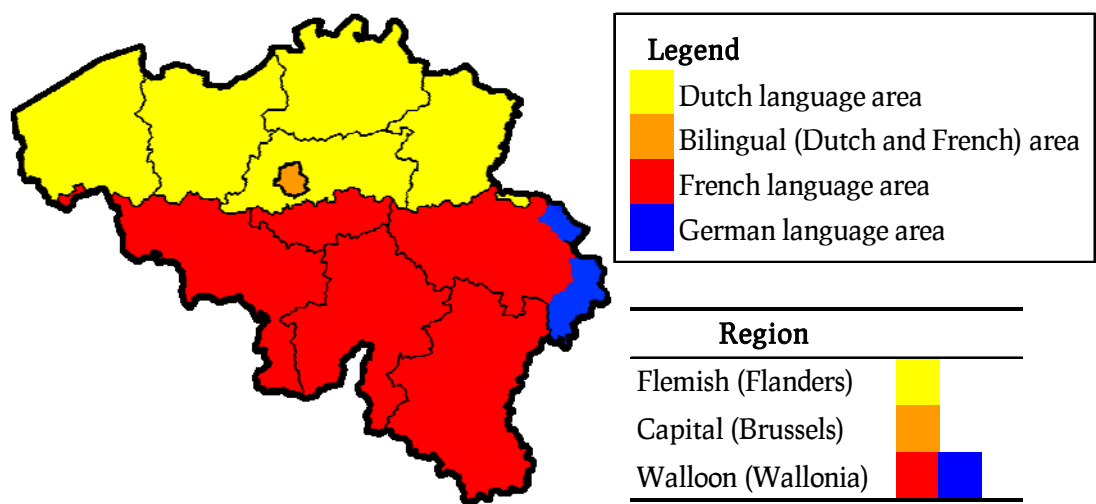


Figure 3-3. Map showing Belgium’s language areas (User:Stevenfruitsmaak, 2006)

For Belgium, major search engines offer different search interfaces (and thus different results), as listed in Table 3-2. As of May 2012, Google provided four language interfaces in its Belgian version (www.google.be): Dutch, French, German, and English. Table 3-2 also details the additional filtering options for each of the non-English interfaces: (1) user-specified language, (2) pages from Belgium and (3) translated foreign webpages. Similarly but with fewer options, the search engine Bing (be.bing.com) offers two language interfaces: Dutch and French. Two parameters, expressed in machine-readable geolinguistic identifiers, control such geolinguistic variations: The parameter (x) in Table 3-2 shows that Google preselected four language versions (nl, de, fr and en) for users in Belgium(BE), while Bing preselected two (nl-BE and fr-BE). The parameter (y) in Table 3-2 controls the further filtering options varied by languages and countries.

Table 3-2

Search engine variants: Belgium

Google Belgium			Bing		
Options	parameters ^a		Options	parameters ^b	
(original text)	(x)	(y)	(original text)	(x)	(y)
Dutch (nl)					
Internet (Internet)	nl		View All (Alles weergeven)	nl-BE	all
Pages written in Dutch (Pagina's geschreven in het Nederlands)	nl	lr=lang_nl	Only Dutch (Alleen Nederlands)	nl-BE	lf
Pages from Belgium (Pagina's uit België)	nl	cr=countryBE	Only websites in Belgium (Alleen websites in België)	nl-BE	rf
Translated foreign pages (Vertaalde buitenlandse pagina's)	nl	tbs=clir:1			
French (fr)					
The Web (Le Web)	fr		View All (Tout afficher)	fr-BE	all
Pages in French (Pages en français)	fr	lr=lang_fr	Only in French (Seulement en français)	fr-BE	lf
Country: Belgium (Pays : Belgique)	fr	cr=countryBE	Belgium only (Belgique seulement)	fr-BE	rf
Foreign language pages translated (Pages en langue étrangère traduites)	fr	tbs=clir:1			
German (de)					
The Web (Das Web)	de				
Pages in German (Seiten auf Deutsch)	de	lr=lang_f			
Pages from Belgium (Seiten aus Belgien)	de	cr=count			
Translated pages (Übersetzte Seiten)	de	tbs=clir:1			
English (en)					
The Web (Das Web)	en				
Pages from Belgium (Seiten auf Deutsch)	en	cr=count			

^aURL: [http://www.google.be/search?hl=\(x\)&\(y\)](http://www.google.be/search?hl=(x)&(y))^bURL: [http://be.bing.com/search?setmkt=\(x\)&filt=\(y\)](http://be.bing.com/search?setmkt=(x)&filt=(y))

By offering various geolinguistic interfaces, search engines construct different web spheres. These varied arrangements, including variants created by search engine platforms (e.g., Google, Bing, etc.) and geolinguistic identifiers (e.g., nl-BE, fr-BE, etc.), can be used in systematic analysis. They are also critical in thinking about the effects of Internet connectivity. While it is true that users, regardless of their actual locations, can use these Belgian variants, it is also true that users face an information environment that is already organized by certain pre-selected combinations of geolinguistic factors. Effectively, information systems are “geolinguistically” configured. Recall how web spheres filter the Web and shape users’ experience in Figure 3-1; such configurations indicate how they are designed to sort both information and users.

In contrast to the national search engine interfaces (e.g., www.google.be and be.bing.com), there is no national Wikipedia version for Belgium. Wikipedia arranges its users and information mostly through language versions. Thus, Belgian users are expected to use any language version of Wikipedia. In contrast to search engines such as Google and Bing, Wikipedia organizes its users and information less along the lines of national differences and more along the lines of language differences. According to various traffic reports provided by the Wikimedia foundation, users from Belgium contribute to viewing and editing activities mostly in its Dutch, French and English versions.² For Belgium, Google and Bing thus each offer a national space (further varied by geolinguistic options) while Wikipedia offers different language versions. Thus, researchers need specific research questions and theories in order to analyse the construction of web spheres as they vary by geolinguistic factors. In a stepwise fashion, researchers should first specify how geolinguistic factors shape and are shaped

² See <http://stats.wikimedia.org/wikimedia/squids/SquidReportPageEditsPerCountryBreakdown.htm#Belgium> and <http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm#Belgium>

by various platforms, as has been illustrated here for Belgium with Google, Bing and Wikipedia.

Concrete research questions can thus also be formulated in relation to web spheres: How do the web spheres of search engine variants reflect the national boundaries of Belgium? Are the Belgian versions of Dutch, French and German results (nl-BE, fr-BE and de-BE) similar to one another because of their commonality of sharing a nation-state? Or are they similar to the respective Dutch, French and German language versions just across the national boundaries of Belgium, such as Dutch in the Netherlands (nl-NL), German in Germany (de-DE), and French in France (fr-FR)? Specifically, for the two geolinguistically identified user groups of nl-BE and fr-BE, which correspond to the two major language areas (Dutch and French) in Belgium, do they have similar or different online activities? Do the two groups still share similar conversations online because of their bilingual capacity and practices, and/or because of the availability of bilingual content or machine-translation technologies? Do they share as much as they share with other French-speaking or Dutch-speaking people outside of Belgium? These questions can be asked by specifying the geolinguistic web spheres both within Belgium (nl-BE, fr-BE and de-BE) and outside Belgium (nl-NL, fr-FR and de-DE).

Further questions, including counterfactual ones, can be raised to examine the cultural-political constructedness of web spheres. For instance, what if search engine companies organized user interfaces first by languages and then by regions—and why don't they? What would be the cultural-political implications if search engine companies clustered French in Belgium (fr-BE), French in France (fr-FR) and French in Canada (fr-CA) on the same web page, similar to the way the Wikimedia foundation organizes the various language versions of Wikipedia? Likewise, what if search engine companies put all Chinese or Arabic variants on the same web page? In this way, researchers can recognize the constructedness of web spheres, and also avoid methodological

nationalism (which will be further discussed in section 3.1.3 on cultural thickening).

After discussing how web spheres can be constructed, we can now turn to platforms as a further factor. As mentioned earlier, the platform perspective has been used to refer to different segments of the Web such as Wikipedia space, Twitterverse, Facebook nation, and the like. Websites (such as Google.com) and their platforms (e.g. books.google.com, news.google.com) are basic sites to observe how web spheres are constructed.

As discussed earlier, a major research effort has been made by Rogers (2012, 2013) and other researchers in digital methods to conduct “cross-spherical” research that basically compares and contrasts different outcomes and features across different information or communication platforms such as the “blogosphere”, the “newssphere”, the “tagosphere” (sphere of ‘tags’) and the like. It is not wrong to define web spheres according to the different functions of the websites, yet it will also be necessary to take other features of platforms into consideration, such as their user base and market share, and also languages and their users. Since major user-generated content websites such as Wikipedia may have different editorial communities for different language versions (Liao, 2009b), for example, it makes sense to consider differences and contrasts across these language-differentiated web spaces (e.g., Sanchez & Mesa, 2011) before making generalizations based on platforms alone. Indeed, previous research on other websites has suggested that the factor of language has had a major effect on the patterns of interaction on the platform (e.g., Hale, 2012; Hong et al., 2011). Thus, depending on the research question and the scope and unit of analysis, researchers may consider perspectives that incorporate both language and platform factors, such as “Arabic Twitter”, “Chinese Weibo”, “Arabic Wikipedia”, and the like.

Likewise, websites and platforms can be either explicitly or implicitly regional. As listed in Table 3-1, websites such as Encyclopaedia of Taiwan, the

Encyclopaedia of Virtual Communities in Hong Kong, and Craigslist Hong Kong are examples of user-generated content that are region-focused. Different search platforms such as Google Hong Kong, Baidu Japan, Yandex Russia, and Google News China are yet another set of region-focused examples.

Altogether, researchers can better define their research objects by clearly specifying the language, region, platform and other factors. However, it would be a mistake to assume that different websites or their varied service platforms automatically construct distinctly separate bounded subspaces or web spheres - without overlapping in content, hyperlinks, and users. For example, what a Wikipedia article recommends as a relevant web page for a particular topic can be the same search result item recommended by Google Search. Another example is how a user may visit different websites and/or platforms (e.g., switching from general search to news search). Thus, researchers must be careful when they use a perspective that distinguishes between platforms in relation to research on web spheres. While it is true that search engines like Google or social media like Twitter and Facebook may gather information together into one place, this is not the same as the idea that Google, Twitter or Facebook bring different users together.

I suggest that researchers need to pay attention to various demarcations of web spheres in cases where a website or platform handles more than one region and/or language: One type of website prioritizes regions over languages. For example, major search engine companies such as Google and Bing first get specific on the level of nation-states or regions that have their own top-level domain names, such as “BE” for Belgium. Then they provide different language interfaces on the second level of languages within the region: nl-BE, fr-BE, de-BE, and the like. In contrast, a second type of website demarcates web spheres first by language and then by region. For instance, Wikipedia projects are first divided into different language-based communities, and then each of these communities

sets and enforces their own region-neutral or region-specific policies (Liao, 2009b).

I argue that the first type of website reintroduces national boundaries more so than the second one. As indicated in the case of search engine options for Belgium, national boundaries are first prioritized by the demarcation of the (often nation-state) top-level domain (TLD) names, and then reinforced by the selection of supported (often national/official) languages of that country/region. Such an arrangement reflects the ways in which web spheres are constructed to correspond to nation-state markets. We can call these TLD-delineated web spheres. In contrast, the second type encourages (by its very design) interaction across and within nation-state boundaries for different languages. For instance, within Chinese Wikipedia, in order to promote cross-regional interactions, specific Chinese-speaking regions are recognized as providing appropriate language interface and support to handle variations in language script and terminology (Liao, 2009b). Also, different editorial communities maintain Chinese Wikipedia, Cantonese Wikipedia and Hakka Wikipedia, which are separate projects despite the fact that they are all ethnic Chinese languages. This type of arrangement serves language platforms first (often justified and identified by the language codes of ISO-639 used in Wikipedia) and then their respective regional variations. We can call these language- or language-code-delineated web spheres. Hence, the TLD-delineated web spheres are expected to reintroduce national boundaries, whereas the language-delineated web spheres are expected to construct geolinguistic regions (recall the discussion of regions in section 2.1). In terms of their interface and information designs, both types of web spheres produce different demarcations.

Implications of and limitations for research. The discussion of the construction of web spheres from various perspectives can now be summarized in terms of the research strategies it suggests: First, the proposed concept of web spheres brings together a number of factors involved in their construction.

Rather than rushing into the debate by using concepts such as “networked public spheres” or “online public spheres” that are based on normative values, the concept of “web spheres” focuses first on examining what is empirically the case. The discussion so far has concentrated on three factors (platforms, languages and regions) that generate different perspectives on how web spheres can be constructed and which will be useful in research. Second, the notion of web spheres can help researchers to identify and justify the scope of the research and/or the unit of analysis. For example, in order to pin down the research object of “Chinese social media”, researchers may want to first specify whether the research focus is on the Chinese-language space or China’s political space (and thus raise questions, for example, as to whether Hong Kong is included). Also, researchers may want to delineate and justify how such a space can be analytically observed on the basis of “a collection, or selected parts, of certain website(s) and their content (hyperlinks included)” (which is my definition of web spheres). Researchers should also make their selection rationale based on user bases and/or market shares: For instance, to construct a web sphere that represents Hong Kong as a region, the selection of websites and their localized parts designed for Hong Kong users must be justified based on researchers’ assessment of its user bases and market shares. Moreover, while this discussion of web spheres has been limited to factors of platforms, languages and regions, future researchers can also add other factors. Since web spheres can be constructed differently, comparative analysis can thus be carried out on different web spheres, as Rogers (2012, 2013) and other researchers on digital methods have done using the notion of “cross-spherical analysis”.

Third, the notion of web spheres highlights the constructedness of online spaces. For instance, the comparative nature of “cross-spherical analysis” can lead to questions such as how this type of construction of spaces can integrate or fragment information or users. In addition, it is possible to gain additional insights because researchers can construct somewhat different notions of web

spheres against the actual web spheres that are designed and implemented (recall ‘counterfactuals’). Take the major search engines as an example—although their interface design is often first demarcated by nation-states and then by various language options within each state (as shown in the above example of Belgium), researchers can think about alternative constructions which lead to ideas about how search results and their social effects are predominantly determined first by the factor of languages, and then by regions. Indeed, the next sections will build on this account by examining how “geolinguistic processability” can produce different “cultural thickening” patterns.

3.1.2 Processability. In order to characterize the common features of web spheres, this section will further discuss the concept of “processability” introduced in section 2.3. As part of the historical development of human information and communication, the increased level and sophistication of processability is at the heart of the mediating power of the Internet, and holds the key to understanding how national boundaries are introduced and reintroduced on the Web.

The development of natural language processing, information systems and computer-mediated communication have a number of implications for social scientists. I propose the term “processability” to conceptualize the enabling conditions, impacts and design of the media-language system (which include related writing, information and communication systems). The capacity to process information by both technical artefacts and their users is not just limited to the development of the Internet. This section will compare and contrast media-language systems to highlight particular aspects of processability.

Sorting information and users. To illustrate how processability is tied to human writing and media systems, we can observe how information and users are sorted and categorized using symbols. Consider the index for modern search engines and of printed books. The sorting and ordering functions of modern search engines can be seen as an improved version of the alphabetic collation

system used by back-of-the-book indices in the print era. Just as back-of-the-book indices guide readers to relevant pages for a certain key phrase, a search engine can be seen as a super index that dynamically presents users with web links pointing to relevant web pages for a specific search query. In fact, many search engines offer keyword suggestions (also known as “autocomplete”) to recommend further search keyword suggestions based on users’ initial input of linguistic symbols. What remains constant are the ways in which users and artefacts sort information using writing systems (e.g., finding information alphabetically), and what is new for modern search engines is that such efforts of indexing and searching information can be facilitated by computing devices, and these are potentially not just limited to mere alphabetic writing systems. This means that researchers must examine how different kinds of processability sort information and, arguably even more importantly, users, in different ways.

Consider the following example of Chinese-language search keyword suggestions recommended by Google. Two different orthographic forms of the Chinese term for “Communist Party” (i.e., simplified: 共产党; traditional: 共產黨) produce two sets of keyword suggestions, as shown in Table 3-3. The differences between the two lists are of cultural-political significance. Users of different orthographic preferences (which correspond to different Chinese-speaking regions) can be expected to see different suggestions. Those who use a simplified Chinese query are expected to find more negative suggestions (e.g., “History Revisionism”, “Mafia”, “Evil Cult”, “lost the hearts of people” and even the Chinese slang for “bastards”) than those who use traditional Chinese in this case (e.g., “Manifesto”, “Flag”, “History”, “Song”, “Chapters”, “Resistance to Japanese”, etc.). Implicitly, both types of information and users are categorized and sorted according to the ways in which both Google and its users process Chinese-language information. The nature of the modern Web recommender system is expected to suggest the dominant practices of users who have recently submitted the same or similar queries. Battelle (2005) conceptualizes modern search

engines as “databases of intentions” because they collect search queries that represent users’ intentions. How search engines depend on human writing systems and practices further suggests that such databases are likely to be structured by these different writing systems as well as by users’ practices in relation to other writing systems (for example, how they use book indexes, or library catalogues). Thus, we can see that the processability of different Chinese orthographic forms (for example) reveals how the processability of the Web sorts both Chinese-language information and users. Such processability and its sorting are more complex than back-of-the-book indices in the print era.

Table 3-3

Different keyword suggestions for the Chinese query of “Communist Party”

Simplified Chinese Results			Traditional Chinese Results		
Suggestions	N	English Translation	Suggestions	N	English Translation
__篡政历史	89,300	History Revisionism	宣言	992,000	Manifesto
__黑社会	690,000	Mafia	__Wiki	220,000	Wiki
宣言	1,070,000	Manifesto	宣言_全文	955,000	Manifesto Fulltext
__倒台	179,000	Collapse	黨旗	450,000	Party Flag
__邪教	454,000	Evil Cult	歷史	30,700,000	History
不得民心	123,000	lost people's hearts	腐敗	1,900,000	Corruption
腐敗	1,220,000	Corruption	__歌	1,940,000	Song
__国民党	2,800,000	KMT	口號	1,630,000	Slogan
__王八蛋	56,200	“bastards”	黨章	1,910,000	Party constitution
__评论	9,900,000	Commentaries	抗日	2,320,000	Resistance to Japan



Processability thus is at the heart of the mediating power of the Internet.

Consider the following quotes:

The Internet is international.

With the international Internet follows an absolute requirement to interchange data in a multiplicity of languages, which in turn utilize a bewildering number of characters.

-Excerpts from "IETF Policy on Character Sets and Languages" (Alvestrand, 1998)

Seen from a Google's eye view, in fact, the Web is less like a piazza than a souk -- a jumble of separate spaces, each with its own isolated chatter. The search engines cruise the alleyways to listen in on all of these conversations, locate the people who are talking about the subject we're interested in, and tell us which of them has earned the most nods from the other confabulators in the room ...

-Quote from a linguist (Nunberg, 2003)

Search engines such as Google aggregate and process conversations on the Web. They must process information of a linguistic and social nature in order to bring the "isolated chatter" together. Before they can do that, human languages must be supported digitally for conversations to be processable. Search engines mediate the Internet by processing digital texts of various languages on the Web: basically matching strings of text symbols that represent digitally articulated voices across the world. Without this processability of the human articulation of messages, the organization of human affairs in relation to the Internet would be difficult.

People themselves also sort and order information, of course, which leads us to the importance of literacy (which, in turn, depends on writing systems). As reviewed in section 2.3, regardless of their technological forms, encyclopaedias, reference works, office filing systems, and modern computer collation systems

all rely on sorting rules that demand linguistic literacy from users as well as language support from technical systems. The alphabetic literacy theory would argue that the complex symbol system of Chinese languages constitutes a major challenge in the design of viable sorting and ordering devices for information processing. This might indeed be the case for print media. However, as more languages become processable in the digital networked environment, it is reasonable to assume that the function and capacity of sorting information and users will be agnostic to the choice of writing systems as long as it is technically supported by machines and fits with the practices of users of these machines. The processability in any information and communication system thus requires technical support on the machine side and literacy skills on the human side.

Geolinguistic processability. Just as various factors influence the construction of web spheres, so the notion of processability is also influenced by the main factors of language, region and platforms (most likely in that order). I use the term “geolinguistic processability” to refer to the geographic and linguistic conditions, impacts and design of the media-language system in question.

The existence of the geolinguistic processability of the Web is arguably most evident in the way language tags (or geolinguistic codes) are used on the Web. As discussed in Section 2.3, language tags are geolinguistic identifiers used to configure users’ devices so as to signal their geolinguistic preferences and/or default settings, and these signals form the bases for various kinds of information processes and services that are tailored for a “locale”. Geolinguistic processability has been instrumental in content and service delivery on the Web. For software to work for different geolinguistic groups of users, a certain level of geolinguistic processability must be achieved. Websites such as Google further allow users to switch from one geolinguistic code to another (Google Support, 2011), including among about 200 language interfaces and 200 countries (Google, 2011b). Sometimes users can take up the role of developers by building interfaces for

their own languages (Google, 2011a). With a do-it-yourself or serve-yourself approach towards localization, Wikipedia had 276 language versions as of 2011, while Facebook has just over 100 language interfaces (M. Anderson, 2011). These all exploit the geolinguistic codification of their interfaces, users and content. It is consequently likely that national boundaries will be reintroduced onto the Web through the ways in which geolinguistic processability operates.

For example, search engines provide different locale services for Belgium, as listed in Table 3-2, suggesting that geolinguistic identifiers play a dominant role in providing different search outcomes for different groups of users. In addition, Chinese geolinguistic identifiers of "zh-CN", "zh-SG", "zh-HK", and "zh-TW" signal variants of Chinese language to users in China, Hong Kong and Taiwan, respectively. While we can draw some parallels between these Chinese-speaking geolinguistic regions and other English-speaking ones such as "en-US", "en-UK", and "en-IN" (English used in US, UK and India, respectively), the codification and categorization of the Chinese-speaking regions are also more controversial and problematic, as discussed in Chapter 1. Hence, researchers must examine how geolinguistic processability contributes to the division or integration of these regions. The other immediate question is, do linguistic factors matter more than geographic factors?

I argue that the demarcation of web spheres proceeds: writing systems, geography, and then platforms. This is because, first, it makes more historical sense, as reviewed in section 2.3, to see the role of writing systems as essential to expanding human communication on a larger scale for more extensive and complex social activities and organizations. Second, as a practical matter for users, it is easier to learn a new platform than to learn a new foreign writing system. In fact, the number of writing systems a user can learn in his or her lifetime can be expected to be much lower than the number of platforms he or she can master. Thus, writing systems should take priority: users are more likely to be confined by the writing systems they are able to master than by the

platforms they can choose. Third, the technical properties of digital writing systems, as also discussed in section 2.3, suggest that information-sorting practices are more limited in bridging the gaps between different writing systems - as compared to bridging the gaps among geography or platforms. In sum, the demarcation of first writing systems, then geographical regions, and finally platforms - makes analytical sense.

It is also important to discuss how geolinguistic processability works on the side of users. The ways in which users can process information depends on their proficiency level of literacy and knowledge skills, which are often limited in terms of languages and geography. For instance, the use of Chinese-language search engines requires the basic skills of typing in Chinese characters and a background knowledge of Chinese-language keywords. Corresponding software and information parameters are set according to different “locale” settings that identify and differentiate between various Chinese-speaking regions by their specific language tags, such as zh-CN, zh-TW, zh-HK and zh-SG. In addition to these processing skills, which are more technical, the education, media and knowledge backgrounds of users from these regions are expected to influence the ways in which these users process cultural-political knowledge in relation to China and Chineseness. Thus, the notion of user geolinguistic processability should include not only the ways these users can or may process information externally on the websites, but also the ways they internalize their Web experiences as part of the formation of their education, media and knowledge backgrounds. Put briefly, user geolinguistic processability skills include both technical and cultural-political literacy.

The concept of geolinguistic processability thus reveals some further important features for research into web spheres and online spaces. Consider the Chinese language, a major language that now enjoys considerable technical support and a strong user base. The notion of a Chinese-language web sphere is expected to highlight the wide variety of design and use practices online because

of the diverse writing systems and literacy education backgrounds of its users. The notion of a “Chinese cultural sphere”, as discussed by Yang (2003), thus demand that researchers notice fundamental differences in language support and literacy backgrounds across different Chinese-speaking regions. As the geolinguistic variation of Chinese-language information has been codified and built into digital networked environments, various Chinese geolinguistic web spheres are constructed accordingly. Thus, researchers can see, or even visualize with data, the “Chinese cultural sphere” as the aggregation of several Chinese geolinguistic web spheres. By tracing how geolinguistic processability is designed into platforms and relates to the practices of users, researchers can examine how national boundaries may be reintroduced or overcome on the Web.

National boundaries versus general geolinguistic regions. As geolinguistic processability plays an important role in the reintroduction of national boundaries onto the Web, it is necessary to discuss its boundary-making effects in greater empirical and analytical detail. National boundaries can be reintroduced onto a website or platform in at least two ways, in addition to the usual considerations regarding the jurisdiction of the websites and/or website owners: One approach exploits country codes, language identifiers, and other information system measures to construct a nationally bounded web sphere. Another approach counts on managing the potential and incoming visitors or users in a way that is limited to a certain (national) group of users with similar (national) literacy backgrounds. These two approaches converge with the use of language tags to sort information and users, as previously discussed. The construction of web spheres therefore occurs through the arrangement of geolinguistic processability, as exemplified by the Google Search keyword suggestions shown in Table 3-3. Hence, it is very likely that the processability on the side of websites and that on the side of users are mutually influencing or mutually reinforcing. A geolinguistic design begets a geolinguistic profile of users.

Such mutually influencing dynamics of geolinguistic processability between websites and users have several implications for understanding national boundaries as cultural-political practices. This point can be elicited by way of a brief discussion on writing systems in relation to print and digital media. As summarized in section 2.3, writing (and thus reading) technologies have multiple social, cultural and political implications. For understand national boundaries, one needs to examine the "processability" of writing systems more generally: If the printing press and print capitalism foster some kind of national consciousness based on the routine sense of "fellow-readers" (B. Anderson, 1983), one can also speculate on the possibility of "fellow-users" as constructed and shaped by websites and their routines. Put into such a longer-term historical perspective, processability as a concept is not limited to the digital networked environment. Just as manuscripts and print products engendered a certain sense of "fellow-readers" - be they literati, members of the "Republic of letters", or a mass readership that can be demarcated by language-differentiated print markets – so too users of digital texts can read, write, search, filter, organize and even execute computer programs in a collaborative or collective fashion. Readers now become users. Researchers must therefore examine how increased processability reshapes human media-language systems in such a way that daily routines of national boundaries are reintroduced onto or they are transformed by the Web. Since new features of digital texts, notably hyperlinks and hashtags, have shaped the concepts of sociability and social affordance online (Wellman et al., 2003), it is important to examine how the processability of digital texts shapes social interactions: For instance, how is the sense of "fellow-users" made possible through sharing the same platform, geolinguistic identifier or writing system?

There are obvious tensions in shaping the sense of fellow users for certain languages such as Chinese and Arabic because these languages have regional variations (which can be identified by geolinguistic codes) despite the fact that these regions share a writing language. Consider the following quotes:

Social media is all about connections. ... There's no stronger connection than literally speaking the same language.

Quote from a digital marketing specialist (M. Anderson, 2011)

While the ISO 3166 directory of names of countries and territories has the TW code as Taiwan, Province of China, most major commercial Web sites such as Amazon.com, Nytimes.com, and Barnesandnoble.com took the time to remove the "Province of China" reference in their country and territory address forms, rather than to blindly use the ISO 3166 standard.

Quote from a readers' letter to Taipei Times (Chang, 2005)

The first quote made by a digital marketing specialist reminds us of the primary role of language, whereas the second quote alerts us to the fact that some country and/or region may have certain cultural and political urges to assert its identity by making sure the country code is implemented properly. In relation to the concept of “geolinguistic regions” from global TV researchers (e.g. Sinclair et al., 1996), there is a tension between a generalized and a more specific sense of “geolinguistic regions”: the more general sense of this refers to a shared spoken and/or speaking language such as the Chinese geolinguistic region, the Arabic geolinguistic region, and the like. In contrast, the more specific sense refers to those regions that can be identified and coded with added country codes, such as the Chinese geolinguistic regions of Hong Kong (zh-HK), Singapore (zh-SG), Taiwan (zh-TW), Macau (zh-MO), and mainland China (zh-CN). The former relies more on the strength of connection made possible by sharing a language, and the latter depends on the differentiation engendered by geolinguistic codes and practices. Both relate to the question of how geolinguistic processability reintroduces national boundaries. Researchers need to recognize the existence of such tensions and evaluate which sense is more important.

Before answering these research challenges, it will be helpful to characterize how information and communication spaces (or simply web spheres) are constructed differently from other kinds of media - such as print and TV. Like the printing press, web spheres can be bounded within different written language systems, and thus cross-lingual translation is still needed—although on the Web, machine and user-generated translation may have reduced barriers. Like TV, web spheres can be bounded within pockets of geographic regions. However, while TV is geographically bounded (often by national regulators) and major global TV networks rely on satellite to import and export programmes, web spheres and their traffic are not as geographically bounded or regulated. In sum, I argue that web spheres are bounded mostly by geolinguistic processability, but at the same time, they are also generally more porous than TV and print simply because of the nature of the Web. The Web has been theorized in terms of “network gatekeeping” (Barzilai-Nahon, 2008) or “relevance filtration and accreditation” (Benkler, 2006, p. 169). Thus, web spheres may be bounded by what they have processed, but they are potentially more porous because of normally unbounded visits and hyperlinking practices.

The next and final section, on the concept of “cultural thickening”, will add to and tie together the concepts of processability and web spheres.

3.1.3 Cultural thickening. Borrowed from mediatization and medium theories (Hepp, 2013; Hepp, Hjarvard, & Lundby, 2010), the concept of “cultural thickening” is useful for researchers in tackling the porous yet bounded nature of the Web. When integrated with the concepts of web spheres and processability, the concept can be applied to the comparative study of Baidu Baike and Chinese Wikipedia.

Identifying cultural patterns while avoiding methodological nationalism.

The main benefit of using the concept of “cultural thickening” is that it provides a unique perspective in identifying cultural patterns in a way that avoids methodological nationalism. Responding to the call to “internationalize” media

studies (Thussu, 2009), some researchers have proposed a transcultural approach (Couldry, 2012; Hepp, 2009; Hepp & Couldry, 2009) that builds upon the concept of “cultural thickening” (Löfgren, 2001). Consciously against a reductive national framework for research, they try to ground the research, for national or transnational media research alike, on a “translocal” understanding of media cultures. Thus, national and transnational activities are both observed as translocal ones. Instead of seeing media culture practices as always and immediately territorial, they see media cultures as a “thickening of translocal processes of meaning articulation that themselves are more (or less) locally specific” (Couldry & Hepp, 2013, p. 254). The translocal approach can be more helpful than the conventional “territorial” view if media cultures are seen as “the ‘sum’ of the classificatory systems and discursive formations on which the production of meaning in everyday practices draws” (Couldry & Hepp, 2013, p. 256). Thus, the actual thickening of culture may or may not correspond directly with territorial borders. Indeed, this approach features a clear shift of research focus from territorial borders to “translocal” cultural patterns that can later be shown to be national, transnational or neither.

Although the concept of cultural thickening remains untested, at least to the author’s knowledge, for any research objects on the Web, it is helpful for researchers in analysing cultural patterns on the Web for several reasons. The following paragraphs will show its usefulness by applying it step-wise, along with my previous discussion on the concepts of web spheres and geolinguistic processability, to the comparative study of Baidu Baike and Chinese Wikipedia.

The first step is to delineate and then analyse “cultural patterns”, which according to Couldry & Hepp (2013) include “patterns of thinking,” “patterns of discourse,” and “patterns of practice” or “doing” in the tradition of social constructivism. For the Web, plenty of activities on the level of practice or doing can be observed, including copying, deleting, linking and filtering information. The goal in this first step is to identify, in specific cultural contexts, typical

“ways” of action, discourse or thought. For the current research question about the cultural-political boundaries across Chinese-speaking regions, researchers need to identify specific forms of cultural patterns that are articulated through reading, writing, or interacting with user-generated encyclopaedias and their users.

The second step is to compare the patterns that are found beyond mere national or transnational comparisons, and geolinguistic factors, example, can be examined in such a comparative way (as previously discussed in the case of Belgium; see the previous section on Web spheres.). To avoid “understanding each cultural pattern as an exclusive *expression* of a national media culture” (Couldry & Hepp, 2013, p. 257), researchers should avoid confining or aggregating data nationally from the start of the research process. Instead, they should first organise data in terms of socio-cultural entities, then categorize different cultural patterns, and finally organize the findings comparatively and accordingly. Such a comparative approach is useful because it allows researchers to compare and analyse cultural patterns beyond the conventional and already reductive national frame (e.g., Belgium’s or China’s national Internet).

Researchers can instead identify relevant socio-cultural entities on the Web (e.g., geolinguistic identifiers such as nl-BE, fr-BE, zh-CN, zh-HK, zh-TW, etc.) and then observe whether and how their findings can be justifiably reduced by national or transnational frames - and why. For example, do cultural patterns amount to nationally bounded thickening? If so, how they are bounded and why?

The third and final step is to make sure that researchers take a critical approach to the construction of the research object and the construction of the research itself. Researchers must expose various “centring” aspects of cultural thickening and the related underlying questions of power (Couldry & Hepp, 2013, p. 258) by taking a “de-essentialized” approach. By not interpreting findings as being essentially national or transnational, researchers need to conduct a cultural-political analysis of the cultural thickening patterns that are found so as

to see how these patterns make certain ideas, information sources, and patterns more prominent and more central. For the purposes of this thesis, the notion of “centring” has an additional potential for the analysis of Chinese cultural politics: the term “China” in the Chinese language literally means “the centre state” (or conventionally “the middle kingdom”). The notion of “centre” thus has an additional ideological bearing in modern China as both political parties—the Chinese Nationalist Party and the Chinese Communist Party—have popularized a Chinese-language term for the “centre” (zhōngyāng 中央) to describe the political centre of the party (dǎng zhōngyāng 黨中央) or the state (zhōngyāng zhèngfǔ 中央政府). Thus, the notion of centring highlights the construction of cultural-political centrality through cultural thickening.

Building specific cultural patterns by constructing web spheres. In sum, “cultural thickening” can be defined, for the purposes of this research, as the intensified and/or routinized processes and patterns of communicative and symbolic ties. The following paragraphs will now explain how this concept is useful for analysing web spheres. First, as already illustrated (at the bottom of Figure 3-2), each geolinguistic web sphere corresponds to respective geolinguistic regions. Ideally, each geolinguistic sphere should represent the aggregated outcome of the online activities for users and information originating from each region. Each of the web spheres is thus formally specified by the combination of its linguistic features (assumed to be Chinese language: zh) and its regional features (signified by each of their respective country codes: CN, HK, TW, etc.). These geolinguistic web spheres are expected to be more or less congruent with their own media systems as discussed in Chapter 1, including varied features of their writing and political systems. Such congruence is shown by the bottom two rows of Figure 3-2. In other words, Chinese-speaking regions should be seen as reintroducing existing boundaries onto the Web.

Second, both Chinese Wikipedia and Baidu Baike, as major platforms of Chinese-language user-generated encyclopaedias, have platform-specific web

spheres. These may potentially engage users and information from all Chinese geolinguistic web spheres and beyond. Both encyclopaedia platforms are bounded in the sense that both filter information that merits the status of being part of the sum of human knowledge as documented in the form of encyclopaedia articles. However, both are also porous in the sense that they are open to new users and new input of information that is deemed admissible by each of their editorial communities. Because both sites exhibit the main features of web spheres in being bounded and porous at the same time, the web spheres constructed by each of the encyclopaedia websites should have certain cultural thickening effects that bring certain information and users together in such a way that reflects on their respective tendency of “centring”. For instance, Figure 3-2 shows an imaginary scenario that clearly indicates that the pattern of “cultural thickening A” focuses and centres on the geolinguistic web sphere of zh-TW (and thus its respective media and political system), while “cultural thickening B” seems to not only overcome existing boundaries that are reintroduced onto the Web but also to position its thickening patterns most thickly over the geolinguistic web sphere of zh-HK. Hence, by analysing the intensified and/or routinized cultural thickening patterns exhibited by the two websites comparatively, researchers can assess whether and how the existing boundaries between the geolinguistic regions are overcome, reinforced or ignored by the two cultural thickening patterns produced by the two encyclopaedia websites.

The thesis thus conceptualizes the two websites as different web spheres constructed for the domain of written knowledge in Chinese. By aggregating and analysing the empirical data to avoid methodological nationalism, the thesis will assess the cultural thickening patterns they exhibit, as well as the underlying power relationships that are expected to reflect on the cultural politics of what Yang (2003) discussed as the “Chinese cultural sphere” on the Internet (Yang sees this as a transnational Chinese cultural sphere, though this is not always socially

organized into collective action). The proposed framework embodies Yang's concept in a more concrete fashion, and the Chinese language here serves as the most basic and fundamental cultural repertoire.

The three main concepts proposed here complement Yang's (2003) notion of a Chinese cultural sphere in several ways. First, while Yang used a cultural approach to examine the rise of a transnational Chinese cultural sphere enabled by the Internet, the idea of cultural thickening proposed here does not assume the rise of a certain type of cultural sphere, be it national or transnational. Second, while Yang's ethnographic approach captured the activities of using and visiting the same websites as sharing a common cultural and emotional repertoire, it may have overlooked the geolinguistic processability that may facilitate or hinder such sharing patterns, especially across different Chinese-speaking regions. The proposed concept of processability takes into consideration both the human and machine sides of this processability of such a Chinese cultural repertoire. Third, for Yang, the Chinese cultural sphere, while falling short of clear organization or collective action, "goes some way toward exploring the meaning of being Chinese in the global context" (2003, p. 486). The proposed framework goes one-step further by analysing cultural patterns and exploring how they provide distinct answers about the meaning of being Chinese. Fourth, Yang seemed to believe that the Internet, despite being the very product of the modern nation-state and capitalism, has also encroached upon the power of state and market institutions by "loosening territorial and human barriers" (2003, p. 485). However, rather than expect that barriers would be loosened, the proposed framework is agnostic on this point, because the way certain patterns of thickening can occur depends on how existing boundaries are reintroduced or overcome.

To demonstrate the usefulness of the proposed concepts for understanding the shifting cultural- political boundaries within the Chinese-Internet, I now reformulate the research question of this thesis as follows:

How do major cultural thickening patterns, produced by the web spheres of major websites (e.g., user-generated encyclopaedias), overcome, reinforce or ignore the existing cultural-political boundaries that are reinstated by the geolinguistic processability (such as geolinguistic codes zh-HK, zh-TW and zh-CN)?

3.2 Mixed methods: geolinguistic analysis of cultural patterns

To analyse cultural patterns, an overall mixed-methods research design is proposed to integrate both quantitative and qualitative data. Three aspects of user-written encyclopaedias will be studied in each of the following chapters. The selection of methods are informed by Web social science (Ackland, 2013), digital methods (R. Rogers, 2013), and digital social research (Marres, 2012), resulting in a multi-method comparative case study of structured, focused comparison (George & Bennett, 2004; Drozdova & Gaubatz, 2009). The following section summarizes how each aspect fits in the overall research design.

3.2.1 Editorial processes and conditions: policing, participating and prioritizing. Chapter 4 will seek to answer how Baidu Baike and Chinese Wikipedia develop rules and practices for editorial development. The chapter begins with an editorial analysis of the two encyclopaedias, comparing how they process user inputs differently. Because their content depends largely on user contributions, the chapter then contextualizes the historical development of the two encyclopaedias within the larger story of the filtering and diffusion dynamics of information in the Chinese-language internet and its users.

It is important to note that, in view of the concept of geolinguistic processability, Hong Kong, Taiwan and mainland China are treated as separate geolinguistic units: Each has its own geolinguistic web sphere signified by its geolinguistic identifier of language code online; at the same time, each refers to its own specific media and political system that have historical and ongoing tensions with those of the others. Thus for the purpose of this thesis, these units are not units with essentialized ideological, cultural or political differences but

dynamic units that are open to change while also having specific historical paths of development in their respective media cultures. If essentializing units is to be avoided and “cultural thickening” to be identified, we must treat these geolinguistic units as the outcomes of historical construction, specifically in relation to the two encyclopaedias, the information they provide (and its reliability) and users.

As both websites are user-generated, the web spheres constructed by them can be expected to be porous: originally not internal to the websites, certain pieces of information and knowledge are accepted and made part of the websites. In addition, certain users (contributors and readers alike) contribute their judgement to the editorial processes. Thus, for the purpose of this research, both the internal and external data of the two encyclopaedias should be collected in considering their internal editorial processes and external conditions. To observe how they cope with boundaries and differences across different geolinguistic regions, it will be important to analyse how units are socially constructed in the overall editorial process. Alternatively, in terms of the three major concepts proposed, how do these two websites construct their own web spheres by arranging their geolinguistic processability such that it results in cultural thickening patterns? Also, inasmuch as the web spheres of encyclopaedias are “bounded” in the sense that only certain types of information are admissible and permitted to stay, and that only certain type of users have the literacy skills and knowledge backgrounds to contribute, theories of “relevance filtration and accreditation” (Benkler, 2006, p. 169) and “network gatekeeping” (Barzilai-Nahon, 2008) will be used to identify cultural thickening patterns.

Data collection. To do this, first, I collected primary and secondary data on the editorial policies of both websites to examine how they filter information. The data sets include:

- online documents—editorial policies, discussion pages, self-reported data by the “power users” (users who are granted extra privileges);

- media reports—mainly newspaper articles and some television reports;
- descriptive statistics on power users and all encyclopaedia entries—including the number of external links and the length of articles, based on the snapshot data set collected in June 2010 (the same set that is also analysed in the Chapter 5 (on webometric and content outcome)).

Second, I collected contextual data that concerns the flowing and blocking dynamics across Chinese-speaking regions:

- Internet diffusion rates—time series data (1990-2011) for Chinese provinces and East Asian countries;
- media reports and user commentaries—mainly those concerning the two encyclopaedias and related online censorship and blocking incidents;
- second-hand research—concerning the behaviours and activities of Internet users

Because user-generated encyclopaedias are often seen as “emerging” and “self-organized”, the data collected must address the dimension of time to capture the origin and changing evolution of the websites, especially in relation to their contributors and hosts. Thus, this portion of the research considers the time dimension from two perspectives: First, for the editorial policies and practices, the research will consider the historical context where Baidu Baike emerged as an alternative to Chinese Wikipedia. It will also examine which self-appointed contributors are given extra privileges. It will also study how geolinguistic processability (or geolinguistic factors) is integrated into the platforms and editorial development. Second, because Wikipedia was blocked in mainland China most of the time from 2005 to 2008, which coincided with Baidu Baike’s launch in 2006, it is important to consider how the competition between Chinese Wikipedia and Baidu Baike can be understood in the context of the growth patterns of Chinese-speaking Internet users.

Methods used. Most of the data, particularly the textual data, can be analysed using textual analysis, i.e., providing the most likely interpretation based on the author's own personal experience as a user of Chinese user-generated encyclopaedias and as an observer of media systems across Chinese-speaking regions. The comparative aim is to identify the underlying power relationships governing cultural thickening activities on the two websites so as to see them as the outcome of the media systems of certain Chinese-speaking region(s). I will provide a number of perspectives on these interpretations to avoid biased interpretations. In addition, relevant statistical and survey data is provided as a supplement to the interpretations of the textual data. Altogether, this mix of methods will show how the two encyclopaedias filter information differently, and how such different filtration and accreditation processes are contextualized when Chinese Wikipedia was blocked in mainland China from 2005 to 2008. This will yield findings about how the existing boundaries among Chinese-speaking regions are reinforced, ignored or overcome by the cultural thickening activities within the two encyclopaedia websites and in relation to their users, processed information and website hosts.

The cultural-political dynamics within and between the editor/contributors are expected to reflect the dynamics of Chinese-speaking Internet users, and cultural thickening patterns are thus the direct result of editorial norms and practices of such "relevance filtration and accreditation" or "network gatekeeping" processes made by self-appointed users. The overall editorial patterns will be considered - not just limiting them to some extreme cases - in order to detect the differences between the two encyclopaedias. A comprehensive comparison of the outcome of the editorial processes will be presented. Although such an effort is the main focus of chapter 5, which uses mainly quantitative data to identify patterns of cultural thickening, some of the findings of the mainly qualitative analysis of Chapter 4 on editorial processes are also relevant to these patterns. This demonstrates the value of the mixed-method

approach and of integrating qualitative textual analysis with quantitative webometric and content analysis.

Cultural thickening can avoid methodological nationalism, but the broad topics of Chinese national culture, media cultures, and media and educational institutions, previous research can also provide background knowledge on the cultural-political differences between the various Chinese-speaking regions, and in particular between Hong Kong, Taiwan and mainland China because of the notions of the “Chinese nation”, “Greater China” and “Cultural China”. It is important not to essentialize these notions but rather incorporate them as possible types of cultural-political construction of the two encyclopaedias. Thus, the methods must be inherently comparative in several dimensions: They must be comparative for both encyclopaedia sites as web spheres, for the different construction outcomes of “Chineseness”, and for the underlying cultural-political power dynamic of the Chinese-language Internet.

Expected outcomes. Based on both internal and external data for the two encyclopaedias, I examine their different editorial practices, from policing “unwelcome” user contributions and encouraging user participation to prioritizing editorial goals. Different user-generated encyclopaedias can be seen as different web spheres that process all incoming edits and information. Their processability is configured differently in view of different user interfaces, editorial policies, participating members, and historical paths of development. By contextualizing these differences in the complications of the wider Chinese Internet development, we can gain insights on the boundary questions for these Chinese-speaking regions from the two contrasting “cultural thickening” patterns. These in turn are possibly indicative of larger cultural-political dynamics of the Chinese-language Internet.

The findings, when taking into account the historical development of Chinese Wikipedia in mainland China, should also provide important insights into the effects of Beijing’s filtering and censorship regime. However, the

comparative perspective adopted here will also go beyond the traditional research concerns over the effects of the filtering and censorship regime on mainland Chinese users, and examine its effects on restructuring Chinese web spheres on different geolinguistic and cultural-politically significant paths. By examining how the two encyclopaedias mediate between their hosting organizations and users, different cultural thickening patterns can be identified.

3.2.2 Webometric and content outcome: covering the world and defining Chineseness. The quantitative analysis of the content of the two encyclopaedias will be complemented with an in-depth qualitative textual analysis. In the first half of chapter 5, the geographic coverage and linguistic preference of all entries will be analysed, treating all their online citations (external web links) as the proxy for content sources and the filtered outcome of line sources, thereby comparing their preferences in substantiating knowledge claims. In the second half of chapter 5, three entry articles that are central to any notion of Chineseness are compared: “Han Chinese characters” (hànzì 漢字 hereafter “Hanzi”), “Han Chinese Ethnicity” (hànzú 漢族 hereafter “Hanzu”), and “All under heaven” (tiānxià 天下 hereafter “Tianxia”), with the aim of providing textual evidence for the perspectives evident in the articles.

Data collection. Various webometric and visualization will be used and tailored for geographic and linguistic analysis. I adopt an integrated webometric, bibliometric and textual approach towards the produced outcome, combined with an additional geolinguistic perspective, in order to examine the geolinguistic preferences evident in the two encyclopaedias. I have collected for further analysis a comprehensive data set of all entries from the two encyclopaedias in June 2010, which include millions of articles, the web links they contain, and external web pages they link to.

Methods used. Chapter 5 conducts a comprehensive webometric comparison using various webometric and visualization techniques. These techniques, some of which are generalizable for other geographic and linguistic

contexts and some custom-made for Chinese-language internet, will help to extract the relevant information. To identify the overall geographic and linguistic preferences, all clickable external sources of the content (i.e. web links linking to external sources) are analysed, thereby comparing where and how two encyclopaedias cite their sources of knowledge claims. Second, for the qualitative textual comparison, the elements of sources, messages and narratives will be identified for an in-depth comparison of the ways in which different cultural patterns are articulated in various discourses, which are in turn examined in the light of the cultural-political dynamics of being Chinese in national, regional and global contexts.

Expected outcomes. Based on both quantitative and qualitative data, I can thus answer the question of how the entry articles of the two differ. The findings, based on a comprehensive quantitative analysis of all entry articles and an in-depth qualitative analysis of three articles central to any notion of Chineseness, will indicate how the two encyclopaedias incorporate certain perspectives from the Chinese-speaking worlds while marginalizing or even excluding others. If online encyclopaedias provide a convenient lens through which their users “know” the world, then the research should also indicate how the two respective worldviews of the two encyclopaedias converge or diverge.

In terms of cultural thickening patterns unique to encyclopaedia projects, both are examined as projects of engaging with knowledge- or information-sources that connect knowledge sources within and beyond the Chinese-language Internet. They bring certain webpages and their content into the focus of attention on their encyclopaedia platforms, thereby producing certain cultural thickening effects in relation to some frequently cited sources. By examining the top-linked websites in each region, the findings should also indicate whether and how they provide adequate cultural thickening patterns within and across each Chinese-speaking geolinguistic regions. It can be expected that while Baidu Baike is less diverse than Chinese Wikipedia in terms of its geolinguistic outlook, and

that Chinese Wikipedia provides substantial cultural thickening across various Chinese-speaking regions, whereas Baidu Baike exhibit cultural thickening patterns only within mainland China. Thus despite some criticisms that Chinese Wikipedia is “foreign”, the findings should provide better quantitative and qualitative indications about the level, extent, and content of “Chineseness” substantiated by the linking patterns to online sources.

3.2.3 User reception: encountering and using online encyclopaedias. To compare how Chinese-language Internet users from different regions encounter and use the two online encyclopaedias in chapter 6, two sets of data have been collected and analysed: search engine result pages (SERPs) and the social media of Twitter and Sina Weibo. From the first dataset, the measurement of visibility scores indicates the likelihood for users to encounter the two websites when they use search engines to find topics of cultural-political significance. In addition, with further network data analysis and clustering techniques, it is also feasible to detect the system boundaries among major Chinese-speaking regions based on the varied visibility patterns given to websites by search engines, thereby showing not only the central role of search engine geolinguistic choice in presenting the two encyclopaedias, but also the central role of major user-generated encyclopaedias as the most visible websites across a diverse set of search keywords. From the second dataset of social media texts mentioning the two encyclopaedias, various text analysis techniques are conducted with the aim to analyse users’ experience and perception on the two “brands” of Chinese-language user-generated encyclopaedias.

Data collection. To compare how likely it is that users use the two encyclopaedias, I take a user reception and mixed method approach that borrows ideas from television/media audience reception theories and Internet marketing industry practices that focus on search engine result pages (SERPs) and social media texts. First, I have collected SERPs, first in 2011 and then in 2012, each of which contains about 20,000 web links. The two snapshots of the same SERP

dataset should indicate what remains unchanged over time to make sure that the findings are not just ephemeral phenomena. Second, using the WeiboScope and DiscoverText tools, I have collected the social media texts from Twitter and Sina Weibo around 2011 and 2012 that contain specific keywords of “Baidu Baike”, “Chinese Wikipedia” and “Wiki” in both simplified and traditional Chinese characters, with the aim of providing a textual analysis of users’ experience of the two user-generated encyclopaedias.

In addition to the above data, I have collected some user log and user interview data during my field trip in various Chinese cities and towns in 2010, which will be incorporated as supplemental data for the mainly webometric and textual analysis of SERPs and social media content.

Methods used. Since previous research has indicated that search engines play a central role in directing traffic to user-generated encyclopaedias such as Wikipedia, it is essential to see how encyclopaedias are visible for Chinese-language Internet users who are already segmented into different search engine markets. By drawing on making connections between how search engine industry targets users and how the television industry targets audiences, I develop a user-reception method to examine SERPs as proxy for user traffic, similar to the practices of online marketing analysis for search engines. Based on industry practices, I employ webometric measurement and visibility scores, with the aim of comparing how websites ranked in different search engines across major Chinese-speaking regions over a select sample of about 3000 search keywords. The visibility scores enable both network visualization and analysis to show the effects of the geolinguistic choices of search engines (which can be called search engine variants) based on the selection of the most visible websites included among the top 10 search results. Note that these methods of extracting and analysing webometric data may likewise be generalizable for other geographic and linguistic contexts.

Second, for the dataset of social media texts mentioning the two encyclopaedias, I use basic spreadsheet and database software to code the message texts for users' perceptions of using Chinese-language encyclopaedias. I also conduct textual analysis to identify the recurring narratives about the user-generated encyclopaedias in general and the two "brands" in particular. Whenever possible, geographic and linguistic information is extracted to examine whether users' perceptions vary because of geolinguistic differences, thereby making inferences about the boundary effects of user-generated encyclopaedias.

Expected outcomes. Based on these data, I answer the question of how Chinese-language Internet users encounter and use the two encyclopaedias. For experiences and perceptions that are shared and agreed upon by the majority of users across different geolinguistic regions, it can be said that the observations converge. For experiences and perceptions that diverge across certain different geolinguistic regions, it can be said that the boundaries are reinforced. Further, because social media Twitter has been blocked in mainland China for a long time, it can be expected that users from mainland China may have limited access and thus limited presence. Thus, the difference between Sina Weibo and Twitter needs to be taken into account when making inferences about the boundary effects of user-generated encyclopaedias among different geolinguistic groups of users.

These findings should provide additional layer of analysis about major Chinese web spheres such as search engines and microblogs. Although the data is limited and focused on how these web spheres of search engines and microblogs make which online user-generated encyclopaedias more visible and more widely disseminated, this diverse set of website platforms should also provide an indication about how online Chinese-language web spheres are geolinguistically configured. By examining how these platforms mediate between their respective users and which user-generated encyclopaedias, further cultural thickening patterns can be identified.

3.3 Overall expected outcomes

As two instances of web spheres, Baidu Baike and Chinese Wikipedia may process information differently, thereby contributing to distinct cultural thickening patterns. Indeed, by examining how both connect Chinese-language users and information sources, researchers can conduct cultural-political analysis of the two websites. The findings can show which trans-local exchanges are encouraged (or discouraged) and which articulations of meanings are prioritized (or marginalized). Specifically, the patterns of “cultural thickening” across regions, identified by geolinguistic codes such as zh_TW (Taiwan), zh_HK (Hong Kong) and zh_CN (mainland China), will not only fill the current knowledge gap about the dynamics of “transnational Chinese cultural sphere” (G. Yang, 2003), but also contribute to a more general understanding of “Internet connectivity effects” (Haythornthwaite, 2005).

Table 3-4 summarizes possible patterns of cultural thickening for three selected regions, assuming a thickening pattern spans over at least one region. The thickening of pattern 1 covers all three, suggesting that all boundaries are overcome. The patterns 2 to 4 cover two of the three, suggesting different specific boundaries are overcome. The patterns 5 to 7 cover only one, suggesting boundaries are reinforced around respective regions. The patterns that will be found about the two encyclopaedias may match any one of these patterns.

Table 3-4

Possible cultural thickening patterns

	CN (Mainland China)	HK (Hong Kong)	TW (Taiwan)
pattern 1	thickening a		
pattern 2		thickening a	
pattern 3	thickening a		
pattern 4	thickening a		thickening a
pattern 5	thickening a		
pattern 6		thickening a	
pattern 7			thickening a

Table 3-5 summarizes plausible factors that influence cultural-political boundaries. The first row shows that each region has its own language tags when using respective country codes, suggesting a reintroduction of boundaries among them. The next two rows point to the political and geographic facts that mainland China and Hong Kong belong to the same Republic and share a border. (The Hong Kong-Shenzhen Corridor is one of the busiest channels of goods, capital and people in East Asia, with Shenzhen being a new mega-city of mainland China hosting many factories). If these factors matter, Hong Kong should converge with mainland China. The last two rows highlight the factors of Chinese characters and the Internet filtering regime imposed by Beijing: if they matter most, it can be expected that mainland China will be increasingly separated from Hong Kong and Taiwan.

Table 3-5

Factors influencing cultural-political boundaries

	CN (Mainland China)	HK (Hong Kong)	TW (Taiwan)
Language tags	zh_CN	zh_HK	zh_TW
Republic	PRC	PRC	ROC
Geographic bordering	Yes (Hong Kong- Shenzhen corridor)	Yes (Hong Kong- Shenzhen corridor)	No (separated by Taiwan Strait)
Chinese characters (Language tags)	Simplified Chinese (zh_simp)	Traditional Chinese (zh_trad)	Traditional Chinese (zh_trad)
Internet filtered by Beijing	Yes	No	No

The proposed theoretical framework can thus be applied to two Chinese-language encyclopaedia websites, identifying the boundaries of thickening patterns *within* and *around* web spheres that are geo-linguistically coded. After all, human- and machine-processable digital texts must work to aggregate, filter, and prioritize information that can be expected to overcome, reinforce, ignore, or create boundaries across different groups of users that are identified by their geolinguistic profiles. Geolinguistic processability, arguably the most salient aspect of processability for many major services on the Web, including search engines and user-generated encyclopaedias, enables cultural-political connectivity that may or may not overcome existing boundaries. While language boundaries constitute major barriers for online interaction (and thus boundaries *around* language web spheres), it remains an open question whether region markers (e.g. country codes) indicate another kind of barrier (thereby creating geographic boundaries *within* language web spheres). Thus, the theoretical framework allows researchers to conduct geolinguistic and cultural-political analysis of Chinese-language Internet and beyond.

A major feature of this research, and potentially a limitation, is the focus on the geolinguistic aspect of processability. Although the importance of

geolinguistic processability is evident in the industry practices and technical standards of localization, it may be that other important aspects of processability are overlooked. These should be also be studied elsewhere in due course. For the purpose of the thesis, the geolinguistic lens should adequately capture substantial and significant evidence of the dynamics of Chinese cultural thickening contributed by the two encyclopaedia websites, which could be national, regional or transnational. The proposed framework and methods in this chapter thus provide a much-needed integrated perspective that enhances our understanding of a language web sphere whose users amount to a substantial portion of the world's population and beyond.

Chapter 4 Editorial processes, Internet control, and Internet diffusion

The prospect of harnessing the potential of online users to work collaboratively on a project like Wikipedia has attracted much research attention, with concepts such as “peer production” (Benkler, 2006), “wikinomics” (Tapscott & Williams, 2008), “cognitive surplus” (Shirky, 2010) and the “democratization of innovation” (von Hippel, 2005). Users must be recruited and editorial practices must meet certain quality standards. How does Baidu Baike differ from Chinese Wikipedia as regards editorial practices? Moreover, how has the filtering and censorship regime imposed by the People's Republic of China (hereafter Beijing) influenced the take-off of the two websites, a question that will be relevant to how many people engage with this collaborative project?

Focusing on the websites' internal editorial practices and the historical development of users and potential users (especially those in mainland China), this chapter applies the concept of gatekeeping to describe two specific phenomena. The first kind of gatekeeping refers to the editorial policies and decisions internal to the two encyclopaedias. The second kind of gatekeeping refers to the Internet filtering by Beijing that prevent users in mainland China from accessing websites outside mainland China, including Chinese Wikipedia. For user-generated websites, the first type of gatekeeping (gatekeeping of information) often occurs before the second type (gatekeeping of user-contributors).

I use the same concept to describe two seemingly different phenomena mainly because I want to highlight the link between information and user-contributors: The two encyclopaedias must be developed in ways that will engage the production of information content as well as engaging user-contributors. The editorial policies for keeping or removing edits can be analysed as rules for information engagement. The blocking and unblocking of mainland Chinese users can also be examined as cultural-political policies influencing user engagement with the two websites in mainland China. Consistent with the

recent literature, mainly in the field of information science (Barzilai-Nahon, 2008) and other disciplines, including “information systems, management, political science and sociology” (Jucquois-Delpierre, 2007), the concept of gatekeeping can help us to better understand information control in network settings through the selection of information and of users.

By examining the editorial processes and policies, first for user contributions and then for potential users, this chapter will compare how, starting from a pool of Chinese-language users and content online, Baidu Baike and Chinese Wikipedia began to develop differently in regards to their respective mutually-reinforcing cycles of growing participation, content, and readership. Specific questions include the following: How do their editorial policies and technical practices reflect different gatekeeping mechanisms? How does Beijing’s filtering of Chinese Wikipedia constitute a gatekeeping factor in preventing users in mainland China from editing it? Finally, how do the two encyclopaedias contribute to different “cultural thickening” patterns?

Before answering these questions, I have summarized some basic information about Baidu Baike and Chinese Wikipedia in Table 4-1. First, Baidu Baike is hosted by a for-profit company in Beijing, Baidu.com, whereas Chinese Wikipedia is hosted by a non-profit charitable organization in San Francisco. The geographic and organizational aspects of their respective hosting organizations may influence the websites’ editorial processes because of differing laws and regulations. Second, although Chinese Wikipedia started as the first and only Chinese-language user-generated encyclopaedia, Baidu Baike nevertheless has a much larger number of entry articles than does Chinese Wikipedia (about ten times more in September 2012). Baidu Baike surpassed Chinese Wikipedia in the number of articles on roughly the third day after its initial launch on April 20, 2006. Third, Baidu Baike supports web pages encoded in the standard of gb-2312, a Chinese national standard that permits only simplified Chinese. In contrast,

like all other language versions of Wikipedia, Chinese Wikipedia uses Unicode, an international standard that supports all the languages in the world.

Table 4-1

Basic information about Baidu Baike and Chinese Wikipedia

Features	Baidu Baike	Chinese Wikipedia
Operator	Baidu.com, Inc. (a NASDAQ-listed web services company known for its search)	Wikimedia Foundation Inc. (a nonprofit charitable organization)
Physical location	Beijing, China	San Francisco, USA
Number of entries (September 2012)	over 5.3 million	around 527,000
Date of 1st article	20 April 2006	17 November 2002
Date of 100,000th article	23 April 2006*	12 November 2006
Date of 1,000,000th article	16 January 2008	N/A
Language policy	Simplified Chinese Only (gb2312)	Both simplified and traditional Chinese (Unicode)

Note. Based on self- and mutual-description of Baidu Baike and Chinese Wikipedia as entry articles in both encyclopedias.

*Interpolated value based on Baidu Baike's articles' sequential serial identification numbers

This chapter will first analyse the editorial policies and practices of the two encyclopaedias. Then it will explore the impact of Beijing's Internet filtering regime on gatekeeping as it affects the growing number of mainland Chinese Internet users. The following abbreviations and conventions will be used when referring to pages and users of the two encyclopaedias: BB (Baidu Baike), zhWP (Chinese Wikipedia), enWP (English Wikipedia), and User:userID (referring to a user who registers herself/himself with a userID).

4.1 Editorial policies and processes: filtering user contribution

Any submission by the users of Baidu Baike must go through an internal review process conducted by Baidu employees, before it can take effect on the content outcome (NetEase, 2012).

There is not really any editorial team inside [Chinese] Wikipedia. The knowledge level of its editors (not teams) is not much better than that of Baidu Baike's. The main reason why the quality of articles appear to be higher in Wikipedia is that the low-quality edits are filtered at the first moment (Z. Chen, 2012).

Rejecting bad edits while keeping good ones, as explained in the quotes above, is the primary task for the two major online Chinese-language encyclopaedias (Liao, 2009c). As quoted above, Baidu Baike uses “an internal review process conducted by Baidu employees” and “low-quality edits are filtered” by Chinese Wikipedia’s editors. Such observations seem to confirm what has been reported about Wikipedia projects: for wikis to be run as successful collaborative platforms, certain policies, norms and a technological architecture are required (Forte & Bruckman, 2008; Konieczny, 2009). In clear contrast, Baidu’s salaried employees review new edits to Baidu Baike, whereas user-contributors remove “low-quality” edits among themselves on Chinese Wikipedia.

While both encyclopaedia websites claim that “any one can edit”, further analysis is required on the actual network boundaries of “any one”. Some descriptive comparison of their editorial practices exists (Liao, 2008; Suo, 2007), but detailed in-depth analysis is lacking. The first half of the chapter will provide a systematic comparison to fill this gap. Direct evidence is first drawn from both explicit and implicit rules, which provide editorial and stylistic guidance that serves to structure the content and collaboration of user-generated encyclopaedias (Butler, Joyce, & Pike, 2008). To avoid arbitrary interpretation of rules, additional data from user testimonies, media commentaries and outcome

are included whenever necessary. The findings can be categorized under three themes: editorial priorities, power users and geolinguistic arrangements.

4.1.1 What do they delete/revert differently? The two encyclopaedias process new edits differently through addition, deletion and selection.

First, salaried Baidu employees do the bulk of the Baike internal review work, whereas the same work for Chinese Wikipedia is done by volunteer user-contributors, a process in which Wikimedia employees rarely intervene. According to the earliest and subsequent versions of “Baike's Basic Rules,” users must “subordinate completely to the unified management of Baidu Baike” to avoid deletion of their edits and commentaries (BB, 2006, 2012a). Thus by default and in practice, new edits must go through Baidu’s internal review, a process that is not transparent to other user-editors. In direct contrast, for Chinese Wikipedia any new edit will, by default, take effect transparently and go through the open editorial process where decisions of reversion (undoing new edits) or deletion are made. Thus, although both websites support basic functions such as editing, commenting and edit-history tracking, the actual editorial practices differ in processing new user contributions: Baidu Baike adopts a model of largely internal review (by Baidu employees), and Chinese Wikipedia uses an open review (by users) model.

The two models also process copyright-dubious materials and politically sensitive materials differently.

Copyright-dubious materials. Based on evidence from various sources, Baidu Baike’s internal review system is rather lax in screening copyright-dubious materials, whereas Chinese Wikipedia’s open review is much stricter. Indeed, since its launch in 2006, Baidu Baike has been reported to be the “worst copyright violator” of Wikipedia’s copyright (Nystedt, 2007). This scale of violation is also corroborated by evidence from Chinese Wikipedia users, including obvious searchable digital traces of English and Chinese signature phrases, such as finding “From Wikipedia, the free encyclopaedia” and “citation

needed” in the content of Baidu Baike, and a sample letter of copyright complaint has been prepared for sending complaints to Baidu (zhWP, 2012c). Based on copying content from Chinese Wikipedia, Baidu Baike quickly (within three days) overtook Chinese Wikipedia in terms of the number of entries, raising the issue of unfair competition (User:430072, 2006).

That Baidu Baike copies Chinese Wikipedia also highlights Wikipedia’s copyleft norm for Wikipedia in content sharing. The copyleft licenses demand that any derived works must be released under compatible copyleft licenses to ensure further sharing. Since I was myself one of the copyleft advocates in Taiwan and the person who came up with a Chinese translation of the term “copyleft,” I agree with the legal opinions given by some Chinese Wikipedians that Baidu Baike has violated the copyright of Wikipedia by refusing to remove the copyright symbol (e.g., “©2012 Baidu”) found at the bottom of each article page (zhWP, 2012c). Baidu Baike can legally copy any content, in full or selectively, from Wikipedia, but when doing so it must both give attribution to Wikipedia and release the final content of Baidu Baike with a copyleft license, instead of claiming ownership of the content as if it is the property of Baidu. This violation prompted some Chinese Wikipedia contributors to declare, as a gesture of protest, that “any use of this image by Baidu Baike is not permitted” on their contributed images (User:ljq513, 2012).

Still, it should be noted that “copy and paste” activities are more common among users around the globe than a basic knowledge about copyleft practices. As reported by user-contributors, at least 10% of newly added entries to Chinese Wikipedia also include copyright-dubious materials, mostly contributed by new users. A taskforce has duly been organized to prevent such copyright violations (zhWP, 2012e). In addition, copyright violation constitutes one possible basis for article deletion (in open review by user-contributors), along with promotional content and advertisements (zhWP, 2012b). In contrast, for Baidu Baike to review copyright violation complaints, copyright owners must send the

complaint via snail mail to Baidu's address in Beijing for internal review (BB, 2012a). It has been reported by contributors of Chinese Wikipedia that Baidu produces no substantial reply to the copyright-infringement complaints that Chinese Wikipedia has filed against Baidu (zhWP, 2012c).

Politically sensitive materials and comments. This lax/strict contrast is reversed when the two encyclopaedias deal with politically sensitive materials and comments. Baidu Baike has a record of censoring content whereas Chinese Wikipedia has an explicitly stated policy that "Wikipedia is not censored" (zhWP, 2012f).

In terms of politically sensitive materials and comments, Baidu Baike rejects any content that may offend the Chinese authorities in Beijing, whereas, again, Wikipedia's policy is not to censor anything. Baidu the company has defined seven categories for deletion from Baidu Baike (BB, 2012a). Among them, the third category refers to "reactionary content" (fǎndòng nèiróng 反动内容), an ideologically and politically loaded term in the context of mainland China, which means any content that "maliciously criticizes the current system of the state", "disrupts social and public order", "provokes disputes over nationalities, ethnicities, religions and regions", "maliciously attacks state agencies and officials", "promotes superstition and cults", or "provides any hyperlinks to the aforementioned content", all of which are in line with Beijing's priorities in policing and censoring online content. (Though not specifically explained in the website's rules, the "state" refers to Beijing.) In contrast, Chinese Wikipedia has no such policies. Instead, it maintains that it is a "Chinese-language" encyclopaedia that serves users around the world beyond just mainland China and Taiwan, and user-contributors must be vigilant against China-centric or Han-Chinese-centric viewpoints when paraphrasing sources (User:Lorenzarius & other contributors, 2010).

Furthermore, many politically sensitive comments exist on Chinese Wikipedia, mainly because contributors do not need to go through the internal

review process required by Baidu Baike when editing articles. Based on my own observations of the discussion pages on both encyclopaedias, politically sensitive comments are fairly common in Chinese Wikipedia but not in Baidu Baike. Some users openly declare their political leanings and attitudes in user pages, ranging from pro-Communists to pro-Falun Gong, and from Han-Supremacist to East-Turkestan independence supporters. I have yet to find another Chinese-language website that hosts users as politically diverse as those of Chinese Wikipedia.

Although researchers cannot directly examine Baidu's non-transparent internal review processes, several leaked internal documents provide convincing details of the day-to-day "internal monitoring and censorship" operations. For example, there are documents that are claimed to be leaked from Baidu the company by Chinese overseas political activists (Xiao, 2009) which include employee logs, performance evaluations, and the directives issued by the authorities. These documents reveal the constant censorship efforts of many of Baidu's services, including Baidu Baike and its discussion forums. This evidence is consistent with other reports of Chinese censorship of user-contributed content (R. MacKinnon, 2008; Woo, 2007).

In contrast, while Chinese Wikipedia's open review model does not guarantee that all political views will be included, the politically sensitive edits and discussions are recorded by default for public review. Formal policies on "edit wars" are established to guide normal users and administrators (zhWP, 2012d, 2012g). In fact, a non-factual satirical article called "The Political Edit War in Chinese Wikipedia" describes the history of Chinese Wikipedia as a war between supporters of the People's Republic of China and supporters of the Republic of China (Taiwan), using the "weapons of vandalism" to impose what they perceive to be a "neutral point of view" (zhWP, 2012h). The fact that such a wide array of politically sensitive articles exist on Chinese Wikipedia allowed an independent researcher (Chu, 2013) to conduct a study testing China's Internet

filtering regime by accessing all of its entry articles, thereby differentiating between blocked and unblocked articles and reverse-engineering the filtering mechanisms.

In 2005 and 2007, there were proposals discussed on Chinese Wikipedia to practice self-censorship, which generated lengthy and heated debates (Chu, 2013). These proposals were rejected by the community, but they sought to ensure that a self-censored “apolitical” encyclopaedia would make Chinese Wikipedia accessible again for users in mainland China. A group of editors who supported these self-censorship proposals reportedly migrated to Baidu Baike after the proposals were rejected. Chu has praised the public nature of discussion on Wikipedia as an “extremely valuable resource to study the cultural phenomenon of self censorship in China” (Chu, 2013). Thus, the explicit “no censorship” policy has faced some challenges from the self-censorship culture brought by users who are sympathetic to the Chinese authorities, but their actions on Wikipedia were publicly recorded and scrutinized by users who hold different opinions about editorial policies. The transparency of edits and editorial discussion which explain content-deletions could be construed as documenting ‘censorship’; but otherwise, edits can be reverted to restore the deleted content on the grounds that “Wikipedia is not censored” policy (zhWP, 2012f).

Hence, the different ways in which the two encyclopaedias handle politically sensitive materials and comments present a clear contrast, suggesting two distinct approaches in processing user contributions: Baidu Baike employees purge politically sensitive content based on Beijing’s directive, whereas Chinese Wikipedia contributors engage one another openly, sometimes sparking heated edits and discussions among themselves.

Indeed, the contrast between the two websites in their attitudes towards copyright and politics was also observed by a user-contributor of Chinese Wikipedia, around the time when Baidu Baike was launched (User:430072, 2006):

I took a walk there when Baidu Baike was first launched ... Today [May 5, 2006, shortly after Baidu Baike's launch in April 20] I checked again. It has now more than 82,000 articles. The growth speed is astonishing. Some contributors there have no sense of copyright, pushing copy/paste to an extreme, without citing where the content comes from. Since [that time] much content has been moved to Baidu Baike from other websites, Wikipedia included. ... In particular, many high-quality non-political articles and images of Chinese Wikipedia became part of Baidu Baike without attribution. (Content written in traditional Chinese was converted to simplified characters). ... If Baidu continues to let this happen, Baidu will declare itself the biggest Chinese-language encyclopaedia in the world.

This quote not only encapsulates what has been discussed on different attitudes in processing copyright-dubious and politically sensitive materials, but also points out the conversion of traditional Chinese character content to simplified Chinese, a point to which the thesis will return when considering the geolinguistic factors.

Quantity and quality considerations. The quote above also confirms the contrast between quantity-minded Baidu Baike and quality-minded Chinese Wikipedia, as reported by user testimony (Z. Chen, 2012), Chinese online media (NetEase, 2012) and users of Chinese Wikipedia (User:430072, 2006). Even users of Baidu Baike agree: “Baidu Baike has a large amount of low-quality articles,” one of its senior contributors commented in an informal online survey of more than 100 contributors, as part of a long essay summarizing his five-year experience with Baidu Baike since its launch (User:gaoyunhuimin, 2011, 2012). These reports also mentioned the fact that various performance metrics are imposed and implemented by Baidu, as if users were playing role-playing games online: users need to earn points to climb up the ladder of contribution. Such a point-based system has been reported as being abused by technical inflation.

The quality-versus-quantity contrast is independently confirmed by the webometric findings based on the 2010 dataset of all articles and their external

links, the details of which will be further explained and analysed in Chapter 5. Using the webometric measurement of word counts and the number of external links⁵, Figure 4-1 shows the distribution of all articles. First, assuming that quality articles must be of a certain length, Chinese Wikipedia does indeed have better quality control measures in place to filter very short articles. The two graphs about the distribution of word counts in Figure 4-1 show that Baidu Baike does not filter short articles, whereas in Chinese Wikipedia those articles with a length shorter than one hundred words (10^2) will be removed. This distinction can be explained by Chinese Wikipedia's policy on "stub" articles, by which short articles are flagged for the attention of users so that they should expand the content or delete the entry, a common policy also in other versions of Wikipedia.

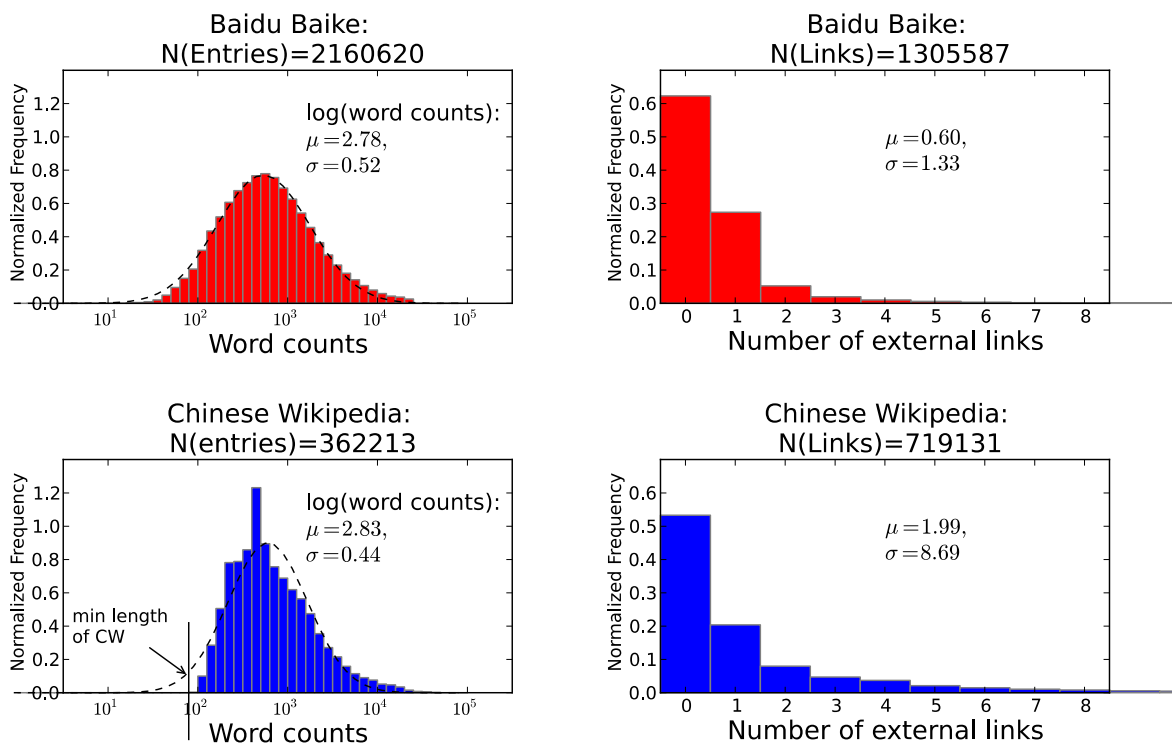


Figure 4-1. Number of entries and external links, and their normalized frequency

⁵ The word counts were calculated by removing HTML markup tags and common punctuation symbols of the body texts using programming scripts written in Python. The external links were collected and filtered to exclude the links to their respective sister websites. Please refer to section 5.1.2 for details.

Second, although not always reliable measure of quality due to potential spams, the number of external links can nevertheless provide a rough indicator of the richness of content sources. Chinese Wikipedia has slightly more external links per entry and is thus perhaps better sourced than Baidu Baike is. As shown by the two graphs on the right in Figure 4-1, Chinese Wikipedia has 1.99 external web links per article, which is significantly larger than Baidu Baike's 0.60. The fact that Baidu Baike has proportionally more article entries (over 60% of total external links) that contain no external links at all indicates that contributors fail to support these articles with proper online sources. Note also that although Baidu Baike has six times the number of articles as Chinese Wikipedia, it has less than twice the number of external links. Chinese Wikipedia is thus on average relatively better sourced than Baidu Baike.

In conclusion, Baidu Baike is more restrictive of new user contributions on the basis of political content whereas Chinese Wikipedia is more restrictive on the grounds of copyright and quality. The following subsections will continue to analyse their internal hierarchies, and their relation to their respective external environments.

4.1.2 Who has extra privileges? As for the question of who constitutes the gatekeepers and “who to trust” among users, who themselves facilitate “relevance filtration and accreditation” (Benkler, 2006, p. 465), the following paragraphs detail the contrasting features of those users with extra privileges, also called “power users”.

Power users: between employees and normal users. Pivotal to the internal hierarchies of the two sites are the “power users”: “Baike Kedou” (bǎikē kēdǒu 百科蝌蚪) for Baidu Baike and “administrators” for Chinese Wikipedia. Here I use the generic term “power users” loosely to describe a group of users who are given extra powers and privileges beyond what “normal users” have. They are the most privileged group of users. Note that I do not intend to argue that Baidu Kedou

and Chinese Wikipedia administrators play similar or equivalent roles (in fact they do not). I merely aim to contrast how these “privileged” users operate in the power hierarchy of their respective editorial settings.

The power to change rules and select power users is limited to Baidu employees who exercise several managerial powers. Baidu employees select around 100 users, based on their performance and application, to become members of the “Baikē Kedou Group” (bǎikē kēdǒu tuán 百科蝌蚪团). Although any Baidu Baikē user can apply, the selection and review processes are not open. The Kedou-specific rules outline the extra rights and obligations for these power users (see Figure 4-2): they enjoy an exclusive group platform, a status icon in front of their user ID, individual editing platforms, and other undisclosed powers and benefits (BB, 2009, 2012b), including the privilege of “green channels” and receiving monthly and seasonal gifts (BB, 2012b; BB contributors, 2012; NetEase, 2012). Thus, the power of the Kedou group of users is still subordinate to the Baidu employees. While Kedou members’ new edits can normally go through the “green channels” without being first reviewed by Baidu employees, the rules clearly indicate the hierarchical power relationship. Figure 4-2 details the Kedou Group Rules (kēdǒu tuán zhāngchéng 蝌蚪团章程) by which the Kedou members are evaluated by Baidu employees, who manage the “eligibility management” (zīgé guǎnlǐ 资格管理), “appraisal system” (kǎohé zhìdù 考核制度), “leave system” (qǐngjià zhìdù 请假制度), and “reward and penalty system” (jiǎngchéng zhìdù 奖惩制度) - as if Kedou members were workers managed by Baidu in a working environment.

<p>🔍 百科团队帮助</p> <ul style="list-style-type: none"> • 蝌蚪团章程 <ul style="list-style-type: none"> ▶ 总则 ▶ 权利与义务 • 附录 <ul style="list-style-type: none"> ▶ 管理与考核 <ul style="list-style-type: none"> ▶ 资格管理 ▶ 考核制度 ▶ 请假制度 ▶ 奖惩制度 • 团员操作手册 	<p>? <i>Baike Team Help</i></p> <ul style="list-style-type: none"> • <i>Kedou Group Rules</i> <ul style="list-style-type: none"> ▶ <i>General provisions</i> ▶ <i>Rights and obligations</i> • <i>Appendix</i> <ul style="list-style-type: none"> ▶ <i>Management and evaluation</i> <ul style="list-style-type: none"> ▶ <i>Eligibility management</i> ▶ <i>Appraisal system</i> ▶ <i>Leave system</i> ▶ <i>Reward and penalty system</i> • <i>Manual for team members</i>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 4-2. Baidu Baike team help: Rules outlined for Baike Kedou members

In contrast, normal Chinese Wikipedia users can, among themselves, change rules and select power users in a transparent fashion. These power users, or “administrators” for Chinese Wikipedia, once nominated and elected by users themselves via open discussions, have extra powers, including the powers to temporarily ban other users from contributing (Butler et al., 2008; zhWP, 2012a). Unlike the pivotal role played by Baidu employees in Baidu Baike, the employees of Wikimedia Foundation rarely intervene in the day-to-day editorial processes and have no say in electing administrators for Chinese Wikipedia. Hence, acting between the employees of the hosting organization and fellow users, the power users of Baidu Baike and Chinese Wikipedia differ in the extent, level of transparency, and internal relationship of their given extra powers.

To conclude, the source of managerial authority as gatekeepers is arranged differently: for Baidu Baike the source of authority is kept mostly inside Baidu the company and only a few users are selected as Kedou members; for Chinese Wikipedia the authority is kept among the users of Chinese

Wikipedia, rather than with the Wikimedia employees in San Francisco. The next subsection will further examine the role of the hosting organizations (Baidu and Wikimedia) and their geo-political context.

Beijing and San Francisco: the geo-political context. The following paragraphs will contextualize how the hosting organizations fit into the wider geo-political context.

On the level of hosting organizations, Baidu Baike is hosted by a for-profit company, Baidu.com, in Beijing, China whereas Chinese Wikipedia is hosted by a non-profit charitable organization, the Wikimedia Foundation, in San Francisco, USA (see Table 4-1). Baidu Baike is one of many services found on the web domain name Baidu.com, whereas Chinese Wikipedia is one of many different language versions of Wikipedia.org. Since Baidu Baike is a service hosted by a search engine company and Chinese Wikipedia is a service hosted by a global encyclopaedia project, they differ in many respects. Financially, whether Baidu Baike by itself (as a unit which is managed inside the larger company Baidu, but whose finance cannot be assessed separately) is financially sustainable remains a mystery; Wikipedia is funded by annual fund-raising efforts run by the Wikimedia Foundation. Encyclopaedia pages in Baidu Baike do have ads promoting products and services provided by other companies, whereas Chinese Wikipedia has no ads. Politically, as mentioned earlier, Baidu the company has editorial policies and practices of purging content that is deemed politically sensitive by Beijing. In contrast, although users of Chinese Wikipedia must comply with the “Terms of use” drafted by the Wikimedia Foundation (2012a) for all language versions of Wikipedia, the foundation does not engage in similar practices of purging politically sensitive content. Also, the footprint of the hosting organizations on editorial politics differs. In an opaque fashion, Baidu the company unilaterally sets the policies, which expanded from seven subsections on a single page in 2006 to seventeen subsections in 2008, not to mention its internal editorial review model by its salaried staff. In Chinese

Wikipedia, it is the user-contributors themselves who set and enforce editorial policies. Even the copyright violating materials are purged by user-contributors. Hence, Baidu Baike operates in a corporate setting with political oversight by Beijing, whereas Chinese Wikipedia governs itself among its user-contributors.

On the level of the wider geo-political context, a contrast exists as to whether and how political authorities influence or are involved in the operations of the hosting organizations. As mentioned earlier, the political influence by Beijing on Baidu Baike is evident from multiple sources as regards the “internal monitoring and censorship” routines (User:430072, 2006; Woo, 2007; Xiao, 2009). In contrast, there is little to no evidence that the US authorities have exercised their influence over the Wikimedia Foundation so as to shape the content on Chinese Wikipedia. On the contrary, there are a few cases where certain US agencies, politicians and corporations are exposed (and thus embarrassed) for contributing edits as normal editors (Fildes, 2007; Hafner, 2007). Some leaked documents regarding Baidu’s daily internal monitoring and censorship indicate the editorial influence on Baidu’s content (including Baidu Baike) by the government authorities in Beijing (Xiao, 2009), while there is less of a likelihood that government officials in Washington DC have sought to shape the development of Chinese Wikipedia through the operations by the Wikimedia foundation. In fact, some Western observers have suspected that certain articles in Chinese Wikipedia are undermined by self-censorship, thereby presenting viewpoints sympathetic with Beijing (French, 2006), suggesting that the editorial processes in Chinese Wikipedia may be more easily influenced by Beijing than by Washington or San Francisco. There are also unverified reports that some Chinese Wikipedia articles are edited by paid online commentators, often known as the “50 Cent Party,” in order to defame pro-democracy opinion leaders (Lam, 2012).

To sum up, the above assessment of the wider geo-political contexts of the hosting organizations and their editorial processes presents many contrasts.

Baidu the company has a much larger footprint on Baidu Baike than the Wikimedia Foundation has on Chinese Wikipedia. Also, several sources of evidence indicate the direct influence of the authorities in Beijing on Baidu Baike's editorial processes, whereas this link of influence is missing between the authorities in Washington and the Wikimedia Foundation in San Francisco, and their influence on and intervention in Chinese Wikipedia is found to be limited and inconsequential. In addition, some sources suspect that Beijing's interference in Chinese Wikipedia is conducted via paid online commentators. In sum, the wider geo-political contexts, largely set by Beijing and Washington, have quite different impacts: Baidu Baike is influenced by and embedded in the national legal, regulatory and cultural framework of the People's Republic of China, whereas Chinese Wikipedia is more self-governed and transnationally-structured. Hence we can now also examine how these geo-political contexts are reflected in the geographic distribution of power users.

Power users from Hong Kong and Taiwan: the major difference. To further compare the geographical locations of power users, I collected and coded self-disclosed data for the years of 2009 and 2012. The disclosure rate for Chinese Wikipedia is a bit higher (over 90%) than Baidu Baike (over 79%), suggesting that the majority of power users do disclose their place of residence. The findings in Table 4-2 show how power users are distributed across various regions, including mainland Chinese regions (zh-cn), regions with ethnic majority Chinese populations (zh), and the rest of the world. The absence (versus the presence) of power users from Taiwan, Hong Kong, Macau, and Singapore distinguishes Baidu Baike from Chinese Wikipedia.

Table 4-2

Geographic distribution of the power users

	Baidu Baike		Chinese Wikipedia	
	2012	2009	2012	2009
Total	97	99	78	78
Disclosure ratio	79.38%	70.71%	92.31%	97.44%
<i>Mainland China (Zh-cn)</i>	<i>77</i>	<i>69</i>	<i>25</i>	<i>24</i>
Beijing	10	3	7	6
Shanghai	4	6	5	3
Guangdong	8	7	4	4
Other provinces	55	53	9	11
<i>Zh Regions (Zh)</i>	<i>0</i>	<i>0</i>	<i>33</i>	<i>37</i>
Hong Kong	0	0	13	16
Macao	0	0	2	2
Taiwan	0	0	16	18
Singapore	0	0	2	1
<i>Asia Pacific</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>2</i>
Japan	0	1	0	0
Australia	0	0	1	2
<i>Americas</i>	<i>0</i>	<i>0</i>	<i>10</i>	<i>11</i>
United States	0	0	9	8
Canada	0	0	1	3
<i>Europe</i>	<i>0</i>	<i>0</i>	<i>3</i>	<i>2</i>
Germany	0	0	1	1
United Kingdom	0	0	1	0
Sweden	0	0	1	1
<i>Missing or undisclosed</i>	<i>20</i>	<i>29</i>	<i>6</i>	<i>2</i>

As a proxy for active contributors, the regional distribution of the eighty-eight Chinese Wikipedia administrators in 2012 also reflects a similar diversity: 27 (31%) are from mainland China, 20 (23%) from Hong Kong/Macau, 2 (2%) from Singapore, and 18 (20%) from Taiwan. In contrast, all disclosed power users of Baidu Baike come from mainland China. Figure 4-3 shows the world map for the year of 2012, first for Baidu Baike and then for Chinese Wikipedia. It clearly shows Baidu Baike's focus on mainland China and Chinese Wikipedia's spread across the world, including North America, Western Europe and Asia Pacific.

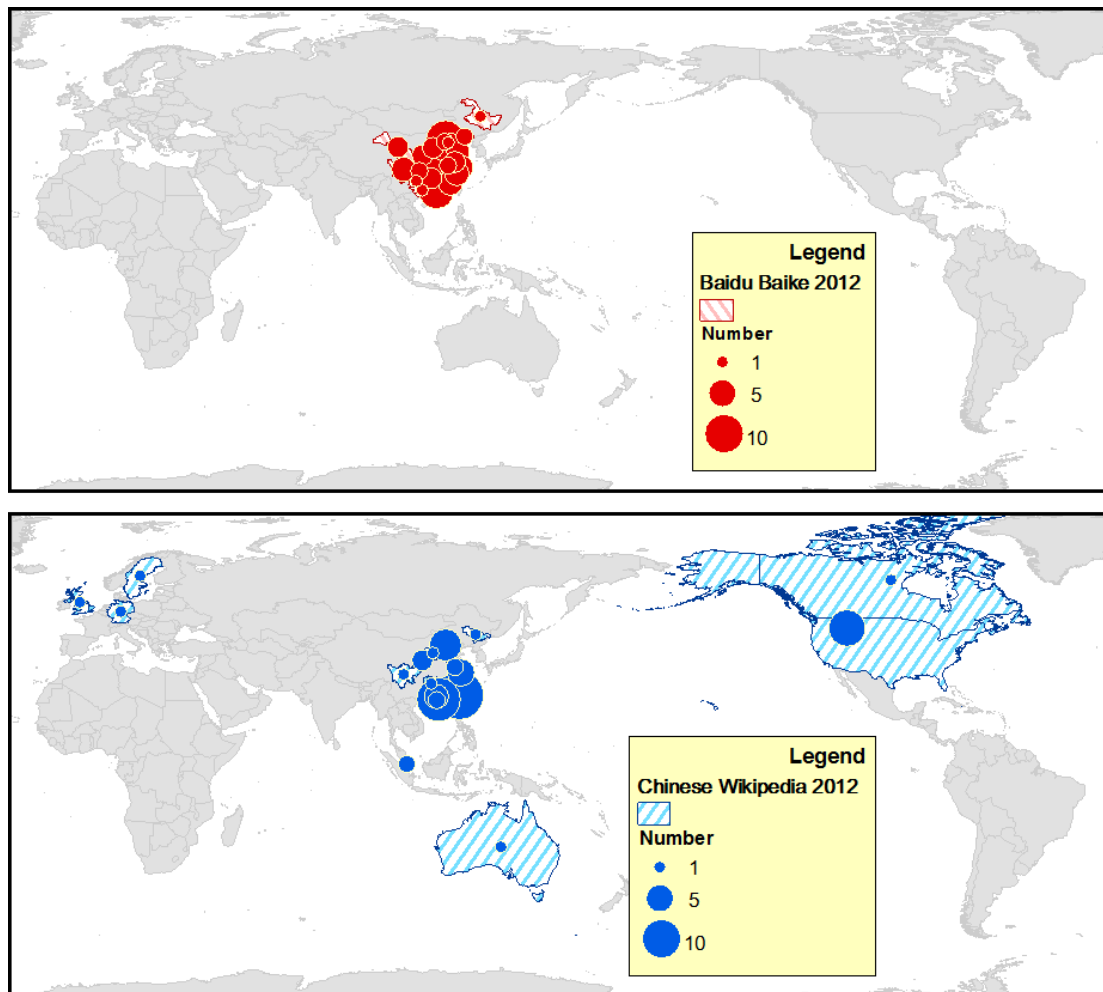


Figure 4-3. Distribution of power users of Baidu Baike and Chinese Wikipedia in 2012

Figure 4-4 further shows greater detail across the Chinese and East Asian regions. Chinese Wikipedia clearly has significant numbers of power users not in only the major cities of mainland China, such as Beijing, Guangdong and Shanghai, but also Hong Kong and Taiwan (and two in Singapore). In contrast, Baidu Baike's power users are bounded within mainland Chinese regions. The reason why Hong Kong and Taiwan constitute the major difference between the two encyclopaedias may be closely linked to the organization of local chapters. Both Wikimedia Taiwan and Wikimedia Hong Kong were founded and then recognized by the Wikimedia foundation around 2007 and 2008. Both have been active and organized enough in 2007 and 2013 to host Wikimania, an international event held annually since 2005 with all Wikipedians in the world.

This is in direct contrast to Wikimedia's efforts in mainland China. The founder of Wikimedia, Jimmy Wales, visited with China's censors in Beijing in 2008 and in 2009 with Hudong's CEO to discuss issues of censorship and Chinese trademark of Wikipedia, showing the tricky and adverse environment Chinese Wikipedia faces in order to grow in mainland China (Farrar, 2009).

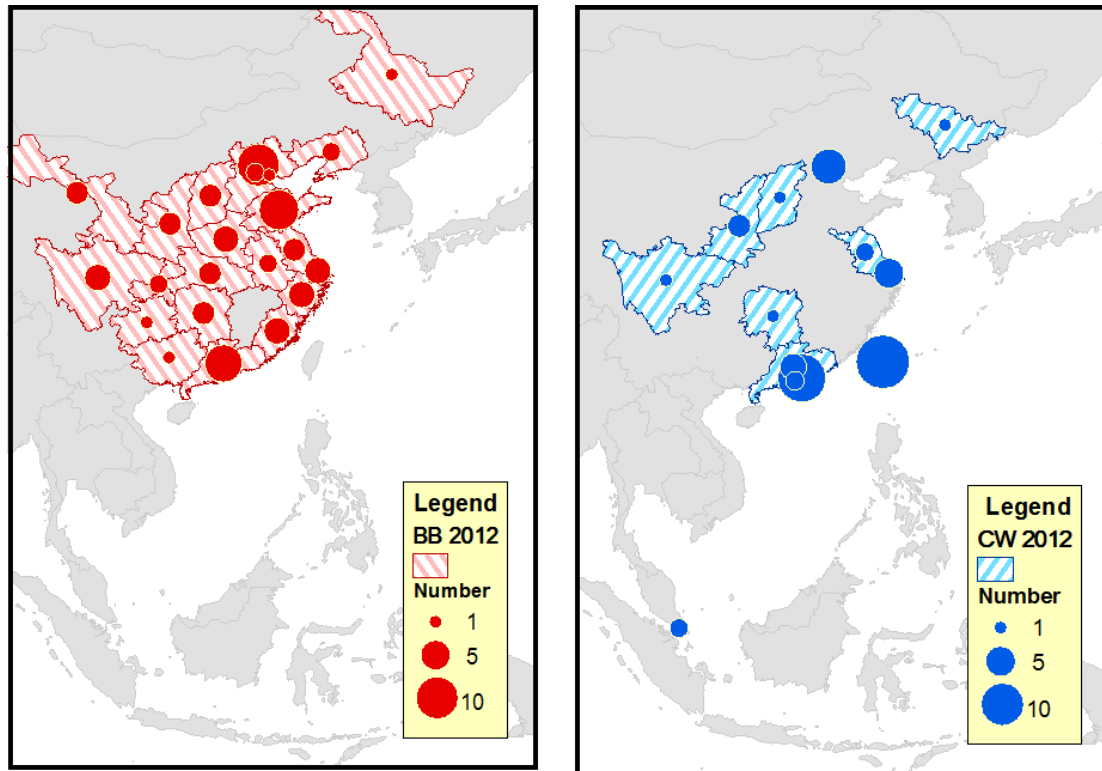


Figure 4-4. Distribution of power users of Baidu Baike (BB) and Chinese Wikipedia (CW) in 2012: East and South-East Asia

While some may interpret this distinction as a part of, or simply a reflection of, the geopolitical struggle between Beijing and Washington, I argue instead that it is the direct result of the way their distinct editorial processes attract different users when these are put into the larger geo-political contexts of Beijing, San Francisco, Hong Kong and Taiwan. In the next sub-section, I will present findings to show how the configuration of geolinguistic factors also contributes to the editorial processes.

4.1.3 How and why do the geolinguistic factors matter? Localization

(L10n), the process of adapting computer software or information systems for a

group of users usually defined by national boundaries or geolinguistic profiles (Liao, 2011; McKenna & Naftulin, 2000), can also be discussed under the rubric of contributing to the “internationalization mechanisms” of gatekeeping (Barzilai-Nahon, 2008). Since what constitutes as legitimate “localization” in a Chinese-language context is highly contested for media scholars (Huang, 2002), I have chosen the term “geolinguistic” factors and their configuration as a relatively more neutral term for this part of the geolinguistic analysis of media and Internet research (Liao & Petzold, 2010; Sinclair, 1996). The findings below provide plausible answers as to why the major differences in the geolinguistic profile of power users lies in the presence versus the absence of Hong Kong and Taiwan: this difference is mainly caused by practical editorial arrangements which involve a more complex Chinese history than the struggle between Beijing and Washington.

Language encoding: Baidu Baike excludes traditional Chinese users. By using simplified Chinese only with the encoding standard GB-2312, Baidu Baike effectively excludes users of traditional Chinese characters from contributing, whereas Chinese Wikipedia and other websites (including Baidu Japan) use the Unicode encoding standard, which can accommodate all the languages in the world.

The choice of language encoding is significant for both the wider context of Chinese-language Internet and the specific context of these two competing online encyclopaedias. As previously discussed in Chapter 3, the development of language encoding standards and the later widespread adoption of the Unicode encoding standard have been essential elements for digital networking technologies to be diffused in East Asian regions, including Japan and China. Indeed, before the all-inclusive Unicode was developed fully enough for wide adoption, traditional Chinese content was often encoded in Big5, a standard set by the computing industry in Taiwan and Hong Kong, and simplified Chinese content as encoded in GB-2312, a national standard set by authorities in Beijing.

These two standards are not compatible, and thus for both traditional and simplified Chinese content to coexist, the global Unicode standard or compatible Chinese national standard of GB18030 must be used. In other words, the wider context of Chinese-language Internet has experienced a history of competing incompatible language standards before Unicode and Unicode-compatible standards were widely adopted. Thus, the decision to use a certain language-encoding standard has important geolinguistic significance and thus cultural and political significance for Chinese language users.

For the specific context of the two competing online encyclopaedias, a brief historical review shows that Chinese Wikipedia was among the early adopters of Unicode, while Baidu Baike adopted the simplified-Chinese-only standard of GB-2312 even though Baidu the company has both capacity for and experience with using Unicode.

The early attempt and decision by Chinese Wikipedia to implement Unicode support can be traced back to the year 2002. In contrast, at the time of this writing in 2013, Baidu Baike uses only GB-2312 for all of its encyclopaedia article pages. For Chinese Wikipedia, the first documented user from mainland China, user:Mountain, used Unicode standard in 2002 to solve the issue of processing Chinese characters for the then-one-year-old global Wikipedia project, thereby creating the first Chinese-language article. It took around ten months for user:Mountain and other users to localize and translate the English-based software and documents, thereby providing fundamental components for Chinese Wikipedia at the domain name of zh.wikipedia.org. It was a significant cutting-edge move by the volunteers to globalize Wikipedia: this becomes clear when we consider that at the same time, in 2002, even popular commercial software such as Microsoft Outlook 2002 was struggling to provide Unicode support for multilingual solutions. Judging from my own computer science training based in Taiwan, Chinese Wikipedia has been a leading website in Unicode support in the Chinese-language digital world, where most of the other

websites use either Big5 or GB2312 (Liao, 2009b; Yunker, 2002). As a result, as early as 2002, Chinese Wikipedia had the capacity to process both types of Chinese writing scripts on the same page.

In contrast, when Baidu launched Baidu Baike in 2006, its choice of the GB-2312 encoding standard effectively precluded traditional Chinese users (mainly from Hong Kong and Taiwan) from participating. Since Baidu provides search services that include traditional Chinese content on the Web, it is unlikely that Baidu lacks the technical capacity to accommodate traditional Chinese characters in its encyclopaedia. In fact, it remains a telling decision that Baidu supports Unicode for its search engine and input method services for Japanese users (MasPoster, 2009), but not for Baidu Baike.

In addition to the choice of encoding standard, Chinese Wikipedia has established a conversion mechanism and editorial rules to ensure it does not preclude certain groups of Chinese-language users as Baidu Baike does, as will be detailed in the following paragraphs.

Language platform: Chinese Wikipedia develops a multiple script-writing system. Collaborative editing implies the integration of different writing systems used by contributors. Although the Wikipedia platform is able to support all the languages in the world when using Unicode standard, there are multiple separate language versions of Wikipedia, which are run by different communities. Thus, an issue arises for the Chinese-language content contributors: should there be a single version or multiple separate versions of Chinese Wikipedia? For the modern Chinese language, the gradual decision to merge the two originally separate versions (one traditional, the other simplified) into a single version was reached around 2003 or 2004 (enWP, 2013; zhWP, 2008a, 2009, 2012k). Users of different Chinese language-scripts now collaborate on the same version of articles.

After the decision to merge, a language platform emerged to accommodate simplified Chinese users and traditional Chinese users on an equal

footing. It mixed different contributed texts on the back stage but presented the overall content in several options on the front stage. This process began as a working automatic Chinese-Chinese conversion system in late 2004, which converted the content between the two language scripts: traditional Chinese and simplified Chinese. It subsequently evolved into a system that formally recognized four Chinese geolinguistic regions: Mainland-simplified, Hong Kong-Macau-traditional, Taiwan-orthodox, and Singapore-Malaysia-simplified. Basically, the platform's front end provided user-readers with several geolinguistic options, so as to present the content consistently in a fashion that matched readers' geolinguistic settings or choices. At its back end, since contributed texts came from individual editors' preferred language scripts, the diversity of scripts was preserved. Thus, it was common for an article to contain mixed Chinese-script content. Contributing and reading content were thus two different experiences: the front end minimized the chances of annoying user-readers with unfamiliar characters or terms; the back end showed basic respect for original user-contributors' choices. This way, the platform bridged the gap among Chinese language variants. It can be seen that the processability of language is intimately bound up with the editorial platform: editing is not just about content, but also about how the content in each script is expressed and represented. This is why the discussion of geolinguistic processability is crucial in the context of editorial processes.

Figure 4-5 shows a typical Chinese Wikipedia page where readers can choose among seven available options (from the fifth tab onwards): Non-conversion, Simplified (i.e., geolinguistic code of zh-hans), Traditional (zh-hant), Mainland-Simplified (zh-cn), Hong Kong-Macau-Traditional (zh-hk and zh-mo), Singapore/Malaysia-Simplified (zh-sg and zh-my), and Taiwan-Orthodox (zh-tw). Note that the actual entry title stored in the Wikipedia database is the term “出租车” (zh-cn). Nevertheless, the content will be presented differently according to the preference or choice of the user-readers: chūzūchē 出租车 (mainland

China), Jichéngchē 計程車(Taiwan), dīshì 的士(Hong Kong and Macau), or déshì 德士 (Singapore and Malaysia), as displayed in Figure 4-5. For user-readers, a familiar linguistic environment is thus provided, and they also have the option to switch between formats..



Figure 4-5. Screenshot of the entry Taxi without the popup box.

The experience of dealing with multiple scripts in the same language version of Wikipedia has been shared and discussed globally since the Wikimania conference in 2005 in Frankfurt, the global annual event for Wiki-projects including Wikipedia, where it has been reported that Arabic, Malaysian and Indonesian language users have also expressed interest in this issue (zhWP, 2008b). These Chinese-language users who helped to develop the solution, mostly from mainland China, have contributed their computer codes back to the free software project on which Wikipedia has been based. At the time of this writing, there are several ongoing implementations and discussions on adopting this platform for other language projects, for languages such as Serbian, Kazakh, Kurdish, Tajik, Uzbek, Gan Chinese, Kyrgyz, Uyghur, Chechen, etc., which use more than one writing script or system (Wikimedia Foundation, 2012b). This successful story mirrors that of Unicode, whereby the global development of

processing Chinese characters leads the way to processing other languages (Burgmer, 2009; Wong et al., 2009). In sum, the language platform developed by users of Chinese Wikipedia, which supports multiple writing scripts with inter-script conversion capacity, has become a model of emulation and adoption for other language versions of Wikipedia.

As we have previously seen, Chinese characters are bound up with questions about Chinese modernity, the Cold War and ongoing geolinguistic differences across Chinese-speaking regions. Such tensions are alleviated by the multi-script platform of Chinese Wikipedia. No Chinese characters are “better” or “more correct” than the others are. I have reported elsewhere that it is arguably one of the most advanced platforms among Chinese-language websites in tackling the inter-script conversion thanks to the fact that conversion mapping tables are constantly updated by the user-contributors of Chinese Wikipedia (Liao, 2009b). While other types of conversion software, such as the one provided with the Chinese version of Microsoft Office, are limited to the pre-determined orthographic and lexical conversion tables, the conversion tables maintained in Chinese Wikipedia evolve through ongoing user contributions. The language platform is more than just the usual simplified-complex conversion; it feeds on the collective intelligence from user-contributors everywhere, making it possible for them to work together on the same article. No specific script is treated as superior to any other.

To sum up, the two encyclopaedias deal with geolinguistic variants of Chinese differently: the language encoding and platform of Baidu Baike implicitly accepts only the simplified Chinese characters, whereas those of Chinese Wikipedia explicitly accommodate various Chinese variants on equal footings.

Regional considerations: editorial policies. To further maintain a balance so as not to favour one writing system over the other, several editorial policies and norms have been adjusted to take regional differences into account. Among

these, the “Avoid-Region-Centrism Policy” (bìmiǎn dìyù zhōngxīn 避免地域中心) is arguably the most significant (User:Lorenzarius & other contributors, 2010). It essentially mandates that China-centric, Han-centric and Chinese-centric statements should be avoided by emphasizing that Chinese Wikipedia is a project written in the Chinese language, and not just a project of, by and for Chinese people. Other policies, such as “Wikipedia: Vandalism”, also take regional variants into account because naming, translation, transliteration and writing conventions may differ from region to region (zhWP, 2012i, 2012j). For instance, if a user-contributor changes others’ contributions from one language script to another, it is counted as an act of vandalism. Inclusiveness is thus ensured.

In contrast, the absence of such geolinguistic considerations in Baidu Baike is expected to discourage participation by user-contributors from Hong Kong and Taiwan. This helps to explain the aforementioned differences in the geolinguistic distribution of power users of Baidu Baike and Chinese Wikipedia. Though actual data on the geolinguistic profiles of all user-contributors of Baidu Baike and Chinese Wikipedia remains a research challenge, the data on the top 100 user-contributors (BB, 2012c; User:Emijrp, 2012) shows consistently that non-Mainland user-contributors are rare in Baidu Baike and user-contributors in Chinese Wikipedia are mixed. Thus, the links are clear between the geolinguistic outcome and the ways in which Chinese characters are configured with encoding standards by Web platforms: Baidu Baike’s simplified-Chinese-only configuration precludes participation by traditional Chinese users, which happen to be the majority of users in Taiwan and Hong Kong, while Chinese Wikipedia’s multi-script support allows contribution from different regions.

The contrast between the two encyclopaedias can thus be better understood from the varied perspectives ranging from Mainland-centric to transnational Chineseness, instead of the perspective of the geo-political struggle between Beijing and Washington. As an ongoing discussion in Chinese Studies, the transnational aspects of Chinese people, media and political institutions

often surrounding the term “Greater China” after the Cold War are essential to understand the modern dynamics of Chinese nationalism, regionalism and transnationalism (Callahan, 2005; Chan, 2009; Guo & Guo, 2010; Hughes, 2004; C. C. Lee, 2001; Rex Li, 1997; Mengin, 2004; Sum, 2004; K. C. C. Yang, 2007). Instead of accepting the notion that Baidu Baike is Chinese and Chinese Wikipedia is American or Western, the evidence presented above suggests two notions of Chineseness: Baidu Baike embodies the Beijing or mainland-centric point of view, whereas Chinese Wikipedia reflects an instance of multi-centric integration of Chinese transnationalism.

Since it has been reported that Chinese authorities in Beijing have rejected the adoption of the term “Great China” on the grounds that it will put Hong Kong and Taiwan on equal footings with mainland China (H. Y. He, 2001), it can be speculated that the way Chinese Wikipedia recognizes four separate but equal geolinguistic regions will not be well-liked by Chinese authorities. Here, the ambiguity of what counts as transnational and what counts as national is evident. Chinese authorities continue to claim that Hong Kong and Taiwan are “part of China” and thus belong to the Chinese nation, whereas Hong Kong and Taiwan seem to have gained and maintained some equal status with mainland China in cultural-political aspects. For example, their preferred language script (orthodox or traditional Chinese) is on par with Beijing’s simplified Chinese. The distinction of their political and media systems from mainland China’s is not highlighted (even if it is not overlooked). Thus, the contrast in the two websites’ respective support for users from Hong Kong and Taiwan can be interpreted as follows: While self-governed Chinese Wikipedia users have reached a working consensus to put these four major geolinguistic regions (including the fourth, Singapore/Malaysia) on equal footings, Baidu Baike’s hosting organization has decided not to do so, even though it offers Unicode support for Japanese-language search engine users in Japan. Here, online writing systems, or the way in which language materials can be solicited and processed by digital network

technologies, are shown to be shaping and reshaping national and transnational dynamics online, an area of research which both Chinese studies and Internet studies need to systematically explore. Geolinguistic arrangements matter, as they constitute the basis for system and content localization.

4.1.4 Discussion: different editorial gatekeeping processes. With different editorial priorities, hierarchies and geo-linguistic arrangements, Baidu Baike and Chinese Wikipedia filter and solicit users' contribution differently, thereby showing different features of editorial gatekeeping. Though both claim to provide open and free (as in freedom) Chinese-language content serving all users, the content and context are different regarding what types of "openness" and "freedom" are provided for which group of "Chinese-language" Internet users.

First, two gatekeeping patterns differ in keeping and removing new edits: Chinese Wikipedia is much stricter than Baidu Baike in preventing explicit marketing, self-promotion and copyright infringement activities. Baidu Baike is much stricter and less transparent than Chinese Wikipedia is in processing content that are culturally- and/or politically-sensitive to Beijing. Second, Baidu Baike disregards edits from non-mainland Chinese characters by design, whereas Chinese Wikipedia integrates them on par with mainland simplified Chinese. Third, during its launch, Baidu Baike copied and pasted content *en masse* from Chinese Wikipedia selectively, avoiding political content, and converted Chinese characters while ignoring copyright/copyleft issues.

All these differences indicate two distinct approaches: One defines the centrality or the centre of gravity with a political core –mainland China, and another integrates the diversity of contribution. Ironically for Beijing which seeks political integration of Hong Kong and Taiwan, by alienating if not outright excluding user participation outside the core, the first approach reinforced the boundary between the core and non-core (outside mainland China). By avoiding region-centric bias, the second approach seems to have relatively more integrative effects, particularly for users from Hong Kong,

Taiwan and mainland China. Altogether, the differences in the aggregation, filtration and accreditation of user contribution constitute two gatekeeping processes that handle user contribution differently.

The media environments and user experiences influences how user contribution is processed online. Online communicative spaces, no matter how new or universal, are shaped by its participants' prior and local experiences. Baidu Baike's internal communicative space, despite its claim in open participation, excludes specific geo-linguistic variants of Chinese language and certain aspects of editorial freedom. This is in direct contrast to Chinese Wikipedia, where users are given significantly more autonomy to communicate and collaborate in a more open and inclusive fashion. Thus, Baidu Baike's communicative space seems to fall in line with Beijing's actions and attitudes towards Internet users within mainland China, whereas that of Chinese Wikipedia reflects both the initial consensus and ongoing challenges in integrating participants' experiences and attitudes towards diverse modern Chinese media environments.

While the core idea of "collaborative filtering and accreditation" in deciding "whom to trust and whose words to question" remains useful (Benkler, 2006, p. 465), the authoritarian state-led or -sanctioned measures may shape or even alter such collaborative filtration, thereby producing different gatekeeping processes that define information orders and communicative spaces. Collaborative filtering and accreditation thus do not necessarily reproduce liberal democratic norms, and it can be used to re-establish other social orders (a set of social structures and practices that reinforce or change social norms). For example, Baidu Baike's communicative space share the same extent and compatible order with mainland China, the main jurisdiction of the People's Republic of China. Chinese Wikipedia managed to integrate varied practices and expectations of users across regions, with a set of general norms of no censorship

and respects for copyright. The overall editorial gatekeeping outcome thus depends on the host, design, participants of user-generated encyclopaedias.

This chapter, then, must address the question how the authoritarian constraints, imposed on the Internet segment inside mainland China, may produce different gatekeeping effects. After analysing the geopolitical and geolinguistic factors internal to the two website, the next half of the chapter will continue to examine the 2005-2008 filtering of Chinese Wikipedia in mainland China as arguably the most important constraint imposed by Beijing.

4.2 Gatekeeping mainland Chinese users from 2005–2008

At this point, we can examine one period of the blocking of Chinese Wikipedia, from 2005 to 2008, which, as we shall see, was a crucial period for the take-off of the two encyclopaedias. User-contributed encyclopaedias need users, and Beijing's Internet filtering of Chinese Wikipedia is expected to inhibit its editorial development in mainland China. To assess the impact of filtering (or lack thereof), this section will examine the historical context of Internet diffusion in relation to the development of Beijing's filtering and censorship regime, so as to contextualize the development of these two encyclopaedia websites within the context of Internet diffusions and filtering.

4.2.1 Background: internet filtering and online competition. On October 19, 2005, Wikipedia began to suffer its third period of being blocked in mainland China: the Chinese government, using various Internet filtering technologies, prevented users in mainland China from visiting the servers of Wikipedia (X. Zhang & Zhu, 2011). Table 4-3 lists all main periods during which Wikipedia was blocked in mainland China. Unlike the two previous periods, which were brief, the third one began a series of blocked periods of longer durations with brief unblocked periods in between. In other words, being blocked became “normal” for Wikipedia in mainland China after 2005.

Table 4-3

Wikipedia in mainland China: blocked periods before 2008

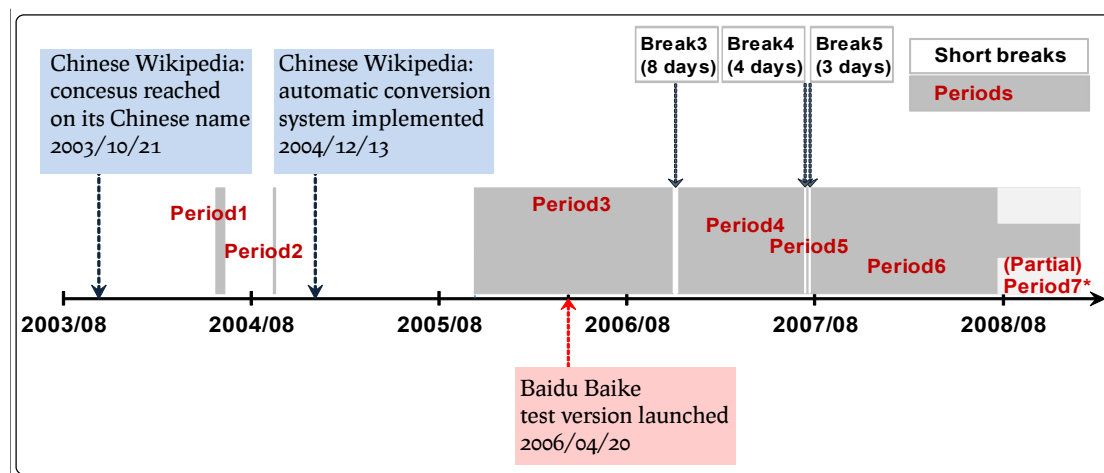
Periods	Begin dates	Days	Speculated causes
Period1	03 June 2004	18	15th anniversary of Tiananmen Square
Period2	23 September 2004	4	(1) Takedown of the YTHT Bulletin Board System (BBS) or (2) Objections to Yaohua2000's requests for adminship of Chinese Wikipedia
Period3	19 October 2005	386	Publication of the white paper "Building of Political Democracy in China" by China's State Council Information Office
Period4	17 November 2006	246	Upgrade of the Great Firewall (GFW) of China, or China's Internet filtering system
Period5	25 July 2007	5	Unknown (a short period inbetween two main periods)
Period6	02 August 2007	364	Unblocked just before the eve of the Beijing 2008 Olympic Games
Period7*	31 July 2008	-	Partial filtering: certain entry articles remained blocked

Note. *Some articles remained blocked after the eve of the 2008 Beijing Olympic Games

The third block in 2005 is significant because it reduced the growth rate of Chinese Wikipedia contributors, which not only curbed the growth of Chinese Wikipedia content but also triggered discouraging social effects on those who were not blocked (typically traditional-Chinese users who reside outside mainland China): their contributions decreased by 42.8% even though they were not affected by the block (X. Zhang & Zhu, 2011).

The timing of the launch of Baidu Baike. The timeline in Figure 4-6 further shows the significance of the third block in context. It marks a breaking point between two periods: mostly unblocked before October 2005 and mostly blocked thereafter till the eve of the 2008 Beijing Olympic Games. It is worth noting that two milestones for Chinese Wikipedia occur before 2005: the

decision to merge the simplified and traditional Chinese versions, and the implementation of its language platform of automatic conversion. For Baidu Baike, the major milestone is its launch in April 2006, about six months after the third block of Chinese Wikipedia. After 2005, mainland Chinese users encountered difficulties in using Wikipedia until the eve of the Beijing Olympic Games in 2008.



Note. * Some articles remained blocked during the Period7

Figure 4-6. Timeline of major events for Baidu Baike and Chinese Wikipedia

The fact that these events coincide has made many observers and users of Chinese Wikipedia speculate that there may have been coordinated efforts by the Chinese government and Baidu to transfer users to Baidu Baike, be it for political or economic motivations (Long, 2006; Old Geng, 2010; M.-J. Yang, 2006a, 2006b). The Baidu founder and CEO Robin Li stated that he was unaware of the block while acknowledging that Baidu Baike was based on the approach pioneered by Chinese Wikipedia (Dickie, 2006). Indeed, as a Chinese-language user-generated encyclopaedia, Chinese Wikipedia was the first mover. Thus, coincidental or not, it is necessary to assess the impact of such Internet filtering on the historical development of the two sites.

Competition between Chinese Wikipedia and Baidu Baike (in mainland China). In their competition for users, Chinese Wikipedia faced consistent blocks from 2005 to 2008. Some of those who support or sympathize with Baidu Baike argue that the impact of these blocks has been overstated and that Baidu Baike was able to gain many more users from mainland China based on its own merits. They believe that Baidu Baike suits (mainland) China better due to its “Chinese characteristics” or Chinese culture-friendly practices (Huynh, 2006; Suo, 2007; X.-Y. Lee & Luo, 2009). The supporters and sympathizers of Chinese Wikipedia tend to believe otherwise, that were it not for the filtering of Wikipedia in mainland China, Chinese Wikipedia’s popularity in mainland China would be much higher (Lih, 2006; M.-J. Yang, 2006b). As mentioned earlier, a group of Chinese Wikipedia editors who supported the self-censorship proposals reportedly migrated to Baidu Baike (Chu, 2013). A more nuanced explanation is thus needed to consider how the two websites have recruited their contributors across Chinese-speaking regions while facing Beijing’s filtering and censorship regime.

Beijing’s filtering and censorship regime constitutes an example of authoritarian gatekeeping mechanisms to “restrict or channel access” to the networks (Barzilai-Nahon, 2008, p. 1499). If one shifts one’s focus from information to users, then effectively the regime also constitutes a gate for websites to access mainland Chinese users. How has the regime’s gatekeeping of mainland Chinese users figured in the competition between Baidu Baike and Chinese Wikipedia? This question leads to the more general question of the regime’s impact on mainland Chinese users’ behaviour.

Internet “access blockage” and user behaviour. Based on the clustering analysis of aggregated traffic data of the top 1000 most visited websites in 2012, Taneja & Wu (2013) argued provocatively that the decade-long “access blockage” had insignificant impact on users’ browsing behaviour. By “access blockage”, Taneja & Wu (2013) seemed to mean the censorship/filtering regime implemented by certain states to block access to certain websites, with the Great

Firewall of China as the main example. Hereafter I only use the term “access blockage” when presenting their arguments.

As shown in Figure 4-7, they argue that the unfulfilled access in 2012 is insignificantly small, and most enacted user preferences were satisfied by the subset of allowed websites. They speculate that around 2001, when the imposition of access blockage began, the unfulfilled access was already relatively small. This speculation was based on a version of the geocultural thesis that Chinese-language content, or “culturally proximate content”, is largely allowed, and thus many of the blocked websites would not be preferred by mainland Chinese users anyway due to their foreign language content or, even if the sites are in Chinese, unpleasant “ideological overtones”(p. 23). The historical difference, Taneja & Wu argue, lies in the expansion of culturally proximate websites hosted in mainland China. Were the access blockage gone now, they contend, the once-unfulfilled user preferences would account for little change to the overall enacted user preferences to the allowed websites.

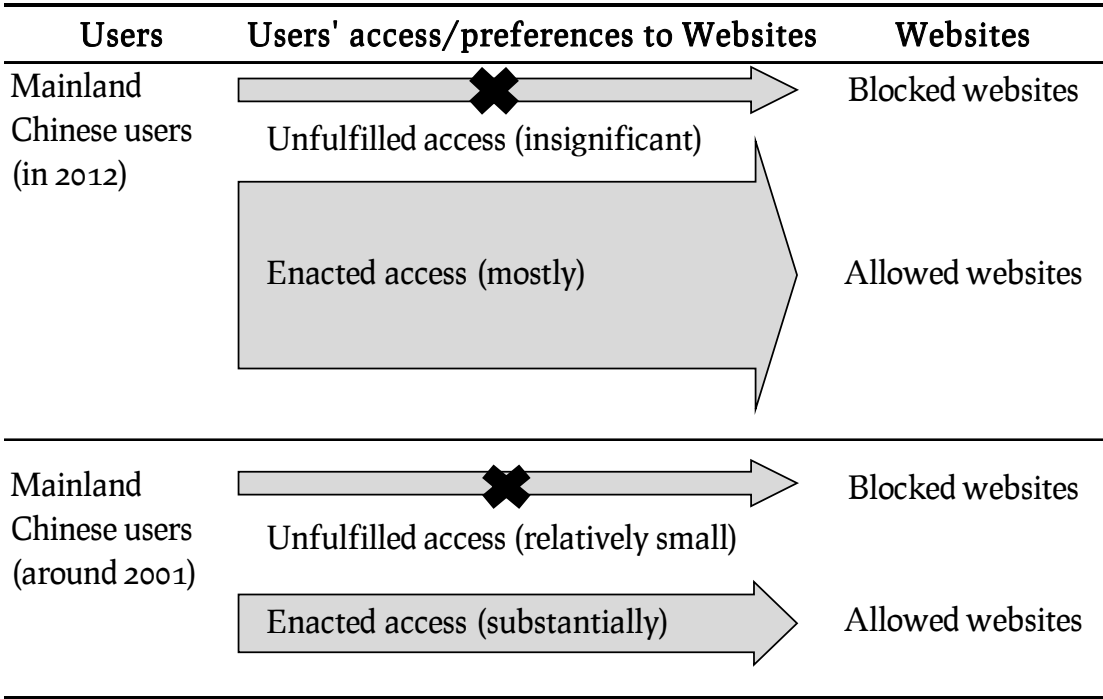


Figure 4-7. Taneja & Wu’s “Internet Access Blockage and User Behavior”

This argument by Taneja & Wu would obviously have major implication for the arguments in this thesis. However, while their approach has merit in correcting against what are sometimes seen as exaggerated impacts of Beijing's censorship/filtering regime on use patterns, they run the risk of underestimating these impacts by conflating data about Chinese-language users within a single Chinese-language cluster. What Taneja & Wu found from the traffic data of the top 1000 most visited websites is essentially the aggregated pattern of anonymous traffic from all over the world. On this basis, it can be expected, given the thesis of a single geocultural cluster of Chinese-language users, that several frequently visited Chinese culturally proximate websites would be widely shared. However, Taneja & Wu failed to disaggregate the data according to different Chinese-speaking regions, or at least to differentiate between those inside mainland China (impacted by the Internet filtering) and those outside (not impacted by the Internet filtering) to determine if the access blockage has an impact *within* the Chinese-language cluster. To be precise, what Taneja & Wu found is that the language factor is more potent than the "access blockage" in clustering Chinese-language users' traffic, but their findings say nothing about whether and how subclusters *within* the larger Chinese-language cluster have been impacted by the censorship/filtering regime. In other words, they concluded too quickly that Beijing's Internet censorship/filtering did not or had not isolated such clusters of Chinese culturally defined markets from the world. As I have pointed out earlier, the bridging websites they found that connect mainland China to the world are the B2B (business-to-business), mostly the English-language website Alibaba.com, some websites in Hong Kong and Taiwan, and Chinese Wikipedia. Thus, it is very likely that when the data contributed by business transactions and Hong Kong/Taiwan users are removed, the remaining data points of mainland Chinese use patterns are indeed isolated by the censorship/filtering regime of China.

What is particularly telling is their interpretation of Chinese Wikipedia. On one hand, they believe that the “access blockage” has little impact on users’ behaviour, but on the other, they suggest that if the “access blockage” were gone completely “[a]nother change we would likely see is that the Chinese Wikipedia, currently at the periphery due to blockage, would move to the centre of the cluster similar to the positions occupied by other language Wikipedias within their respective CDMs [culturally-defined markets]” (Taneja & Wu, 2013, p. 23). Such speculation goes directly against their main hypothesis because it implies a substantial impact of the “access blockage” on users’ choice between Baidu Baike (allowed websites) and Chinese Wikipedia (blocked websites). More research is therefore needed to assess the contrasting impact of the filtering/censorship regime on users inside mainland China versus users outside.

To do so, I propose an explanatory framework combining elements of a number of diffusion theories that draw together previous findings on the Internet development in China and East Asia. By analysing longitudinal data that includes mainland Chinese, Chinese-speaking and East Asian regions, the analysis should provide more substantial and detailed findings regarding the impact of the Internet censorship/filtering than the work of Taneja & Wu.

4.2.2 An explanatory framework. To assess the impact, an explanatory framework is proposed to understand the socio-temporal and socio-spatial dynamics of the diffusion of Internet-based innovations such as user-generated encyclopaedias. Since using the Internet is an essential prerequisite for using user-generated encyclopaedias, the diffusion of the latter depends on the diffusion of the former. This framework allows researchers to use publicly available Internet diffusion rates (also known as the Internet penetration rates) to explain how such diffusion patterns may promote or hinder Internet-based innovations.

A selective review. Previous attempts at understanding Internet diffusion in mainland China have used the diffusion of innovations theory to highlight the

uneven geographic and social patterns of Internet diffusion (Foster & Goodman, 2000; Qiu, 2005; J. J. H. Zhu & Wang, 2005). However, there has been no major follow-up research since 2005 and thus we lack a timely update on the historical development over the last several years. In 2006, the Internet diffusion rate for mainland China was still low (about 10%) though its total national Internet population was already the second largest in the world (Tai, 2006). The gap after 2005 needs an update, which is also essential for the current research since Chinese Wikipedia has been mainly blocked since 2005.

Indeed, within a specific population, the percentages of Internet users are important indicators for the development of user-generated encyclopaedias. Using the Internet and using Wikipedia were new activities for the majority of Chinese people around 2005. For the purpose of this research, I categorized diffusion research from various disciplines into two rough categories: socio-temporal diffusion and socio-spatial diffusion. The socio-spatial perspective comes mostly from anthropology and human and political geography, with major concepts such as “cultural diffusion”, “spatial diffusion”, etc., that deal with topics ranging from the spread of civilization to intercultural diffusion (Blaikie, 1978; Cliff, Pred, & Hagerstrand, 1992; Fløysand & Jakobsen, 2011; Webber, Lutz, & Brown, 2006). The socio-temporal perspective comes largely from management studies (Easingwood, Mahajan, & Muller, 1983; Kalish & Lilien, 1986; Mahajan, Muller, & Bass, 1990; Mcgrath & Zell, 2001; Midgley & Dowling, 1978), sociology and communications studies (Bruce & Yearley, 2006; E. M. Rogers, 2003; Tarde, 1962, 2010). This perspective addresses relatively more modern concerns such as diffusion of innovations, strategic media and marketing campaigns. Despite these different academic backgrounds, innovation diffusion research seems to converge on the topic of Internet adoption and/or use as social, geographical, and strategic diffusion processes (Beilock & Dimitrova, 2003; Dutta, Dutton, & Law, 2011; Dutton & Eynon, 2009; Kiiski & Pohjola, 2002; Loch, Straub, & Kamel, 2003; Morris & Ogan, 2002; Prescott, 1997; Press et al., 1998; Press, Foster, Wolcott, &

McHenry, 2003; Wolcott, Press, Mchenry, Goodman, & Foster, 2001). To understand the Internet diffusion history in mainland China in relation to other major Chinese-speaking regions, I use both perspectives as follows.

Spatial and temporal diffusion. Spatial and temporal diffusion is concerned with intercultural diffusion, the origin and spread of civilization and similar topics where spatial factors dominate (Strang & Tuma, 1993). Often researchers need to define or identify some single elements of practice in a culture - beliefs, institutions, language and technology - in order to study how they spread from one origin (also known as the “cultural hearth”) to another region. These concepts from spatial and temporal diffusion will be useful in framing some of the major geo-political dynamics of Internet development in China and East Asia.

Diffusion of the cultural traits of the Internet: two conventional views. In applying diffusion theory to the Chinese and East Asian regions, there are two polarized views – which must be overcome – that see the development of the Internet as part of an ongoing geo-political struggle between China and the US. These two views focus on different directions and content: One view sees the Internet as a strategic instrument to spread American values for the “peaceful evolution” (héping yǎnbiàn 和平演變) of China and that measures must therefore be taken to save China from this Western plot in installing a new regime that is susceptible to “foreign influences” (Barmé, 2010; Chase, 2011; Zuo, 2006). On this view, any foreign-originated Internet developments should be viewed suspiciously, thereby blocking the unwanted diffusion of values, practices and ideas into China. On the other end of ideological spectrum, the other view, common among pro-Internet freedom observers, holds that the imposed monitoring and censorship regimes not only suppress dissidents, but also cuts China off from the world, thereby creating a “Chinese local-area network” that is docile to the Chinese authorities (China Digital Times, 2011; Human Rights Watch, 2006). This conventional geo-political opposition, whose

rhetoical origin can be traced back to Mao Zedong's response to John Foster Dulles's Peaceful Evolution strategy (R. Ong, 2007; Zhai, 2009), sees the Internet development in China and East Asia as one of the remaining battlefields of ideology and geopolitics.

This discussion puts spatial-temporal diffusion into a geopolitical context, emphasizing the key importance of the origins and directions of diffusion. Seen from the spatial-temporal diffusion perspective, both views can be interpreted as the spread of different values from different “cultural hearths”. One sees the US and American or Western values as the foundation of Internet culture, and thus the diffusion of the Internet into East Asia and China suggests the spread of American or Western values into the region. Another sees mainland China as the centre for spreading Chinese values and powers in East Asia, including its Internet practices. Either way, the development of the Internet in the region becomes part of the grand Beijing-Washington chess game, spreading the cultural traits of American freedom versus those of Chinese “*weiwēn*” (wéiwěn 维稳, literally meaning stability maintenance) between two states, two major geopolitical powers, or even two civilizations.

Diffusion of the cultural traits of the Chinese Internet: an alternative view. Here I propose an alternative, more updated and nuanced view, which reframes the same Internet development within a slightly different spatial-temporal perspective. Informed by a previous historical analysis of the modern Chinese press (S. R. MacKinnon, 1997), I argue that the Internet is one of the major technological forms of media and communication for the Chinese-speaking population of the world, and that different types of spatial-temporal diffusion of media technologies and practices, no matter whether they are owned or operated by domestic or foreign entities, constitute communicative spaces that may promote the exchanges of information, opinions and ideas across boundaries. Thus, this alternative view recognizes the role of several regional centres in processing Chinese-language content, notably the coastal Chinese regions and

regions such as Hong Kong and Taiwan. This view is supported by previous research about Internet diffusion in mainland China which shows that, directly or indirectly, the uneven geographic diffusion within mainland China is linked to the historically significant roles of Hong Kong and Taiwan (Foster & Goodman, 2000; Qiu, 2005; F. B. Tan, Corbett, & Wong, 1999; J. J. H. Zhu & Wang, 2005).

The fact that mainland Chinese Internet users have been blocked from using certain websites hosted outside mainland China, including Chinese Wikipedia, suggests a cultural-political phenomena in opposition to the diffusion (or cultural thickening) from regions of Hong Kong and Taiwan as the cultural hearth of “the outside”, rather than the US. Here, the concept of diffusion contributes to our understanding of cultural thickening patterns: First, researchers can better observe the direction of diffusion as part of the cultural thickening patterns. Second, researchers can better analyse the temporal patterns with time series of diffusion rate data. Third, diffusion theory provides a conceptual dichotomy to the Internet filtering, a feature that is useful for researchers in assessing the impact of the Internet censorship/filtering regime. Thus, contrasting the Internet diffusion rates inside and outside mainland China can indicate the discriminatory impact of the Internet filtering on Internet diffusion patterns (and thus on cultural thickening patterns). In this vein, the next section will introduce another perspective on diffusion research, socio-temporal diffusion.

Socio-temporal diffusion. There are a number of additional useful concepts that taken from the social theory of innovation diffusion (Bruce & Yearley, 2006; E. M. Rogers, 2003; Tarde, 1962, 2010): the S-curve model, its application in business strategy, and the cosmopolitan-local spectrum. This body of work provides the basics for interpreting the data of Internet diffusion rates.

S-shaped curve. To illustrate the social dynamics of diffusion, members of a society (or a market) are divided into five categories, based on the length of time required for an individual member to adopt innovations (E. M. Rogers,

2003). Figure 4-8 shows a typical model of innovation diffusion curves, where members are categorized into groups of innovators, early adopters, early majority, late majority and laggards, depending on the time required for adoption on the horizontal axis. Based on various empirical findings, the diffusion curve is found to be close to a normal distribution (the bell-shape curve), so that threshold values can be approximated based on descriptive statistic values of mean and standard deviations. Further, the cumulative diffusion curve is an italicized 'S-shape' curve, as shown by the second dashed curve in Figure 4-8. The S-shape curve indicates an overall slow-fast-slow growth temporal pattern. The varying diffusion speeds are indicated by the S-shaped curve at different points in time: the adoption of new ideas and practices is slow in the beginning, then rapid if adopted by the majority groups, and then slow again with the laggards.

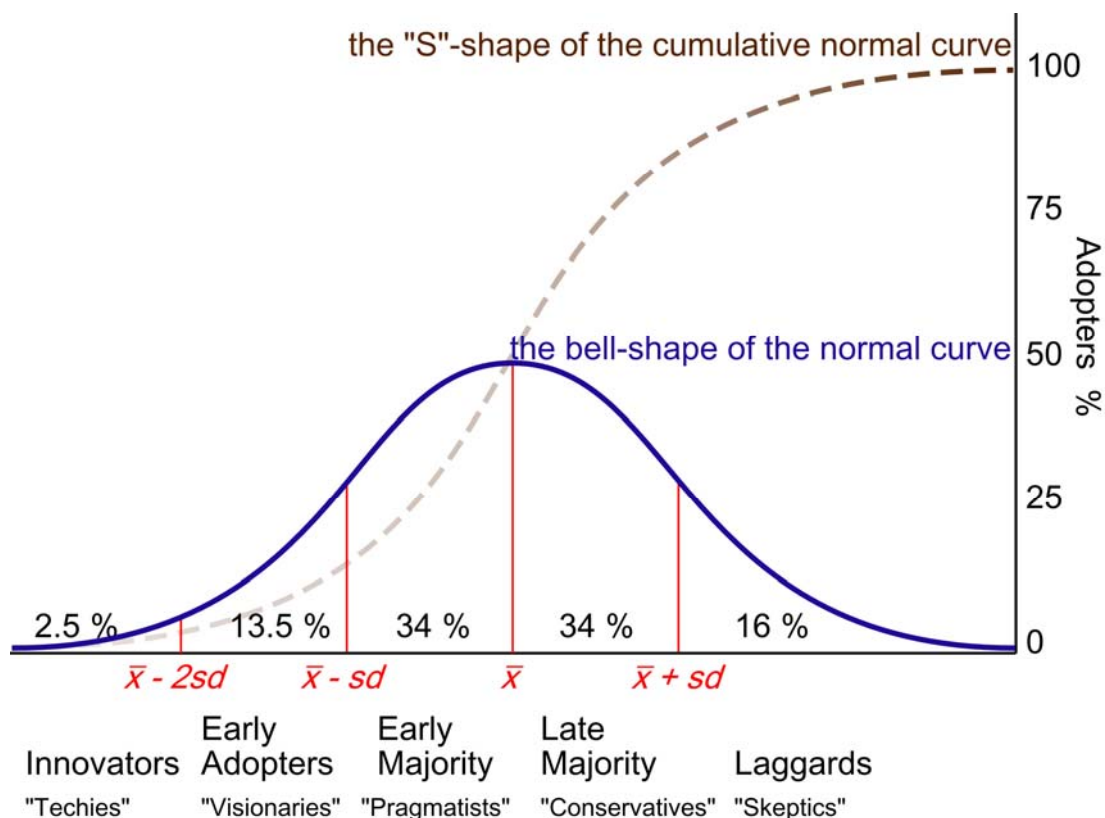


Figure 4-8. The bell-shape curve and the S-shape curve of innovation diffusions

Strategic use. The slow-fast-slow pattern of diffusion implies that the

critical moment occurs when the process begins to expand from the early adopters to the early majority. Before the majority adopts a piece of innovation, the early adopters must persuade the early majority to do so. The critical threshold constitutes the first major test or “chasm” for wider adoption, a major idea in the strategic business use of diffusion theory, especially in marketing technology products to mainstream markets (Moore, 1999). This strategic use is common in high-tech industry and in MBA programs (Iskold, 2007). The theory shapes and reinforces the perceived wisdom that start-up companies, as the “innovators”, need venture capitalists’ marketing to reach and secure the majority. This is central to success if the early adopters are to succeed in playing the role of opinion leaders in persuading the early majority. The vision of “early adopters” as innovators must be put to pragmatic use by the early majority. Whether the majority, and especially the early majority users, can be persuaded becomes the key for wider adoption in the social system.

Cosmopolitan-local spectrum. Various factors influence the rates and paths of diffusion processes (E. M. Rogers, 2003), and among them is the willingness to adopt new products and ideas which is expected to correlate with the cosmopolitan-local spectrum: innovators and early adopters are usually more cosmopolitan whereas other groups tend to be locales with different tendencies to accept (or reject) new ideas.

The S-curve model was applied to forecast the Chinese-language television (J. H. Zhu, 1997). Also, the previous findings on the diffusion of television and telephony in China suggest the difference across the country is only a function of time, with main cities a few years ahead than other regions (J. H. Zhu & Zhou, 2002). Thus, in the wider Chinese context, the cosmopolitan-local spectrum of innovation diffusion has both social and spatial dimensions. The coastal urban areas of China tend to be more cosmopolitan, and thus more likely to adopt new (and sometimes foreign) ideas and products. Thus, it is expected that people in the Chinese coastal regions started using the Internet and

user-generated encyclopaedias sooner. Regions such as Hong Kong and Taiwan can also be considered cosmopolitan because their popular cultures were “very influential in China” (Chan, 2009, p. 33).

Empirical questions. Hence, it can be asked: Does the growth of the number of Internet users in mainland China match the slow-fast-slow S-curve? When did regions in China first begin to see a fast-growing number of Internet users? What does the outcome of this diffusion mean for the two user-generated encyclopaedias? Moreover, how does Internet filtering shift the dynamics of user-contributed encyclopaedias, which require Internet users for their success? Counterfactually, what if Chinese Wikipedia were not blocked, or had been blocked at different historical moments (say, after Baidu Baike’s launch or later)?

4.2.3 Data selection, a natural experiment and comparative methods. To answer these questions, I compiled a longitudinal dataset of Internet diffusion rates for both 17 different East Asian regions and 31 disaggregated Chinese regions at the first administrative level, as shown in Figure 4-9. The dataset included publicly available data from 1990 to 2011, provided by both Chinese and international authorities. Altogether, Figure 4-9 shows 48 curves (each representing a region). Many of these curves fit the S-shaped cumulative curve as shown in Figure 4-8, confirming that Internet diffusion largely followed the predicted slow-fast-slow growth pattern. This selection allowed a more updated, detailed and comprehensive picture of Internet diffusion in the region.

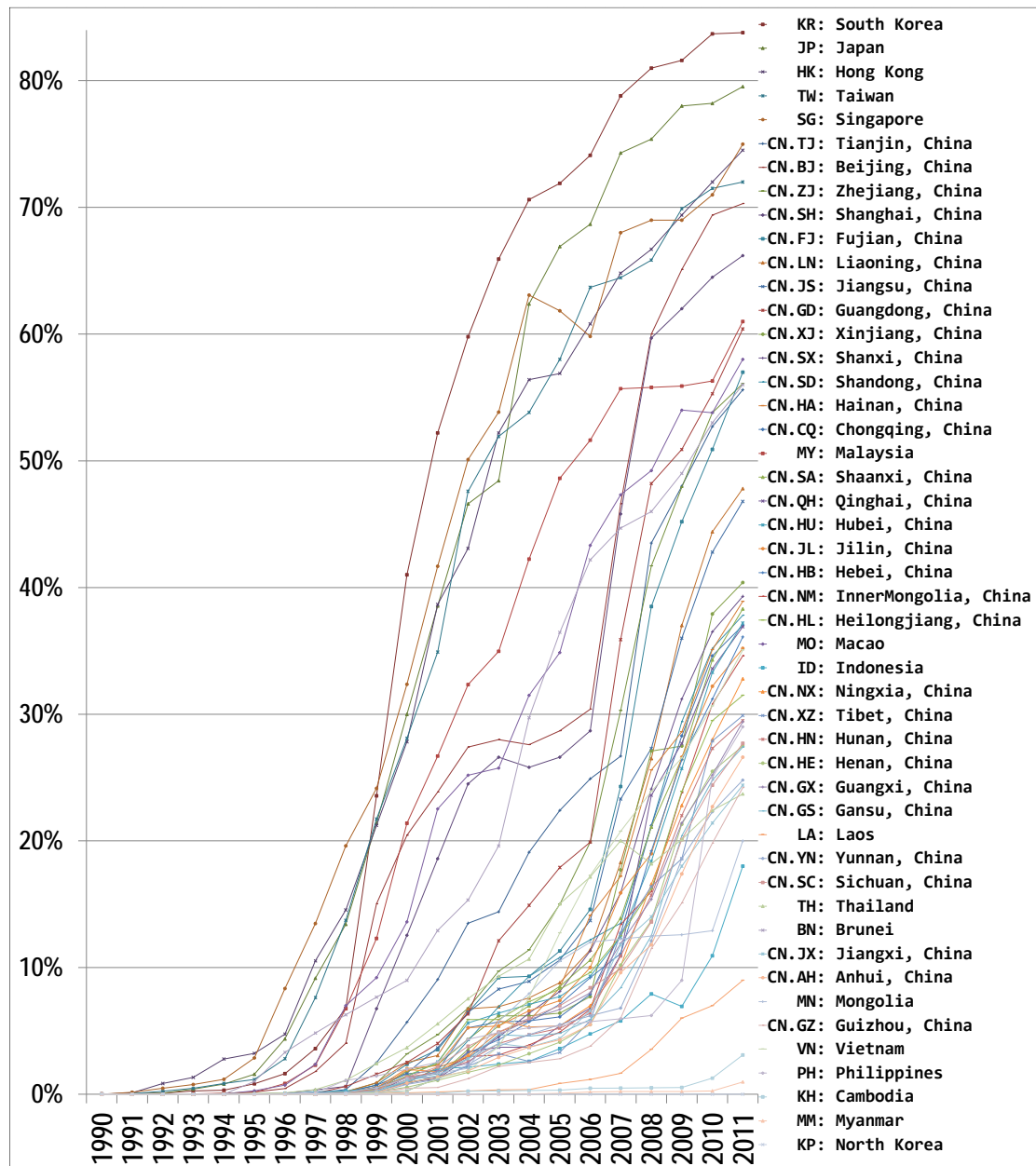


Figure 4-9. Internet diffusion rates for 17 East Asian and 31 Chinese regions

The decision to disaggregate Chinese data has the following benefits: First, it highlights the cosmopolitan-local spectrum within mainland China, differentiating between more and less Internet-developed regions. Second, it overcomes the pitfalls of “methodological nationalism” (Wimmer & Schiller, 2002) by putting the statistical units of mainland Chinese cities (e.g., Shanghai) and provinces (e.g., Guangdong) on par with those of Hong Kong and Taiwan, as well as with other East Asian cities (e.g., Singapore) and countries (e.g., South

Korea). Researchers can thus better assess the impact of the Internet filtering imposed by Beijing on Internet diffusion. This approach is effectively a natural experiment, whereby the East Asian units outside mainland China serve as the control group whereas the mainland Chinese units serve as the treatment group, receiving the treatment of the Beijing-imposed filtering and censorship regime.

I further categorized 48 units, based on 2010 data, into different groups of “region” adopters of the Internet according to their different levels of Internet development as shown in Table 4-4. Several threshold values were derived based on the assumption that the saturated diffusion rate is 80%, which matches the level of developed regions in East Asia in 2011 (79.53% for Japan and 83.80% for South Korea). Thus, other threshold values could be derived by multiplying the saturated diffusion by the accumulated percentages listed in Figure 4-8, resulting in 12.8%, 40%, and 67.2%, respectively. Category I contains “Internet-developed” regions with Internet diffusion rates larger than 67.2%; these are effectively the innovator region units. Category II contains “slow-growing” regions with Internet diffusion rates falling between 40% and 67.2%; these are effectively the early-adopter region units. Category III contains “fast-growing” regions where the rate falls between 12.8% and 40%; these are effectively the late-adopter region units. Categories IV and V contain what are in effect the laggard region units.

Table 4-4

Categorization of different levels of diffusion based on the 2010 data

Category	East Asian units	Mainland Chinese regions	Penetration (% population)
I	Hong Kong, Japan, Singapore, South Korea, Taiwan	Beijing	$x > 67.2\%$
II	Brunei, Macao, Malaysia	Fujian, Guangdong, Jiangsu, Liaoning, Shanghai, Tianjin, Zhejiang	$40.0\% < x < 67.2\%$
III	Mongolia, Philippines, Thailand, Vietnam	Anhui, Chongqing, Gansu, Guangxi, Guizhou, Hainan, Hebei, Heilongjiang, Henan, Hubei, Hunan, Inner Mongolia, Jiangxi, Jilin, Ningxia, Qinghai, Shaanxi, Shandong, Shanxi, Sichuan, Tibet, Xinjiang, Yunnan	$12.8\% < x < 40.0\%$
IV	Indonesia, Laos		$2.0\% < x < 12.8\%$
V	Cambodia, Myanmar, North Korea		$x < 2.0\%$

4.2.4 Results. Since the analytical goal is to compare the data in the treatment group (mainland China regions) and the control group (East Asian regions), the following sections will first compare the statistical units category by category based on the categorization scheme above.

Category by category comparison. Two series of figures can be used to compare the differences between the mainland Chinese treatment group and the East Asian control group: Category by category, the average values of the two groups are compared in the first series of three figures (from Figure 4-10 to Figure 4-12), each of which has a histogram showing the difference each year.

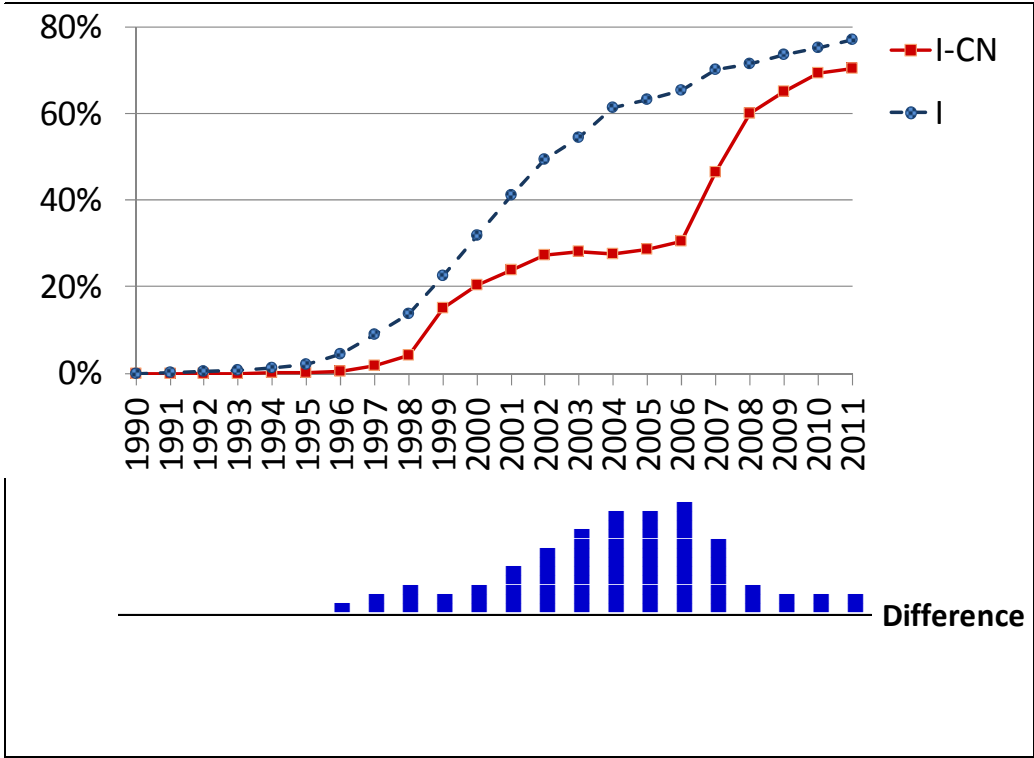


Figure 4-10. Chinese regions and East Asian regions: Category I average

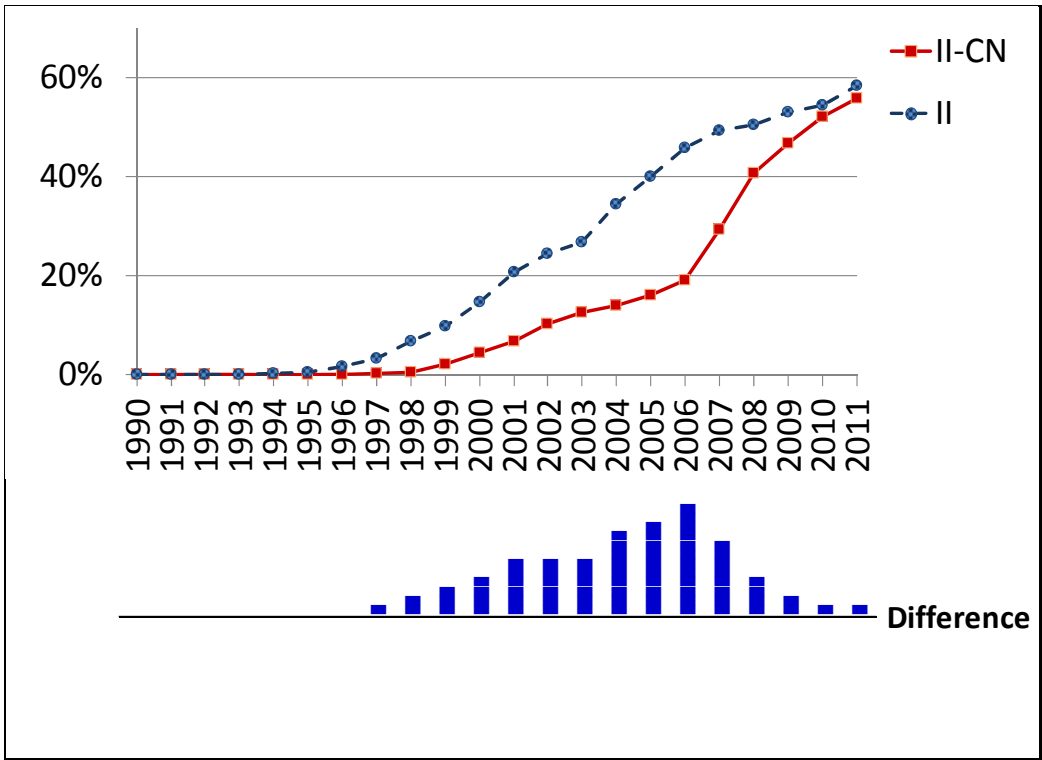


Figure 4-11. Chinese regions and East Asian regions: Category II average

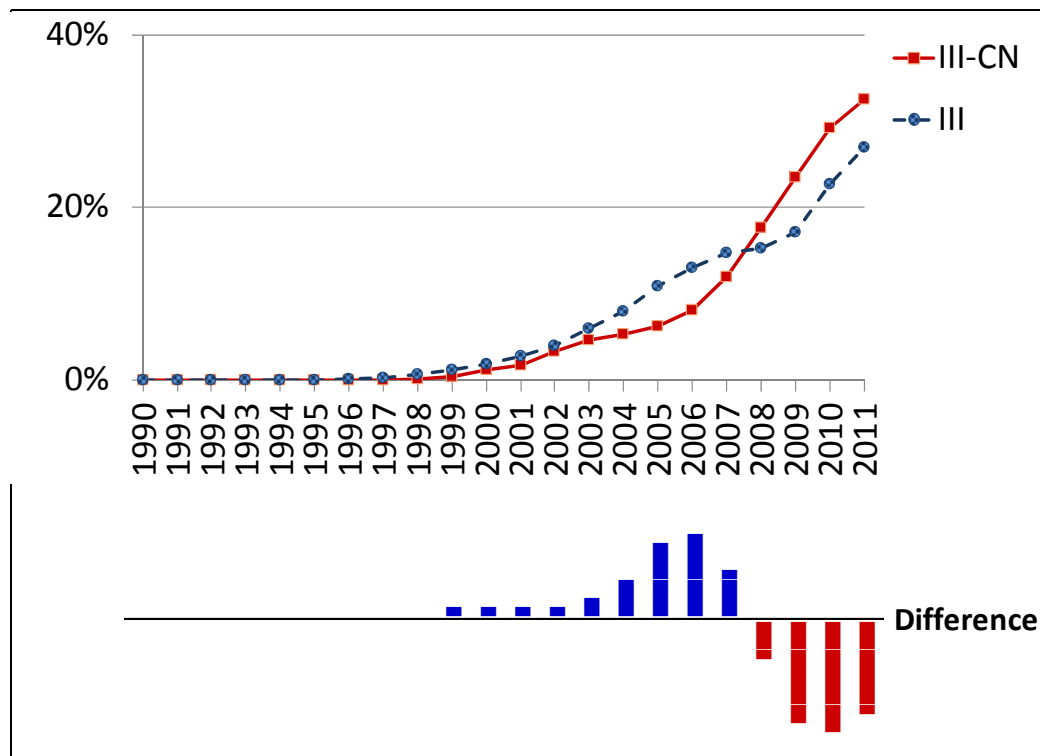


Figure 4-12. Chinese regions and East Asian ones: category III average

What can be observed from the histograms is that the difference in values peaked in 2006. Before 2006, Chinese regions had relatively slower growth rates; after 2006 they caught up or even surpassed other East Asian regions. The year 2006 marks a turning point, which coincides with the year that Baidu Baike was launched.

The second series of figures compares two important selected regions (Beijing and Shanghai) from the treatment group against a similar region selected from the control group. Figure 4-13 shows the contrast between Beijing and Taiwan. Taiwan's curve roughly matches the S-shape model: its growth rate is relatively slower at the early (from 0% to 12.8%) and late (from 67.2% to 80%) stages. In contrast, Beijing departed from the S-shape curve around 2002, signalling its relative slowdown in growth. It reversed this trend in 2006 and finally caught up around 2008. A similar contrast can also be identified in Figure 4-14 between Shanghai and Malaysia in 2006 and 2008.

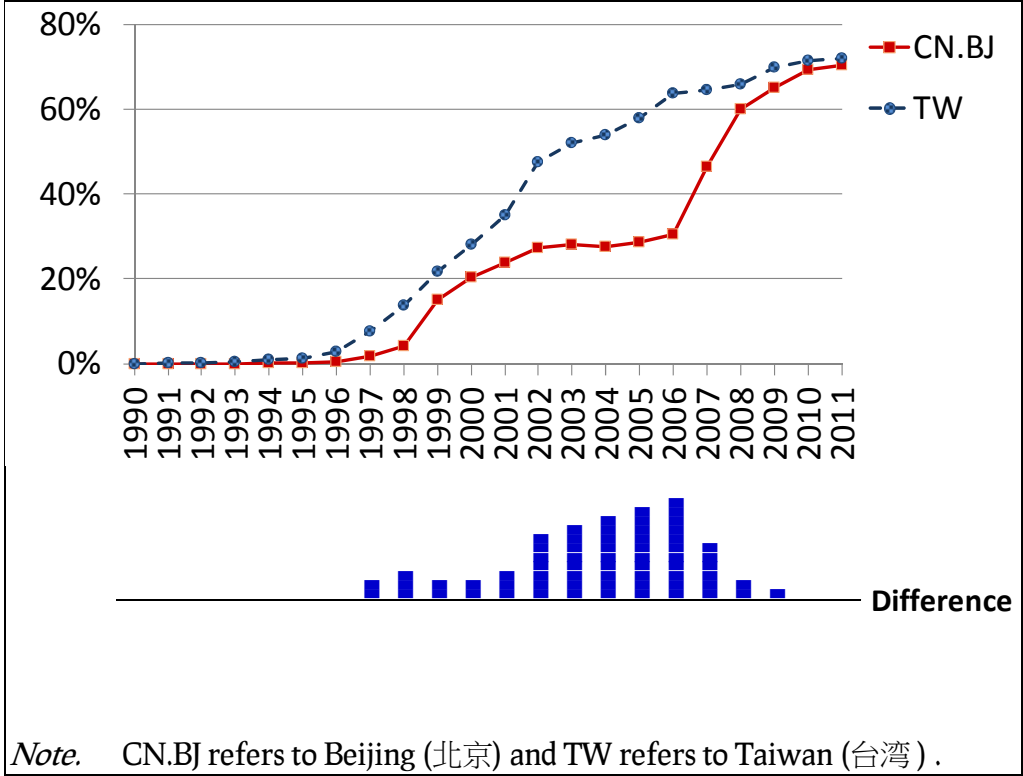


Figure 4-13. Comparison of Chinese regions with East Asian regions: Category I

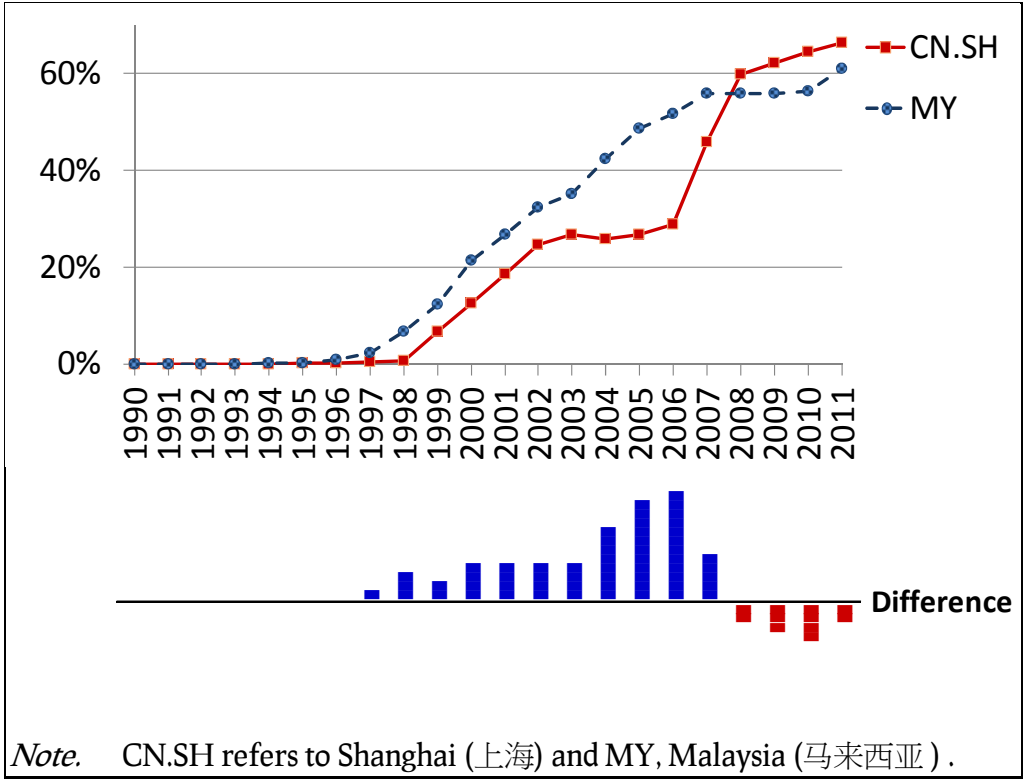


Figure 4-14. Comparison of Chinese regions with East Asian regions: Category II

These findings indicate that, contrary to the expectations of innovation diffusion theory, mainland China indeed underwent a discernible anomaly from 2002 to 2008, with a turning point at 2006. Chinese regions first departed from the ideal S-shape curve of their East Asian counterparts around 2002 with relatively slower growth rates, and then began to catch up again in 2006 with comparatively faster growth rates. The growth after 2006 was rapid enough for these mainland Chinese regions to catch up or even surpass the contrasting East Asian regions around 2008–2009. This contrast strongly suggests that Beijing's censorship and filtering regime has had a discernible impact on Internet diffusion rates, especially from 2002 to 2006, but further analysis is required to explain why from 2006 to 2008 these mainland Chinese regions enjoyed such fast growth and recovery - along with all that this implies for the competition between Chinese Wikipedia versus Baidu Baike.

Socio-temporal and spatial-temporal dynamics. In view of the findings above (especially in Figure 4-10 to Figure 4-12), we can turn to the impacts of the Internet censorship/filtering regime on the two encyclopaedias. First, when Chinese Wikipedia was blocked in 2005, the Internet had not yet reached the following groups of potential Internet users:

- in Beijing: the early majority (most), late majority and laggards.
- in other regions: the early adopters, early majority, late majority and laggards.

When Chinese Wikipedia was largely unblocked in 2008, only the following groups remained unreached:

- in Beijing: the late majority (some) and laggards.
- in seven Chinese coastal regions: the late majority and laggards.
- in other regions: the early majority (some), late majority and laggards.

Therefore, because of the Internet censorship/filtering regime, Chinese Wikipedia not only missed the fastest growth of mainland Internet users from 2006 to 2008, but it was also kept from reaching the critical groups of the early

Internet adopters in all Chinese regions except for Beijing. Launched in the year 2006, Baidu Baike not only rode historically the fastest wave of Internet user growth, but also secured the critical group of early majority users in Beijing and seven other coastal regions.

What does this mean for the historical development of two user-generated encyclopaedias and their cultural thickening potentials? The following paragraphs will first examine the socio-temporal dynamics involved to see how Internet filtering may redirect diffusion, and then it will discuss the related spatial-temporal dynamics to explore the implications of the blocks and diffusion in shifting the boundaries among the Chinese-speaking regions.

A plausible hypothesis: blocks and diffusion. The above findings clearly indicate a phenomenon of suppressing the growth of the mainland Chinese Internet from 2002 onwards, followed by a turning point in 2006 that reversed the once-suppressed growth. The different treatment received by the group of mainland Chinese statistical units as compared to the rest of the groups is likely due to the filtering and censorship regime imposed by Beijing, although other factors cannot be ruled out - for example, the possibility that the data from mainland China is statistically treated differently from the rest of the data. Nonetheless, the mainland Chinese data appears, especially during the critical period where the anomaly is found, to receive consistent treatment according to the CNNIC reports. Thus, while future research may attempt to exhaust all other plausible factors to account for this observed anomaly, this research suggests that the specific historical development of Baidu Baike and Chinese Wikipedia can be accounted for by the establishment of the Chinese Internet filtering and censorship regime.

A plausible hypothesis that I would therefore like to put forward can be called a blocked-then-diffused hypothesis that describes the user-gatekeeping effects of Beijing's filtering and censorship regime. The 2002–2006 suppression of Internet diffusion rates indicates the workings of the regime, i.e., the Internet

filtering regime of largely non-mainland Chinese websites such as Chinese Wikipedia. The subsequent fast growth in 2006–2008 represents the rise of domestic offerings that are censored but not blocked, thus effectively promoting the diffusion of innovation bounded within mainland China, e.g. the launch of Baidu Baike in 2006. One can argue that by the year 2006, the filtering and censorship regime had matured enough to redirect effectively the mainland Chinese domestic demand from foreign to domestic websites. This hypothesis provides a plausible and consistent explanation for other findings presented so far. Moreover, while this hypothesis cannot be fully proven in this thesis, it does provide a plausible explanation that can account for both Internet censorship/filtering regime and Internet diffusion: The regime redirected the course and content of innovation diffusion, but at the expense of the early phase of Internet population growth. The redirection of potential users from Chinese Wikipedia to Baidu Baike then is the likely reason behind the timing of the Internet filtering of Chinese Wikipedia from 2005–2008 and the launch of Baidu Baike in 2006.

Central to the details of this hypothesis is the timing of the Internet filtering and diffusion, which are derived from the results that have been presented. Baidu Baike's timely launch in 2006 enjoys an unprecedented windfall from the rapid growth of Internet users in mainland China with its "early majority" as a potential pool of readers and editors. Note that the windfall during these years is large enough for Chinese regions to catch up or even overtake other equivalent East Asian regions. The windfall of Internet user growth explains why Baidu Baike has picked up many more users in mainland China since its release in 2006 than has Chinese Wikipedia, which was once the leader in Chinese-language user-generated encyclopaedias. Because of the blocks, Chinese Wikipedia missed the window of opportunity of the biggest wave of new Internet users from mainland China, which was exploited by Baidu Baike instead.

This blocked-then-diffused hypothesis is in some ways similar to protectionism in the form of import controls, which many countries have used in the past to protect young domestic industries from foreign competitions (e.g. Lundvall & Borrás, 2005; Steinmueller, 2010).

4.2.5 Discussion: boundaries, blocks and diffusion. Contrary to previous research which argues that Wikipedia and similar wiki sites have failed to gain expected popularity in China because of cultural or national differences (Luo Z.-C. & Fu, 2008; Suo, 2007), the findings presented here suggest - based on the timings of major time periods when Chinese Wikipedia was blocked in mainland China and the windfall of user growth rates - that the impact of Beijing's filtering and censorship regime has been more consequential. The critical moment in competing for potential users out of the pool of the mainland's "early majority" Internet users is confined to the narrow window of 2006–2008, an opportunity that Baidu Baike was able to exploit exclusively while Chinese Wikipedia was blocked.

The plausible hypothesis that has been proposed, a block-then-diffusion hypothesis, gains additional support from the experience of Baidu's founder, Robin Li. It should be noted that Robin Li was involved in the Browser War when he was working in Silicon Valley (on the side of the Netscape browser) and that he devoted a chapter to this topic in his book "Silicon Valley Business War" (guīgǔ shāngzhàn 硅谷商战) (Li, 1999). Since the book was published before he founded Baidu, it can be surmised that he was aware of the strategic importance of timing and distribution channels in competing for new Internet users when Baidu was founded. Here it is relevant to bring to bear the theory of the "second mover advantage": arguably the most discussed case of "second mover advantage" is the browser war between Netscape Navigator and Microsoft Internet Explorer in the late 1990s. This case has been used to answer how and why a second mover (e.g., Microsoft's Internet Explorer) could replace the existing market leader who had made the first move (e.g., Netscape's

Navigator). Despite the early advantage of technological leadership and a large user base, Netscape's Navigator lost to Microsoft Internet Explorer because, according to a number of management and marketing researchers (Bresnahan & Yin, 2006a, 2006b; Windrum, 2004; Yin, 2006), a second mover can still win the overall competition by winning the newer and larger user base before it is too late. For a second mover to prevail, it must exploit its market position in distribution channels (e.g., a near-monopoly of the operating system market by Microsoft) before new users decide which browser to use. In other words, at that time, the main channel of diffusion of Internet browser users was still dominated by the Microsoft operating system, and Microsoft was able to exploit this channel before it was too late to capture new users—in other words, “before the Netscape browser diffused throughout the entire population of potential adopters” (Bresnahan & Yin, 2006b, p. 40). Second movers need to gain the majority of users preferably before their rivals do so, again, usually by using the advantage of having stronger distribution channels.

Baidu Baike can thus be regarded as the timely second mover winner against Chinese Wikipedia in mainland Chinese regions for the following reasons. First, Baidu Baike's launch in 2006 was timely, since most Chinese regions had not yet crossed the threshold between the early adopters and the early majority, with the exception of Beijing. No matter how popular Chinese Wikipedia had been among the innovators and early adopters before 2005, its early advantage was insufficient to guarantee wider adoption in mainland Chinese regions. Second, Baidu Baike exploited the windfall between 2006 and 2008 that its rival Chinese Wikipedia missed because of the Internet censorship/filtering regime, thereby critically establishing the early majority group in mainland China—the essential group for wide adoption. Third, Baidu Baike thus enjoyed at least two advantages in distribution channels: it had Baidu Search, the most popular search engine service in China, and second, it could take advantage of Beijing's filtering regime blocking its major competitor,

Chinese Wikipedia. Baidu's timely launch and advantages in distribution channels thus contributed to Baidu Baike's rapid growth. Thus, while I cannot prove or disprove whether the outcome was the result of conscious strategic decisions made by the Chinese government or whether Baidu was involved, the findings presented are unlikely to be merely coincidental, because Baidu's founder must have been familiar with the concept of second mover advantage as previously and successfully practised by Microsoft against Netscape.

It is worth repeating the key point of the innovation diffusion theory: for any innovation to be diffused among the wider membership of a society or a market, the main challenge to overcome is the gap between the "early adopters" and the "early majority", which is estimated at 12.8% in this research. The findings clearly show that in 2005 no Chinese regions had crossed this gap except for Beijing. Nevertheless, within just three years, all but four regions had passed the 12.8% threshold. This historical anomaly, in retrospect, is critical for new Internet innovations to gain and maintain a solid user base in mainland Chinese regions. For Chinese-language online encyclopaedias, Baidu Baike enjoyed this very windfall in mainland China that Chinese Wikipedia had missed. Even if my proposed hypothesis that the windfall was created largely by the Chinese Internet censorship regime is falsified in the future, the historical existence of the windfall and its impact on the two encyclopaedias are nevertheless an established fact.

These findings, including the trend lines in Figure 4-9 and the 2010 snapshot categorization in Table 4-4, are consistent with the larger spatial diffusion flows of Internet development and other regional developments. Japan, Korea, Taiwan, Hong Kong and Singapore are major and early nodes of networks of global submarine cables (which carry the bulk of the global Internet traffic), popular media content circulation (which constitute the bulk of the East Asian cultural market), and financial capital flows. Equally, the mainland Chinese regions exhibit similar related spatial-temporal dynamics: Internet use diffuses

spatially from Beijing and other coastal regions towards other regions inland. The data also shows that these more advanced Chinese coastal regions caught up with nearby regions from 2006 onwards, which is not surprising given the Internet's geographic topology in this region: Beijing, Shanghai and Guangzhou host the three major domestic Internet exchanges and the major global Internet connections to other more advanced regions in East Asia. Thus, the Internet regime of China, which is responsible for the Internet filtering of Chinese Wikipedia in mainland China, serves as a major valve in the region for the spatial-temporal dynamics of Internet diffusion. It effectively acts as a gatekeeping method against users from mainland China.

The analysis of the timing of major editorial developments of the two encyclopaedia websites in the context of the growing number of new Internet users provides important insights regarding boundaries, Internet filtering and diffusion. First, the “access blockage”, as meant by Taneja & Wu (2013), were part of a larger adaptation of Beijing's filtering and censorship regime that matured within a decade. Second, the diffusion of the Internet and other Internet-related innovations was negatively influenced by the Internet censorship/filtering regime in the early half of the decade among mainland Chinese regions. The websites permitted in mainland China were not ready at that time to serve the early adopters there. Thus, the blocked-then-diffused hypothesis suggests a time-sensitive process in which Beijing's filtering and censorship regime exercised its gatekeeping power to channel the then-future and now majority of Internet users in mainland China to the allowed websites like Baidu Baike. In this sense, the regime has contributed to the formation of a national market in mainland China and prevented a Chinese-language market across regions. Although the findings presented here cannot prove any possible coordinated efforts between the government and Baidu in channelling new Internet users, they nonetheless document a historical shift around 2006, a key period for understanding the competition between Baidu Baike and Chinese

Wikipedia and other websites. The proposed blocked-then-diffused hypothesis is thus consistent with the authoritarian effects of channelling access (Barzilai-Nahon, 2008). New Internet users in mainland China from 2006 to 2008 can thus be seen as being channelled from the first mover Chinese Wikipedia to the second mover Baidu Baike.

Additional spatial-temporal insights, gained by disaggregating Internet diffusion data, suggest the significant roles played by the coastal regions in mainland China in relation to the more advanced East Asian regions, including Japan, Korea, Hong Kong and Taiwan. In fact, these mainland Chinese coastal regions are relatively more urban and connected with these East Asian regions via the East Asian popular culture market, industry production chains and physical Internet connections. Thus, the spatial and temporal diffusion of the Internet in the East Asian and Chinese regions suggests the likely direction of diffusion of new ideas, practices and technologies. They move first from the more advanced East Asian regions, including Hong Kong and Taiwan, to the more advanced cities such as Beijing, Shanghai, Shenzhen and Guangzhou in the Chinese coastal regions, and then to the rest of the Chinese regions. Thus, Beijing's Internet regime has likely redefined a boundary among the Chinese and East Asian regions, thereby shielding most of the Chinese regions from the influence that originated from and passed through regions such as Hong Kong and Taiwan, including Chinese Wikipedia.

Adding these perspectives of diffusions to our understanding of the Chinese-language Internet also refutes some of Taneja & Wu's (2013) arguments on the formation of a Chinese culturally-defined market. First, their argument that the censorship/filtering regime has contributed to the formation of this market is thus at best confusing and at worst misleading. Based on the findings here, a more precise description of the Chinese culturally-defined market would be Beijing's "culturally-and-politically sanitized domestic market" or "censorship/filtering-delineated market" for mainland Chinese users, which is

distinct from the Chinese-language culturally-defined market. Second, based on the overall traffic-based analysis in 2012, their findings are limited to the snapshot situation that year. The proposed alternative blocked-then-diffused hypothesis here highlights the plausible effects of the censorship/filtering regime on users that should not be overlooked. What Taneja & Wu (2013) have described in Figure 4-7 as “Users” in 2001 (“imposition of blockage”) and “Users” in 2012 (“after years of persistent blockage”) are not exactly the same users. For example, the Internet diffusion rate in Shanghai was below 30% from 2001 to 2006 but reached over 70% by 2012.

What Taneja & Wu (2013) overlooked about the “users” then is the dynamic of the growing Internet population before and after Beijing’s implementation of its Internet censorship/filtering regime. The findings here suggest that regime have had negative or suppressive effects on the Internet diffusion rates among mainland Chinese regions. They even acknowledged that Chinese Wikipedia has been marginalized by it: “currently at the periphery due to blockage” (Taneja & Wu, 2013, p. 23). Thus, Chinese Wikipedia serves as a major piece of evidence for its impact on the Internet diffusion process that reshaped the thickening patterns on the Web. While Taneja & Wu made an exception for Chinese Wikipedia by speculating that the lifting of the blockage would have made Chinese Wikipedia the centre of the Chinese cluster, it is likely that other blocked websites would have gained central positions, thereby undermining Taneja & Wu’s argument that the impact of Beijing’s filtering and censorship regime on Chinese-language connections was limited or even negligible.

In conclusion, what Chinese Wikipedia missed and Taneja & Wu failed to consider is the 2006–2008 windfall of Internet population growth that defines the majority group of mainland users. Effectively, Beijing’s censorship and filtering regime channelled potential users from the pool of new Internet users away

from Chinese Wikipedia and to Baidu Baike, thereby producing cultural patterns that are distinct from the ones already produced by Chinese Wikipedia.

4.3 Chapter conclusions

Over the past decade, Chinese-language users have rapidly adopted Internet technologies. How and where these users can become part of the ongoing mutually reinforcing cycle of increased participation, content, and readership are important questions for many websites, especially for Baidu Baike and Chinese Wikipedia. This chapter has addressed the participation component of the cycle, and indeed both encyclopaedias could be seen as different collaborative ecosystems (Okoli et al., 2012) that have grown with the general growth of Chinese-language Internet users.

Table 4-5 summarizes the results of comparing their editorial development and patterns, with the differences clearly indicating a specific context of the growth and control of the Internet users in mainland China. Despite the fact that Chinese Wikipedia was the first mover, its being blocked by Beijing consistently from mid-2005 to mid-2008 gave the second mover Baidu Baike an advantage. The 2006–2008 windfall of the new Internet users in mainland China likely benefited Baidu Baike's launch in 2006. In addition, the editorial policies of the two websites differ greatly in whether they censor politically sensitive content as seen by Beijing, a contrast that is typical between websites hosted inside versus outside mainland China. Despite being rejected by the community, the proposals for Chinese Wikipedia to practice self-censorship in 2005 and 2007, imply a link between censorship and blocking. There are also similarities here to the policy signal sent by Beijing to corporations such as Google, Yahoo! and other Internet companies: self-censor yourself or face blocking. Another difference in information gatekeeping is that Chinese Wikipedia filters marketing, self-promotion and copyright infringement edits - whereas Baidu Baike does not. Geolinguistically, Baidu Baike implicitly allows only mainland Chinese edits, whereas Chinese Wikipedia has made efforts to

accommodate mixed Chinese-language-script contributions by users from different regions and established the "avoid-region-centrism" policy. The policy demands mutual respect for different scripts and terminologies preferred by fellow editors, resulting in geolinguistic profiles of power editors and active contributors that are more diverse. (Chapter 5 will confirm this based on another sets of findings.) I further argue that the patterns of discourse and practice also differ: Baidu Baike has an overall tendency towards the discourse of being an indigenous encyclopaedia (i.e. mainland China's encyclopaedia website) that focuses on growth, whereas Chinese Wikipedia's explicit editorial policies establish the discourse of building a "Chinese-language" encyclopaedia that control content quality and respect differences among users and regions.

Table 4-5

Comparing editorial development and patterns

Patterns	Baidu Baike	Chinese Wikipedia
... of growth	Second-mover advantage	First-mover advantage
... of being blocked by Beijing	None	Consistently from mid-2005 to mid-2008
... of reach to mainland Chinese users (esp. the 2006-2008 windfall of the new Internet users)	More likely through Baidu because of the windfall	Less likely because of the blocks by Beijing, missing the windfall
... of gatekeeping	Filtering/censoring politically incorrect content	Filtering self-promotion, marketing, and copyright infringement edits
... of geolinguistic arrangement	Mainland simplified Chinese only	Mixed and the "Avoid-region-centrism" policy
... of geolinguistic profile of contributors	Mostly mainland Chinese	Distributed across various (Chinese-speaking) regions
... of discourse	Indigenous, not "foreign" (Webster, 2008)	Chinese-language version, avoiding nationalistic biases
... of practice	Growing content and users	Controlling content quality and respecting differences

These two paths in editorial development lead to different kinds of cultural thickening among Chinese-language users. Managed by the oversight of Baidu employees which report to Beijing, Baidu Baike contributes mainly to the “cultural thickening” within mainland China. In contrast, self-governed by active editors and administrators (elected mutually among the editors themselves), Chinese Wikipedia contributes to the “cultural thickening” across major Chinese-speaking regions and cultural-political divisions. Despite the dominant discourse on the China–US opposition, the main tension seems to be between certain Chinese-speaking regions, and especially between the differences in their cultural-political environments ranging from language scripts to freedom of speech. Baidu Baike’s editorial system mirrors the cultural-political environments of mainland China; Chinese Wikipedia features a system that integrates diverse cultural-political experiences from major Chinese-speaking regions. Thus, diverse media practices and environments, which include Hong Kong and Taiwan, are also likely to have resulted in the editors from these regions shaping the differences between Chinese Wikipedia and Baidu Baike. Different cultural thickening patterns indicate different impacts on boundaries: Baidu Baike excludes traditional Chinese, thereby effectively discouraging contributors from Hong Kong and Taiwan from participating. Using geolinguistic identifiers to acknowledge and process Chinese script and phrase variations, Chinese Wikipedia has integrated these users. Thus, Baidu Baike reinforces the boundary between mainland China on one side and the rest of the Chinese-language regions on the other, whereas Chinese Wikipedia has overcome, at least nominally and technologically, the boundaries among the four recognized geolinguistic regions as of early 2013. It is Chinese Wikipedia’s approach that overcomes these boundaries since it recognizes the differences across these geolinguistic regions.

The different extent and centres of the two cultural thickening patterns further show the cultural-political tensions across the regions. Arguably, Beijing's censorship/filtering regime has successfully avoided outside influence from Hong Kong and Taiwan at the expense of potential substantial cultural thickening between the regions inside and outside mainland China. Were it not for Beijing's intervention and gatekeeping of mainland Chinese users, the centre of gravity would likely be spatially located among Hong Kong, Taiwan and other coastal urban areas in mainland China, as shown by the geographic distribution of power users in Figure 4-3 and Figure 4-4. The Internet filtering of Chinese Wikipedia and the launch of Baidu Baike suggest a refashioning of Internet diffusion: Beijing's filtering regime channelled the majority group of users to its preferred allowed websites, thereby shaping the diffusion of new innovative websites. An alternative cultural thickening pattern thus guided the majority of new Internet users from mainland China away from the pre-existing patterns that went across major Chinese-speaking regions. Baidu Baike's dominance in mainland China but limited outreach to traditional Chinese users indicates the re-centring efforts that have aimed to make Beijing *the* cultural-political centre. These cultural-political struggles are inherently *Chinese* rather than a struggle of Beijing versus Washington or Beijing versus San Francisco. The main struggles are about whether Beijing should be the sole centre and source of cultural thickening processes of the Chinese-language Internet and whether Hong Kong and Taiwan could be the other main sources or even alternative centres.

It is plausible, however, that motivations for China Internet policies might well be driven mainly by industrial policies, such as creating domestic businesses and commerce, rather than cultural policy. Regardless of the motivations behind Internet policies, their impact on cultural patterns as observed here is significant.

The findings further suggest that different cultural patterns can be produced or guided by "codifying" users with geolinguistic codes and then

structuring their contributions and activities accordingly, as part of the effort to shape “geolinguistic processability”, a concept that has been discussed in Chapter 3. Beijing’s filtering and censorship regime thus reinforces a cultural thickening pattern within the geolinguistic identifier of “zh-cn”. By blocking Chinese Wikipedia from mainland Chinese users, Beijing channelled some of the early and crucial popularity away from Chinese Wikipedia to Baidu Baike, thereby fortifying its cultural-political power in ways that are possible with Baidu Baike but much less so with Chinese Wikipedia. Such practices lead to the cultural thickening patterns desired by Beijing. Indeed, it has been reported that Chinese authorities in Beijing have “guided” public opinion by encouraging some voices while discouraging others on various media platforms (Brady, 2009; Muncaster, 2012; G. Yang, 2011). What we can see here is the importance of user autonomy which cannot be overstated; hence it is necessary to examine the ways in which user contributions are encouraged or discouraged through website designs and policies.

Regarding Chinese Wikipedia’s integration across regions and Baidu Baike’s sole focus on mainland China, the findings indicate two general cultural thickening patterns. These patterns result from the diffusion and filtering of Chinese-language websites. Neither the perceived notion of China’s “indigenous” rise nor the opposite notion of US-originated “global” Internet values can adequately describe dynamics. The findings tell neither a story of mere US-centric Americanization nor the story of a mere “Chinese local-area network”. It is rather a story of two different cultural thickening developments mainly across and within Chinese-speaking regions. Ultimately, it is a struggle to “centre” the Chinese-language web sphere by promoting certain directions, intensities and normative values via cultural thickening processes among users and such user-generated websites.

To sum up, the contrast in the editorial development of Baidu Baike and Chinese Wikipedia, including the historical growth of mainland Internet users

contributing to this development, demonstrates two distinct cultural thickening patterns, as evidenced by their different geolinguistic preferences and cultural-political centring. The two encyclopaedia websites have engaged Chinese regions differently, producing different cultural thickening patterns. I argue that such distinction is the direct outcome of Beijing's intervention in the exchange dynamics of different media and information environments across Chinese-speaking regions. The authoritarian constraints imposed by Beijing on Chinese Wikipedia could thus be interpreted as a particular form of cultural-political policy that aims to promote Beijing-centric cultural thickening with normative values different from Chinese Wikipedia's. Baidu Baike's Beijing centric practices are in direct contrast to the region-neutral, multi-script, open-ended practices of Chinese Wikipedia that put the various Chinese-speaking regions on an equal footing.

Chapter 5 Citation and content analysis

The previous chapter has compared the editorial processes of Baidu Baike and Chinese Wikipedia, and this chapter will continue by examining the outcomes of these processes in filtering and curating the world's information. In this chapter, the aim is to reveal the two encyclopaedias' preferences for information sources and will present findings based on a systematic analysis of the two million external web links from all article entries of Baidu Baike and Chinese Wikipedia (collected in June 2010) and also analyse specific articles that typically define the notion of Chineseness.

The goal of this comparative analysis is to show how the citations and content reflect the scope, focus, and nature of information gathered on the two websites. I contend that the information outcomes exemplify certain forms of online civic engagement – or the accumulated results of open collaboration where anyone can read and edit. Although both platforms promise more open and equitable knowledge production, the outcome is bound to be shaped by contributors and the ways in which they use knowledge and information sources. In this sense, the two major online Chinese encyclopaedias, Baidu Baike and Chinese Wikipedia, provide important opportunities for researchers to assess online civic engagement in the domain of Chinese-language knowledge and information. Effectively, these contributors/editors of the two encyclopaedias build “civic learning repertoires” for “information engagement” (Bennett & Wells, 2009), including the citations they use (as well as those they remove).

On the topic of civic engagement on the Chinese Internet, there has been some previous research on Chinese language societies (D.-Y. Chen & Lee, 2008; Jiang, 2007; Leibold, 2011; Leung, 2009; Tai, 2006; G. Yang, 2003; Yongnian Zheng, 2007), but none has systematically examined how geographic and linguistic factors shape online engagement. For example, Leung (2009) has explored the link between civic engagement and content generation based on a sample of Hong Kong users. Most research, however, has focused on civic engagement in

mainland China (e.g. Jiang, 2007; Leibold, 2011; Shao, Lu, & Wu, 2012; Tai, 2006; J.-C. Zhang & Qin, 2012; Zheng, 2007; Chen and Lee (2008) focused on Taiwan). Yang Guobin (2003) has argued that the Internet contributed to a transnational Chinese cultural sphere that fulfilled political functions within and beyond mainland China, though without focusing on geographic and linguistic factors. Yet geographic and linguistic differences are important, as we have seen, for the discussion of Chinese media and globalization (Chan, 2009; Guo & Guo, 2010), including practical implications for interface design for websites (Chung et al., 2004; Liao, 2009b). Thus, further research is needed to examine how geographic and linguistic factors influence civic engagement across Chinese societies.

As analysed in the previous chapter, the editorial processes of Baidu Baike and Chinese Wikipedia are influenced by contributors' experiences offline. Civic engagement in terms of citations and content is likely to be similarly influenced by various political and media experiences across Chinese-language societies. Again, we shall see that the ideal notion of universal knowledge confronts local conditions. Two opposite impulses exist between knowing the world and serving a group of users: one is exemplified by Baidu Baike's the slogan of "Let Humanity Know the World Equally" (ràng rénlei píngděng de rènshí shìjiè 让人类平等的认识世界), expressing the ideal notion of encyclopaedias to document universal knowledge about the world. Another is to provide authoritative or trustworthy information, concentrated in certain sources as agreed upon by the user-contributors involved: this comes closer to Chinese Wikipedia's ideal. On this view, the citation and content produced should reflect the overall editorial judgement regarding which knowledge and information sources are deemed reliable. These kinds of "information engagement" are bound to involve exchanging experiences and expectations about knowledge and its sources across Chinese-speaking regions. Chinese Wikipedia, as I have discussed elsewhere, has the potentials of "unbounded citizenship" or "reconfiguring citizenship" mainly because of its efforts in

conjoining and integrating content across four main Chinese-speaking regions (Liao, 2009a). While this chapter will not go so far as to speculate on the possibilities of an unbounded or reconfigured Chinese citizenship, here the focus is on information engagement that may overcome or reinforce existing boundaries which points to the cultural-political tensions in negotiating civic cultures.

Through research concerning online information engagement and offline civic cultures (as represented by geolinguistic differences), this chapter aims to answer a more general question: How do Internet connectivity effects (Haythornthwaite, 2005) realize the potential for enhancing civic engagement and civic culture (Benkler, 2006; Rheingold, 2008; Scammell, 2000; Shirky, 2010), as Wikipedia has been cited as one of the major examples (Leung, 2009; Lih, 2009; Reagle, 2008; Sherrod, Torney-Purta, & Flanagan, 2010)? Indeed, as the largest user-generated encyclopaedia platform that potentially serves users from around the world, Wikipedia may encourage civic discussions related to a more open and equitable production of knowledge for all. From one historical perspective, Wikipedia shares the roots of the 18th-century European Enlightenment, including “the rational impetus to understand and document all areas of the world” (Ayers, Matthews, & Yates, 2008, p. 35), an ideal which applies to encyclopaedias as they go online and constitute new kind of civic engagement or civic cultural practice. The intensity of - and type of content resulting from - civic engagement in both Chinese-language encyclopaedias remains to be researched.

Hence, this chapter will assess the information outcome of online civic engagement within the two encyclopaedias, with the aim of answering the boundary question for Chinese-speaking regions. To do so, several types of data are collected about the citations and content of the two encyclopaedia websites to examine how they cover the world and define Chineseness, which will be an important indicator of the quality and patterns of civic engagement for users across Chinese-speaking regions. The first section will address this question by

means of a geolinguistic analysis (which also corresponds to the geolinguistic processability discussed in Chapter 3) of all external web links of the two encyclopaedias. The second section will advance this analysis by zeroing in on the differentiating regions for comparison. The third section will focus on a few selected key articles that define “Chineseness” for more detailed comparison. Altogether, the citation and content data should answer the question regarding the outcome of civic engagement as a case of gatekeeping and reliable knowledge sources. The chapter conclusion will discuss how the findings relate to the theoretical questions of the thesis about cultural thickening and boundary dynamics.

5.1 Geolinguistic patterns and preferences

The ideal of using human knowledge to engage citizens has often been subject to parochial and national concerns despite the universal ideal of enlightenment encyclopaedias. During the European enlightenment, geographic and linguistic barriers were among the major challenges for knowledge collection and diffusion (Delanty, 2009; Israel, 2001; Roche, 2006). Facing similar challenges, the Wikimedia Foundation, the hosting organization for all Wikipedia projects, has targeted the “Global South” regions of Brazil, India, and the Arabic language countries for engagement (Wikimedia Meta, 2012). Nonetheless, some research has suggested that cultural and linguistic factors have prevented wider acceptance of Wikipedia, especially the Chinese and Korean versions (Shim & Yang, 2009; Suo, 2007). Such questions about Wikipedia projects constitute part of larger and on-going discussions about the Internet and cosmopolitanism (e.g. Ess, 1998, 2002; Jeffres, Atkin, Bracken, & Neuendorf, 2004; Zuckerman, 2013). Geographic and linguistic factors remain crucial.

Analysing and comparing the geographic or linguistic features of online data, including user-generated content such as Wikipedia, has been a useful approach in identifying the boundaries or gaps in knowledge production (e.g. Thelwall & Smith, 2002; Liao, 2008; Hecht & Gergle, 2009, 2010; Graham, Hale,

& Stephens, 2011; Liao & Petzold, 2010; Lowe, 2011; Paolo Massa, 2011; Bao et al., 2012; P. Massa & Scrinzi, 2012; Petzold, Liao, Hartley, & Potts, 2012; Warncke-Wang, Uduwage, Dong, & Riedl, 2012). For instance, in user-generated encyclopaedia research, Liao (2008) has been among the first to compare the geographic and linguistic features of one language version of Wikipedia with another non-Wikipedia project. Hecht and Gergle (2009) have similarly compared the geographic focus of several major language versions of Wikipedia. Coming from various discipline backgrounds with different tool sets, these researchers effectively provide a toolbox for advancing understanding of geographic and linguistic dynamics online. Hereafter, I refer to such research loosely as geolinguistic research that focuses on the systematic analysis of geographic and linguistic information online. Although most of this research does not directly contribute to the question of civic engagement, this collection of research has the potential to provide crucial evidence about where and who is actually engaged and with what content.

To fill the gap in the geolinguistic analysis of civic engagement, especially concerning the Chinese-language world, this section presents a comparative geolinguistic analysis of two million web links collected in June 2010 from all articles of Baidu Baike and Chinese Wikipedia, the two major user-generated encyclopaedias for Chinese-language Internet users. By framing Baidu Baike and Chinese Wikipedia as examples of “network gatekeeping” (Barzilai-Nahon, 2008) or “collaborative filtering” (Benkler, 2006), this research examines whether and how they actualize the ideal notion of encyclopaedias in producing universal knowledge about the world by filtering online sources. In short, the overall geographic and linguistic features of these external links should reflect the outcome of such network gatekeeping or collaborative filtering processes, thereby providing important evidence about the influence of geographic and linguistic factors on Chinese civic engagement online.

External web links embody network gatekeeping effects by providing two major functions for online user-generated encyclopaedias: first, they support knowledge claims with online references. Indeed, Wikipedia's editorial policies of "Verifiability" and "No Original Research" demand verifiable sources, suggesting the important relationship between citation sources and statements (Ayers et al., 2008; Myers, 2010). Second, they provide clickable links that lead readers directly to materials hosted on other websites, thereby serving as gateways to information hosted elsewhere. In addition, these clickable links can be used by computer software programs including search engines, and thus contribute to the overall hyperlinked pattern of the World Wide Web. The online citations and clickable web links in user-generated encyclopaedias should therefore reflect preferred information sources or authoritative knowledge.

How online sources are determined to be reliable by contributors can be a challenge for Chinese-language users from different, largely Chinese societies to agree on, especially as regards cultural-political issues. Mainland China, Hong Kong, and Taiwan have political and institutional differences: Mainland China is "neither fully free nor democratic"; Hong Kong is "very free" and "little more democratic than it was under the British"; Singapore is "democratic ... but ... not free"; and Taiwan has been "both free and democratic" since the 1990s (Mitter, 2008, p. 132). During the Cold War, Hong Kong, Macau, and Taiwan were not ruled by the People's Republic of China (Hong Kong and Macau were returned to China in 1997 and 1999 respectively). These regions, particularly Hong Kong and Taiwan, have played instrumental roles during China's economic reform since 1978 in various domains, including financial investment, Internet development (Damm, 2007; P. Lee & Rice, 1998; Qiu, 2005; G. Yang, 2003), and popular culture (Gold, 1993). Nonetheless, the issue of political integration remains contentious.

A geolinguistic analysis of all external links in the two user-generated encyclopaedias can show the diversity and/or concentration of information

sources of the two across different regions and languages, thereby providing insights about online civic engagement within and across Chinese societies.

5.1.1 Method: geolinguistic analysis of web links in the Chinese context

Extending the fields of webometrics (Almind & Ingwersen, 1997; Björneborn & Ingwersen, 2004; Park & Thelwall, 2003; Thelwall, 2009; Thelwall, 2008; Thelwall & Vaughan, 2004) and citation analysis (Borgman, 2007; Garfield, 1979), this research proposes a systematic approach to analyse both the geographic and linguistic information of web links. Although some previous webometric research on web links has exploited some geographic information (e.g. Ortega et al., 2009; Thelwall et al., 2002) and language information (e.g. Hale, 2012; Hong et al., 2011; Petzold, Liao, Hartley, & Potts, 2012), this research advances an integrated approach that considers both as independent variables. In short, for geographic information, it uses both the country code top-level domain name (ccTLD, not to be confused with generic top-level domain names, gTLDs) and the geographical location of an IP address (geo-IP lookup). For linguistic features, it uses the language encoding standards, complemented with tools tailored to measure the variation between simplified and traditional Chinese.

Similar to the way webometrics and/or citation analysis research attempts to study the underlying structure of human interaction from the Web and/or the citation networks (Hong et al., 2011; Hale, 2012; Ortega & Aguillo, 2009; Thelwall & Smith, 2002), the study of external links of online encyclopaedia should reveal the latent structure of information citation as reflected on the overall geographic and linguistic preferences for reliable information sources.

Generic geographic and linguistic features. For a given web link, certain generic geographic and linguistic features can be extracted or detected.

For geographic categorization, this study uses two separate schemes, ccTLD and geo-IP. Each has its own strengths and weaknesses. Based on country code top-level domain names (ccTLD), such as “uk” (representing the UK), “cn” (China), and “tw” (Taiwan), the ccTLD categorization can capture the nominal

expression at the country level. Based on IP addresses, the geo-IP categorization can detect the geographic location of a computer server, since almost all working web links require a computer server hosted somewhere to deliver content.

Some important caveats need to be addressed here: for the ccTLD scheme, given the historical legacy of domain names, three generic top-level domain names (gTLDs), namely “.edu”, “.gov”, and “.mil”, are categorized as being in the U.S. because these domain names are owned and/or administered exclusively by U.S. educational, governmental, and military institutions. For another, domain names, unlike most phone numbers and mailing addresses, do not have to correspond to the actual geographic location where the web page is hosted. Also, there are some ccTLDs that are repurposed for different uses (e.g. “.tv” as an abbreviation for television and “.fm” referring to FM radio). Given these caveats, this research follows the following guidelines: Acknowledging the fact that ccTLD cannot categorize certain websites (including those with popular gTLDs such as “.com” and “.org” and those with only IP addresses), the geo-IP (to be discussed in the next paragraph) scheme must be used to complement the ccTLD findings. Second, the ccTLD outcome should be interpreted as the “expressed” geographic target instead of the actual physical location of a given web link. Third, those links with three gTLDs (“.edu”, “.gov”, and “.mil”) should be categorized as the U.S. By addressing and acknowledging these caveats, the proposed ccTLD and geo-IP categorization schemes provide two best-effort and complementary views on the geographic distribution of web links.

The second generic geographic categorization scheme, called “the geographic location–Internet Protocol address lookup” (geo-IP lookup), has been widely used for commercial and research purposes in detecting the geographic location of an IP address (the standard numerical address underlying all web applications). The geo-IP lookup process requires two procedures: First, the IP address needs to be looked up from the Web address using Domain Name System (DNS) services. This research uses the Public DNS service provided by

Google. Second, based on the produced IP address, the associated geographic information (e.g. countries, regions, or even cities) can be further looked up in some geo-IP databases such as GeoLite Country (MaxMind, 2012) used in this research. Though these two are not the only sources available, they have been utilised by many users and organizations. The Google Public DNS resolution service promises “absolutely no redirection”, meaning that it will never falsely give the alternative IP addresses, also known as DNS hijacking (Google, 2012b). MaxMind promises 99.5% accuracy for looking up geographic locations based on its free GeoLite Country database (MaxMind, 2012).

For linguistic categorization, a basic generic categorization method is used to produce language-script-level categorization (e.g. Latin, Arabic, Cyrillic) by detecting the “character encoding” standards, as shown in Table 5-1.

Table 5-1

Examples of character encoding standards (character sets)

Languages/Scripts	Character Sets	Aliases
English	ascii	646, us-ascii
Traditional Chinese	big5	big5-tw, csbig5
Traditional Chinese	big5hkscs	big5-hkscs, hkscs
Simplified Chinese	gb2312	chinese, csiso58gb231280, euc-cn, euccn, eucgb2312-cn, gb2312-1980, gb2312-80, iso-ir-58
Unified Chinese	gb18030	gb18030-2000
Western European	latin_1	iso-8859-1, iso8859-1, 8859, cp819, latin,
Central and Eastern European	iso8859_2	iso-8859-2, latin2, L2
Arabic	iso8859_6	iso-8859-6, arabic
Greek	iso8859_7	iso-8859-7, greek, greek8
Hebrew	iso8859_8	iso-8859-8, hebrew
Turkish	iso8859_9	iso-8859-9, latin5, L5
all languages	utf_32	U32, utf32

While linguistic categorization has been performed by language detection algorithms to produce language-level categorization (Hale, 2012; Hong et al.,

2011), this is not suitable for the research here: for the purpose of this research, it is not necessary to distinguish German versus Spanish among Roman scripts or Egyptian Arabic versus Iraqi Arabic among Arabic ones. The language-script-level categorization in Table 5-1 provides a sufficient indication of language scripts based simply on the character encoding systems used. The only unresolved exception is the all-inclusive Unicode, where all scripts are possible. Additional mechanisms of categorization, especially designed for the Chinese context, are thus needed.

Geographic and linguistic differences in the Chinese context. For political and cultural reasons, specific geographic and linguistic differences are crucial in the Chinese context. The following paragraphs describe how the proposed methods of geolinguistic analysis are applied in the Chinese context.

As discussed in Chapter 2, the regions with a majority population of Chinese-speaking users have the following country codes: “cn” (mainland China), “hk” (Hong Kong), “mo” (Macau), “sg” (Singapore), and “tw” (Taiwan). The existence of such a variety of ccTLDs reflects technological and political developments of the past. As for the geo-IP results, those hosted inside mainland China are expected to be subject to Beijing’s censorship regime.

For some major mainland China-based websites, one caveat requires special treatment. Because of the use of “content delivery networks” (CDN) services (Totok, 2009) such as China Cache or Akamai Technologies, their geo-IP results would be the locations of CDN servers rather than the original websites. For the purpose of this research, and based on publicly available information, I have adjusted these geo-IP results to revert to their main servers in China in order to compare the geographic locations of the “ultimate” sources.

The difference between traditional and simplified Chinese character scripts is not only visible but also related to how Chinese modernity and the history of the Chinese Civil War and the Cold War are conceived (P. Chen, 1993; Ji, 2004; L. Z. Lee, 2005; Liao, 2009a; Sui, 2011; Shouhui Zhao & Baldauf, 2007b).

The two major corresponding character encoding standards, gb2312 for simplified Chinese and big5 for traditional Chinese, testify to the history of the separate development of standards between mainland China on one side and Hong Kong, Macau, and Taiwan on the other side. Thus, the generic character encoding-based linguistic categorization scheme works for distinguishing the simplified Chinese gb2312 from the traditional Chinese big5. However, for texts encoded in Unicode, both Chinese scripts can be accommodated. A method is needed to decide whether a given Unicode-encoded text is written mostly in traditional Chinese or in simplified Chinese. To do so, the following formula is proposed:

$$\textit{Deviance from Simplified Chinese} = \frac{\textit{diff}(\textit{orig_text}, \textit{simp_text})}{\textit{diff}(\textit{trad_text}, \textit{simp_text})}$$

What the proposed formula does is to decide how far a Chinese-written text deviates from a simplified Chinese text, thereby providing a reliable categorization method to differentiate simplified Chinese texts from traditional Chinese ones.

To illustrate how the above formula works, in Table 5-2 the first row contains a text with the content of 12 characters, which consists of three non-Chinese characters, four shared Chinese characters (characters that are used in both character scripts), four simplified ones, and one traditional Chinese character. Using conversion algorithms, the original text (*orig_text*) is converted to traditional (*trad_text*, the second row) and simplified (*simp_text*, the third row) texts respectively. The difference between *orig_text* and *simp_text*, or *diff(orig_text, simp_text)*, thus equals one. The difference between *trad_text* and *simp_text*, or *diff(trad_text, simp_text)*, equals five, thereby making the value of *deviance from Simplified Chinese* 20%. This means that the text has 20% of the convertible Chinese characters written in traditional Chinese. The measurement

thus quantifies how far a given text deviates from an all-simplified Chinese text, or equivalently, how close it is to a traditional Chinese text.

Table 5-2

Measuring the deviance from simplified Chinese

Text	Code	Content	Content in Symbols
Original Text	orig_text	1) 人人 <u>对</u> <u>社</u> <u>会</u> <u>负</u> <u>有</u> <u>义</u> <u>务</u> .	# # C C <u>S</u> C <u>T</u> <u>S</u> C <u>S</u> <u>S</u> #
Converted to Traditional	trad_text	1) 人人 <u>对</u> <u>社</u> <u>会</u> <u>负</u> <u>有</u> <u>义</u> <u>务</u> .	# # C C <u>S</u> C <u>S</u> <u>S</u> C <u>S</u> <u>S</u> #
Converted to Simplified	simp_text	1) 人人 <u>对</u> <u>社</u> <u>会</u> <u>负</u> <u>有</u> <u>义</u> <u>务</u> .	# # C C <u>T</u> C <u>T</u> <u>T</u> C <u>T</u> <u>T</u> #

Note. Symbol '#' represents a non-Chinese symbol. Symbol 'C' represents a Chinese character that is used in both. Symbol 'T' represents a traditional Chinese character. Symbol 'S' represents a simplified Chinese character.

5.1.2 Data selection and description

The data set, collected in June 2010, included all external links in the article pages of Baidu Baike and Chinese Wikipedia.⁶

⁶ First, the web page collection software was implemented in Python by using *PycURL*, which used a file transfer library called *libcurl*. Because each article page in Baidu Baike had a sequential number in its URL (e.g. <http://baike.baidu.com/view/48682.htm> for the article “Chinese language”) and each article page in Chinese Wikipedia had a unique page ID, a set of URLs were constructed accordingly to cover all available article pages that were not (or not yet) deleted at the time of data collection. Second, for each article page, the external links were extracted and filtered using further Python-implemented software codes to generate a list of distinct external web links. Third, using similar data collection methods used in the first step, each web object, represented by each of the external links, was collected along with the HTTP response messages for further analysis. I did contribute to Chinese Wikipedia from June 2005 to June 2010, but these edits had negligible impact on the dataset. (For example, the number of my own edits to Chinese Wikipedia article pages during the time period was less than 400 and the majority of edits involve no added citations or links.)

What follows is a brief description of the dataset. Special attention has been paid to the skewed distribution of the data points across geographic and linguistic categories.

All distinct external web links extracted from the article pages constitute the main component of the dataset. The following links were systematically excluded: the internal links, navigation links, links to other Baidu websites (for Baidu Baike) and links to Wikipedia's other sister projects (for Chinese Wikipedia). All other external links were included, including those in the main text and all other sections such as footnotes, references, external websites, etc.

Both the ccTLD and geo-IP categorization methods were then applied to all external links, with tabulated results showing Baidu Baike and Chinese Wikipedia's geographic distribution. The distribution of both categorization results was found to be skewed: A few regions at the top of the distribution had an enormous number of links (hereafter called "link-have-more" regions), whereas massive regions received few links (hereafter called "link-have-less" regions). Both the generic and Chinese-specific linguistic detection methods were also used to categorize each external link into different writing systems and Chinese variants, with missing data rates just above 10%.

Data set. The data set includes three components: (1) the encyclopaedia article pages, (2) the distinct external web links extracted from the article pages, and (3) the externally linked web pages. By the year 2010, the then four-year-old Baidu Baike had more encyclopaedia articles and more external links than the then eight-year-old Chinese Wikipedia, as shown in Table 5-3. Baidu had about six times the number of articles and about twice the number of external links. Thus, Chinese Wikipedia has on average more links per article. Note that not every external web link is well-formed (i.e., following Web standards) and not every well-formed link leads to an external web page that could be accessed for data collection. Still, slightly over 90% of the external links produced web pages that can be analysed.

Table 5-3

Numbers of collected article pages, external web links and pages

	Baidu Baike	Chinese Wikipedia	BB:CW ratios
Articles	2,160,620	362,213	5.97
External web links	1,305,587	719,131	1.82
- well-formed links	1,303,240	719,016	1.81
- retrieved pages	1,174,039	673,790	1.74

External links. Certain external links are systematically excluded from the dataset because they are not considered valid indicators of online citation sources. First, the navigation links, usually outside the main text of the article pages, are justifiably excluded because the research concerns citation and content, not the navigation structure of the website. Second, also excluded are those links that lead to other sister websites hosted by Baidu or the Wikimedia foundation. Thus, as shown in Table 5-4, any web links for Baidu Baike that lead to other Baidu websites are excluded. Similarly for Chinese Wikipedia, any links to other language versions and other sister projects hosted by the Wikimedia Foundation are excluded.

Table 5-4 *Exclusion criteria for the external web links*

Baidu Baike	Chinese Wikipedia		
*.baidu.com	*.wikipedia.org	*.wikitionary.org	*.wikibooks.org
	*.wikisource.org	*.wikimedia.org	*.wikimedia.de
	*.wikinews.org	*.wikiquote.org	*.mediawiki.org

Link-have-more versus link-have-less regions. With the complete external links that fit the research purpose, both the ccTLD and geo-IP categorization methods were conducted to determine the geographic categories for each link. Table 5-5 shows the top 5, top 21 to 25, and top 51 to 55 regions for the geo-IP results. The distribution of the results, for Baidu Baike and Chinese Wikipedia

alike, is skewed. A few “link-have-more” regions at the top have an enormous number of links, whereas massive “link-have-less” regions receive quite few links. Note that proportion-wise, the top five regions already have 93% of Baidu Baike’s total links; for Chinese Wikipedia, the number is 74%. The skewed distribution poses a challenge for analysis. Although on the surface Baidu Baike is dominated by links from China and Chinese Wikipedia is dominated by links from the United States, further unpacking of the data is required to factor out the size and factor in the role of Hong Kong and Taiwan, both of which ranked 3rd or 4th.

Table 5-5

Skewed geographic distribution of links

Ranking		Baidu Baike		Chinese Wikipedia	
1	China	1,017,841	United States	249,106	
2	United States	146,736	China	115,004	
3	Hong Kong	25,427	Taiwan	71,049	
4	Taiwan	8,094	Hong Kong	62,677	
5	Japan	8,051	United Kingdom	39,523	
...		
21	Russia	343	Malaysia	1,611	
22	Pakistan	298	Spain	1,458	
23	Macao	298	Norway	1,396	
24	Malaysia	280	Denmark	1,295	
25	Asia/Pacific Region	251	Belgium	1,230	
...		
50	Estonia	19	Philippines	227	
51	Indonesia	17	Estonia	213	
52	United Arab Emirates	14	Slovakia	206	
53	Bulgaria	10	Iran	194	
54	Iceland	10	Indonesia	187	
55	Romania	10	Ethiopia	185	

5.1.3 Results. Both geographic distribution and linguistic diversities are compared as follows.

Comparing geographic distribution. Showing the contrast in geographic distribution, Table 5-6 lists areas where Baidu Baike and Chinese Wikipedia differ. Each row shows the geo-IP and ccTLD results respectively. Each of the three main columns shows the contrasting categorization of regions, from those where Baidu Baike has more links to the opposite. Since Baidu Baike (BB) has about 1.8 times the number of total external links that Chinese Wikipedia (CW) does, the middle column contains regions where Baidu Baike has more links than Chinese Wikipedia but not proportionally more (i.e. the ratio value $r = BB/CW$ is smaller than 1.8 but larger than 1).

Table 5-6

Comparing the geographic distribution of links

	Baidu Baike has more				Chinese Wikipedia has more	
	Cases ^a	N	Cases ^b	N	Cases	N
geo-IP	CN*, BS†, KY, BZ, NE, SD	6	TH, CU	2	US*, TW*, HK*, GB*, JP*, NL*, DE*, FR†, CA†, IT†, CH†, AU†, SE†, IE†, RU†, MO†, NZ†, MY†, ES†, NO†, BE†, DK†, FI†, SG†, and 158 other countries	182
ccTLD	cn*, la, me, vc, sh, cm, gd, io, ly, ne, dj	11	cc, nu, fm, ru, gg	5	tw*, jp*, hk*, fr†, de†, au†, ca†, kr†, ru†, nl†, se†, it†, mo†, us†, ch†, be†, and 201 other countries	217

Note. All countries are ordered from largest to smallest in terms of the difference. Countries are represented by the ISO country codes (upper/lower case for geo-IP/ccTLD results), except for countries such as the U.K., which has the country code (GB) different from ccTLD (uk).

ISO country codes: AU(Australia), BE(Belgium), BS(Bahamas), BZ(Belize), CA(Cocos Islands), CC(Cocos (Keeling) Islands), CH(Switzerland), CM(Cameroon), CN(China), DE(Germany), DJ(Djibouti), DK(Denmark), ES(Spain), FI(Micronesia), FM(Micronesia, Federated States of), FR(France), GD(Grenada), GG(Guernsey), HK(Hong Kong), IE(Ireland), IO(British Indian Ocean Territory), IT(Italy), JP(Japan), KR(South Korea), KY(Cayman Islands), LA(Laos), LY(Libya), ME(Montenegro), MO(Macau), MY(Malaysia), NE(Niger), NL(Netherlands), NO(Norway), NU(Niue), NZ(New Zealand), RU(Russia), SD(Sudan), SE(Sweden), SG(Singapore), SH(Saint Helena), TW(Taiwan)

^a Both absolutely and proportionally more (i.e. ratio $r = BB/CW > 1.82$)

^b Absolutely more but not proportionally more (i.e. $1 < r < 1.82$)

* Countries where the difference is over 10,000 links

† Countries where the difference is over 1,000 links

Both the number of countries (N) and selected countries are shown for each categorization in Table 5-6. Chinese Wikipedia has more links than Baidu Baike for most regions (over 180 for geo-IP and over 210 for ccTLD).

The asterisk symbol in Table 5-6 signifies the countries where the difference is the widest. In the first column, China marks the extreme advantage of Baidu Baike over Chinese Wikipedia. In the third column, it is the U.S., Taiwan, Hong Kong, the U.K., Japan, France, and many other countries where Chinese Wikipedia has a clear lead. Since Taiwan and Hong Kong are areas where traditional Chinese language scripts are used, it is to be expected that the linguistic analysis of the web links will also reflect these geographic differences.

Comparing linguistic diversities. Two pie charts Figure 5-1 and Figure 5-2 are presented to compare the linguistic diversities of the external link data set.

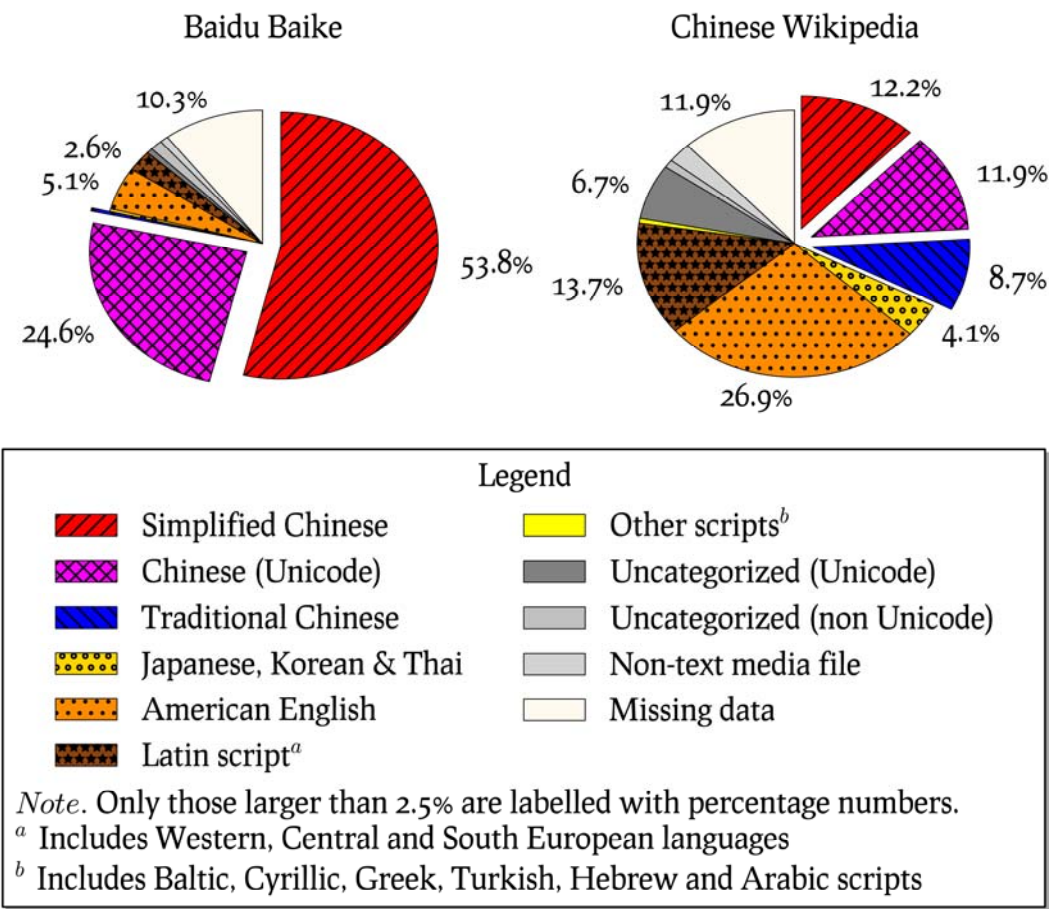


Figure 5-1. Language scripts detected: all language scripts

The pie charts in Figure 5-1 show that more than half of Baidu Baike’s external links point to simplified-Chinese-only content, whereas fewer than 2% of the links point to traditional-Chinese-only content. In contrast, Chinese Wikipedia’s links are more diverse and balanced, with 12.2% simplified-Chinese-only and 8.7% traditional-Chinese-only content. Around 10% of the links appear to be dead or broken, which is a common phenomenon and does not alter the overall comparative findings. Using the aforementioned *deviance from Simplified Chinese* measurement, the Unicode portion can be further unpacked.

Showing only the online sources in the category of East Asian languages, Figure 5-2 confirms the expectation that Baidu Baike prefers simplified Chinese (up to 96.6% = 68.0% + 28.6%), whereas Chinese Wikipedia is more balanced. Traditional Chinese and Japanese take up sizable portions of the external links of Chinese Wikipedia, but much less so for Baidu Baike’s.

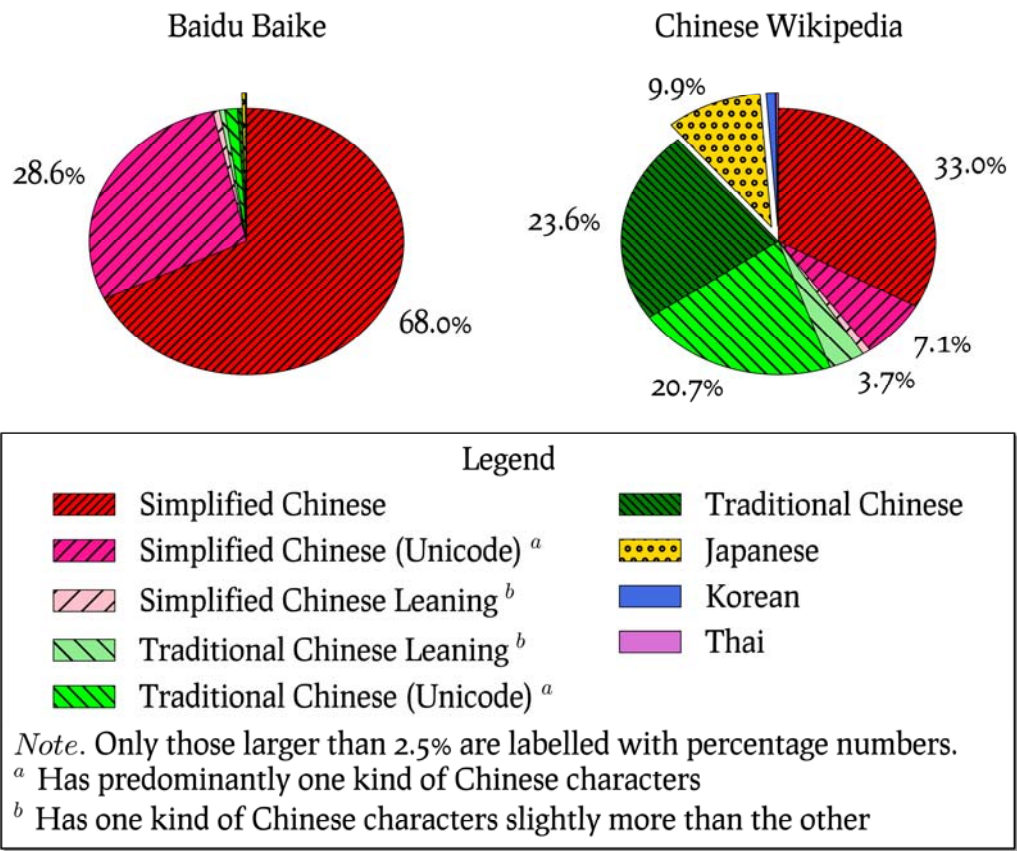


Figure 5-2. Language scripts detected: unpacking the Unicode-encoded results

So far, both the overall geographic and linguistic analysis has been helpful in illustrating one consistent observation: Chinese Wikipedia has more linguistically and geographically diverse online sources than Baidu Baike does. The overall results show how differently the two encyclopaedias link to the world's knowledge sources. The differences indicate the overall editorial preferences of the two user-generated encyclopaedias, suggesting how differently the two websites engage online sources.

5.1.4 Discussion. This chapter has so far sought to explore the influence of geographic and linguistic factors on online civic engagement dynamics across Chinese societies, and presented the findings of the two million web links collected in June 2010 from all articles in Baidu Baike and Chinese Wikipedia. The geolinguistic analysis shows strong and different Chinese localization effects in filtering the world's information. The findings reveal that the difference between the two is less about the "level" of Chineseness (Chinese versus non-Chinese) and more about the geographic and linguistic "extent" of Chineseness (mainland China versus the Chinese-speaking world combined). These two million external links embody the collective editorial judgement on reliable sources for encyclopaedias contributed by online volunteer editors (thus "network gatekeeping", see Chapter 4), thereby showing the reach of civic engagement online.

Thus, the geolinguistic analysis demonstrates how large datasets can be processed to analyse cultural-political patterns wherever geographic and/or linguistic factors are relevant. Although the analysis represents only a partial picture of how user-generated content websites have filtered the world's information, geolinguistic analysis can provide important insights: Baidu Baike's external links are concentrated on simplified Chinese content, mostly hosted in mainland China, whereas Chinese Wikipedia's are more dispersed and diverse, including much more traditional Chinese content and content hosted in regions such as the U.S., Japan, Hong Kong, Taiwan, etc.

5.2 Size and institutional considerations

Additional insights can be gained by further considering the size and institutional factors.

One way to deal with the skewed distribution is to factor out the size. Indeed, the geographic mapping in the previous section, while useful in many ways, has two major drawbacks: Some regions are simply too small to be shown on the global map despite their potential analytical significance. Thus, a degree of normalization is required to achieve some “per-size” comparison that is similar to the “per-capita” comparison commonly used in the social sciences. For the purpose of this thesis, three general and independent size indicators can be used to derive several “per-size” comparison outcomes. Thus, each of the regions can be better compared with one another (e.g. Macau versus Taiwan or mainland China versus the U.S.) and at least three kinds of “size” can be considered to check the results in order to avoid arbitrary comparisons.

Another way to analyse the findings further is to compare the top-linked websites because they indicate preferred choices for online sources. The top 15 websites will thus be analysed for each region that could make a difference in distinguishing Baidu Baike and Chinese Wikipedia. Then, using publicly available information and my own experience as a heavy user of online and offline Chinese-language content, I code the websites into categories of commercial, educational, governmental, or non-profit websites. This way, I can assess the nature and quality of online sources linked to by both encyclopaedias. Special attention will be paid to the fact that many institutions in China, although not belonging directly to Chinese government authorities, are affiliated with them. Since it is not feasible to compare all websites, comparing the top-linked sites should provide adequate results.

It should be noted that websites do not always correspond to the institutions of information sources, and that linking to a website is not always equivalent to citing a source. Linking is citing for certain websites, such as the

official media websites of the *New York Times* or *People's Daily*. Yet linking is not exactly citing for information platforms such as archives, databases, or platform websites, including Google Books and YouTube. Researchers must distinguish linking from citing when interpreting webometric findings, as the difference is likely to be significant to both the subject matter and institutional nature of the website. For example, Google Books is a commercial website (institutional trait) that provides information on books (subject or topical trait), and *People's Daily* is mainly an official medium of the Chinese Communist party (institutional trait) that provides news and official statements (subject or topical trait)

5.2.1 Geographic normalization. Geographic normalization can now be used for cross-country comparisons. Geographic normalization, or simply data normalization, allows data to be compared using a meaningful common denominator, thereby producing measurements of intensity or density, such as population density (American Planning Association, 2006; Cote, n.d.). Such normalization is particularly useful in “factoring out the size” in order to facilitate comparisons across unequal areas or populations (Cote, n.d.). Since the findings aggregated at the country level are likely to be biased toward large-sized countries, some size measurements may be selected as meaningful common denominators against which data normalization can be performed.

For geographic normalization, I use three general size measurements (GDP, population, and Internet population) because a country can be expected to receive more external links if it is larger in size. I use the 2008 data provided by the International Monetary Fund and the International Telecommunication Union, including the Gross Domestic Product (GDP) numbers derived from purchasing power parity (PPP) as the indicator of economic size. These size measurements can then serve as common denominators that can be used to compare per-size intensity, resulting in measurements per GDP, per Internet user, and per capita for cross-region comparison.

Since to compare all the regions of the world would take efforts that are likely unnecessary for the study, only mainland China and twenty other regions have been selected for comparison. Covering a diverse set of regions, the selection includes the following: regions with sizable Chinese-speaking populations such as Hong Kong, Macau, Taiwan, Singapore, and Malaysia; major Pacific countries such as the U.S., Japan, South Korea and Australia; large developing countries such as Brazil, Russia, and India; one small developed country, the Netherlands; ex-Communist countries such as Vietnam, Czech Republic, and Poland; Middle East countries such as Iran, Israel, and the United Arab Emirates; and one African country, South Africa.

Table 5-7 shows the results across the 21 selected regions in four major sections, from top to bottom: pre-normalized outcomes, per GDP outcomes, per Internet user outcomes, and per capita outcomes. To compare the geographic distribution *across countries*, the values of each cell are normalized against the maximum number within that row. Within each section, Baidu Baike and Chinese Wikipedia's respective ccTLD and geo-IP results are compared. The first section (pre-normalized outcomes N) shows that Baidu Baike is concentrated heavily in mainland China, and that the most-salient differences between the two encyclopaedias include regions such as the U.S., Taiwan, Japan, and Hong Kong, all of which receive a significant number of links from Chinese Wikipedia but not from Baidu Baike.

Table 5-7

Comparing distribution of links across selected regions

Pre-normalized and normalized comparisons of Baidu Baike (BB) and Chinese Wikipedia (CW)			United States	China	Hong Kong	Macao	Taiwan	Japan	South Korea	Singapore	Malaysia	Vietnam	Australia	India	Brazil	Russia	Netherlands	Czech Republic	Poland	Iran	Israel	United Arab Emirates	South Africa
Pre-normalized (N)	BB	cc																					
		TLD																					
		geo-IP																					
	CW	cc																					
		TLD																					
		geo-IP																					
Normalized by GDP (N/GDP)	BB	cc																					
		TLD																					
		geo-IP																					
	CW	cc																					
		TLD																					
		geo-IP																					
Normalized by Internet population (N/pop_inter net)	BB	cc																					
		TLD																					
		geo-IP																					
	CW	cc																					
		TLD																					
		geo-IP																					
Normalized by population (N/pop)	BB	cc																					
		TLD																					
		geo-IP																					
	CW	cc																					
		TLD																					
		geo-IP																					

Note. For the purpose of comparison by highlighting the relative differences within each row, the values of each cell are normalized first against the maximum value in each row.

The normalized per-size results, shown in the last three sections of Table 5-7 provide new insights. For the two encyclopaedias, the U.S. ceases to be the differentiating region, whereas regions such as Hong Kong, Macau, and Taiwan stand out. The clear contrast strongly indicates what differentiates the two encyclopaedias: China, on one hand, and Hong Kong, Macau, and Taiwan on the other. This observation confirms the previous overall geographic analysis and is consistent in complementing the aforementioned linguistic analysis of the contrasting factor of simplified versus traditional Chinese scripts. The per-size findings reject the notion that the difference between Baidu Baike and Chinese Wikipedia is along the lines between China and the U.S. and indicates that the difference is along the line between mainland China on one side and the rest of the major Chinese-speaking regions on the other (echoing findings in Chapter 4).

The comprehensive geographic and linguistic findings show that the difference between Baidu Baike and Chinese Wikipedia is less about the “level” of Chineseness and more about the geographic and linguistic “extent” of Chineseness. Indeed, Baidu Baike focuses more on simplified Chinese content and content hosted in mainland China. However, it is not inclusive enough to include Hong Kong, Macau, and Taiwan, the main Chinese-speaking regions that Beijing is keen to influence and integrate politically. In contrast, Chinese Wikipedia is culturally and linguistically Chinese with substantial links to traditional-Chinese content plus websites hosted in Hong Kong, Macau (officially part of China), and Taiwan (claimed to be “Chinese” by Beijing). The findings thus clearly identify the gaps in Baidu Baike’s engagement with Chinese-written information outside mainland China. Chinese Wikipedia is shown to be an example of what Yang Guobin (2003) described as the rise of a “transnational Chinese cultural sphere” as it relates to the Internet. Further research efforts are still required to better quantify and qualify “Chineseness” online in relation to online civic engagement. Still, the methods and findings here fill an important gap in the existing efforts to understand Chinese civic

engagement online, which are often limited by national frameworks. As indications of cultural-thickening patterns, the Web linkage of the two websites clearly indicates two distinct Chinese cultural spheres or web spheres.

5.2.2 The top 15 websites for mainland China and the U.S. To provide a further check on the findings presented so far, it will be useful to examine the five regions that are critical for determining whether the main difference is China versus non-Chinese or mainland China versus the Chinese-speaking world. Since the top-linked websites should indicate preferred choices for sources for each region, they should tell us how the two encyclopaedias filter information from these critical regions. To enable a straightforward comparison, the respective numbers of links for the two encyclopaedias are listed side by side as N, N(CW) and N(BB) for each website listed.

China-hosted websites. Table 5-8 lists the top-15 most linked websites hosted in mainland China.

Table 5-8

Top-linked China-based websites

Rank- ing	Baidu Baike (BB)			Chinese Wikipedia(CW)		
	Website	N	N(CW)	Website	N	N(BB)
China (CN)	1 tushucheng.com	43210	> 4	bioinfo.cn	26008	> 327
	2 sina.com.cn	33842	> 6884	xinhuanet.com	8780	< 12830
	3 yoostrip.com	26878	> 0	sina.com.cn	6884	< 33842
	4 ilucking.com	23011	> 1	people.com.cn	2811	< 6155
	5 tushulian.com	22707	> 0	163.com	1761	< 8372
	6 51966.com	20558	> 0	qq.com	1046	< 8044
	7 xinhuanet.com	12830	> 8780	cntv.cn	1002	< 4730
	8 xzqh.org	12234	> 409	beijing2008.cn	902	> 867
	9 agri.com.cn	11662	> 48	chinanews.com.cn	888	< 1349
	10 worldpersondictionary.com	10955	> 3	dfo.cn	888	> 22
	11 gsdkj.net	10953	> 1	delta-intkey.com	695	> 0
	12 elong.com	9981	> 14	china.com.cn	643	< 1787
	13 beijingtushucheng.com	9467	> 0	tom.com	554	< 2221
	14 163.com	8372	> 1761	sohu.com	554	< 4342
	15 qq.com	8044	> 1046	www.gov.cn	495	< 3724

Baidu Baike. E-commerce websites dominate Baidu Baike's links: four for books: {1}tushucheng.com, {4}ilucking.com, {5}tushulian.com, and

{13}beijingtushucheng.com; three for travel/hotel-booking: {3}yoostrip.com, {6}51966.com, and {12}elong.com; one on agriculture: {9}agri.com.cn (hereafter, the numerical values in the braces indicate the ranking in the table). These eight websites alone contribute to 17% of the total links. Among the remaining top 15 websites, four are major portal websites: {2}sina.com.cn, {7}xinhuanet.com, {14}163.com, and {15}qq.com, and one, xinhuanet.com, is owned directly by the Chinese government. The others are topic-specific websites: {8}xzqh.org (administrative regions), {10}worldpersondictionary.com (a dictionary of famous people), and {11}gsdkj.net (mining related). Comparing the numbers of N and N(CW) clearly indicates that Chinese Wikipedia seems to filter out the e-commerce websites that Baidu Baike has.

Chinese Wikipedia. No e-commerce website dominates Chinese Wikipedia's links. Among those listed are nine major portal websites and news websites: {2}xinhuanet.com, {3}sina.com.cn, {4}people.com.cn, {5}163.com, {6}qq.com, {7}cntv.cn, {9}chinanews.com.cn, {12}china.com.cn, and {14}sohu.com. It should be noted that five of these are news portals run by party and state media in China ({2}, {4}, {7}, {9}, {12}), outnumbering those run by private owners. The remainder includes three mostly sports-related websites: {8}beijing2008.cn (Beijing Olympics), {10}dfo.cn (German soccer), and {13}tom.com (NBA basketball); two academic databases, {1}bioinfo.cn and {11}delta-intkey.com; and one official government website, {15}www.gov.cn. As the findings indicate, what distinguishes Chinese Wikipedia from Baidu Baike is not about China but rather about the issue of whether e-commerce websites are appropriate sources for user-generated encyclopaedias.

Discussion. The findings above thus challenge the perceived notion that Chinese Wikipedia is less Chinese than Baidu Baike is. Baidu Baike's content is flooded with "infomercial content" contributed by e-commerce websites, an observation that is supported by my further examination (below) of Baidu Baike's articles that contain those links. This concentration among the top links

is also supported by the existence of tips on using Baidu Baike for Web promotion (Língcái Web Promotion, 2010). “Infomercial” links even outrank major portal websites among Baidu Baike’s external links. In contrast, nine out of Chinese Wikipedia’s top 15 websites are major Chinese portal websites, with one-third run by party and state officials in China, thereby rejecting the notion that Chinese Wikipedia is not Chinese enough or shies away from citing official Chinese sources. The difference between the two encyclopaedias is less about their Chineseness or about the official-ness of sources and more about their failure or success in excluding or including infomercial links (and the content associated with them).

U.S.-hosted websites. Since many major global and Chinese-language websites are hosted in the U.S., a comparison on the most-linked U.S.-hosted websites is needed. Table 5-9 shows the rather diverging outcomes.

Table 5-9

Top-linked U.S.-hosted websites

	Rank- ing	Baidu Baike (BB)			Chinese Wikipedia (CW)		
		Website	N	N(CW)	Website	N	N(BB)
United States (US)	1	wikipedia.org	10588	> 0	nih.gov	7965	> 25
	2	souezu.cn	2281	> 1	doi.org	6857	> 4
	3	5d6d.com	2209	> 15	sil.org	6809	> 0
	4	nba.com	1569	> 1411	google.com	6034	> 417
	5	hao565.cn	1395	> 0	imdb.com	4751	> 1046
	6	chinaexpertsweb.net	1379	> 0	youtube.com	4359	> 2
	7	asian-chinese-african.org	1346	> 0	nasa.gov	2654	> 97
	8	qdgqtv.cn	1276	> 0	harvard.edu	2570	> 3
	9	xikao.com	1063	> 22	caltech.edu	2458	> 6
	10	imdb.com	1046	< 4751	sed.s.org	2308	> 3
	11	ey800.cn	998	> 0	uefa.com	1835	> 26
	12	world-culture-research.org	979	> 1	wikia.com	1806	> 25
	13	eb.com	910	> 71	nytimes.com	1805	> 66
	14	doudouditu.cn	876	> 0	blogspot.com	1721	> 732
	15	google.cn	867	> 105	skysports.com	1711	> 0

Baidu Baike. For the U.S., Baidu Baike’s top-linked website is Wikipedia, and a further breakdown of the data shows that most of the links are pointing to Chinese, English, and Japanese versions (61.60%, 26.05%, and 9.72%, respectively).

The fact that Baidu Baike links to various versions of Wikipedia provides additional evidence that Baidu Baike has copied content from Wikipedia projects, thereby confirming the findings in Chapter 4.

The remaining websites can be categorized roughly as follows. First, for websites that are linked by Baidu Baike and also heavily linked by Chinese Wikipedia, I found a major sport website, {4}nba.com, and a major movie website, {10}imdb.com. Second, for websites that are rarely linked by Chinese Wikipedia (less than 20), I found ten websites, most of which are primarily Chinese-language websites on different topics: {2}souezu.cn (general search), {3}5d6d.com (free online forums), {5}hao565.cn (youth portal), {6}chinaexpertsweb.net (human resources), {7}asian-chinese-african.org (crafts e-commerce), {8}qdgqtv.cn (online video), {11}ey800.cn (pharma-medical), {12}world-culture-research.org (crafts e-commerce), and {14}doudouditu.cn (maps). Again, Baidu Baike included spam links of infomercial content. Third, for websites that receive some links from both encyclopaedias, I found three: {9}xikao.com (on traditional Chinese opera), {13}eb.com (Encyclopædia Britannica Online), and {15}google.cn (Google China). Hence, Baidu Baike is found to have a substantial number of links to U.S.-based websites, including Wikipedia projects; global sports, movies, and encyclopaedia websites; and, again, other Chinese-language infomercial websites.

Chinese Wikipedia. In contrast, Chinese Wikipedia's top 15 exclude e-commerce websites and include major information and educational websites across knowledge domains. (Note that other Wikimedia projects are precluded by research design). The top 9, as expected, are for encyclopaedic references: {1}nih.gov (a U.S. government-funded health research agency), {2}doi.org (a non-profit organization in charge of the Digital Object Identifier system, {3}sil.org (a non-profit for authoritative linguistic information called SIL International), {4}google.com (a major Internet company that provides search services for books and scholarly works), {5}imdb.com (a major information website on movie

databases), {6}youtube.com (a major global website for online videos), {7}nasa.gov (a U.S. agency on space projects), and {8}harvard.edu and {9}caltech.edu (two major U.S. universities). Further unpacking shows that these websites contain digital archives or catalogues for extensive linking. For example, the top-linked NIH services website is the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov), with 75% of the total number of links to {1}nih.gov; the top-linked Google service is Google books, which garners 34% of the total number of links to {4}google.com; the second top-linked NASA service is the National Space Science Data Center, which gets 11% of the total number of links to {7}nasa.gov. For others, the top-linked harvard.edu service is NASA Astrophysics Data System (adsabs.harvard.edu), with 74% of the total number of harvard.edu links; the top-linked caltech.edu service is NASA/IPAC Extragalactic Database (ned.ipac.caltech.edu), with 92% of the total number of caltech.edu links. Chinese Wikipedia's interest in linking to space-related materials hosted in the U.S. is further confirmed by {10}seds.org, an international student organization focused on space exploration.

The remaining websites include {12}wikia.com (a major wiki-hosting website), {14}blogspot.com (a major blog-hosting website), {13}nytimes.com (a major news portal), {11}uefa.com and {15}skysports.com (two major sports portals). Thus, the findings suggest the concentration of authoritative (e.g. space and astrophysics databases) or comprehensive sources (e.g. movies, videos, and books) on these U.S.-hosted websites. Further examination of Chinese Wikipedia's Google Books and IMDb links shows a substantial amount of Chinese-language content due to the nature of encyclopaedic content. One would expect that Baidu Baike would have a substantial number of links to these websites as well, but comparing the numbers of N and $N(BB)$, in fact, except for {5}imdb.com, Baidu Baike has few or no links to them. Because Baidu Baike has a large number of links to Chinese and English Wikipedia, and it has more links to NBA.com than Chinese Wikipedia, it would be difficult to claim that Baidu

Baike is less “American” than Chinese Wikipedia. The findings also demonstrate the dominant position of certain U.S. websites for encyclopaedia references because they provide the world’s most-comprehensive or authoritative information in certain knowledge domains.

5.2.3 The top 15 websites for Hong Kong, Taiwan, and Macau. Since the per-size comparison has shown that Baidu Baike and Chinese Wikipedia diverge on the distribution of external links to Hong Kong, Taiwan, and Macau, the following sub-sections examines how they differ.

Hong Kong-hosted websites. Table 5-10 shows that Baidu Baike has a much lower number of links for Hong Kong-hosted websites.

Table 5-10

Top-linked Hong Kong-hosted websites

Rank- ing	Baidu Baike (BB)			Chinese Wikipedia(CW)		
	Website	N	N(CW)	Website	N	N(BB)
Hong Kong (HK)	ysbooking.com	1553	> 0	yahoo.com	8433	> 384
	meiweizhongguo.com	1058	> 0	tvb.com	5626	> 256
	zzfanwen.cn	825	> 0	wretch.cc	2272	> 193
	zhuhere.com	716	> 0	info.gov.hk	1540	> 40
	gglsw.cn	516	> 0	rthk.org.hk	1429	> 30
	hxcxgl.com	481	> 0	atnext.com	1156	> 16
	kuliu.com	433	> 0	hkfa.com	1152	> 4
	axdfz.gov.cn	415	> 6	mtr.com.hk	1128	> 17
	yahoo.com	384	< 8433	hkex.com.hk	1028	> 7
	fushantang.com	282	> 30	sina.com.hk	983	> 11
	442.cn	267	> 0	681busterminal.com	941	> 0
	52pk.com	260	> 0	on.cc	800	> 8
	fjqhdmw.com	259	> 7	nextmedia.com	736	> 1
	tvb.com	256	< 5626	legco.gov.hk	731	> 2
	whydp.com	222	> 0	lcsd.gov.hk	683	> 79

Baidu Baike. Again, e-commerce and infomercial websites dominate Baidu Baike’s links: four of them are for travel/hotel-booking: {1}ysbooking.com, {3}zzfanwen.cn, {4}zhuhere.com, and {7}kuliu.com. The rest are for different topics: {2}meiweizhongguo.com (food), {5}gglsw.cn (law), {6}hxcxgl.com (business consulting), {8}axdfz.gov.cn (tea), {10}fushantang.com (Kungfu and

fashion), {12}442.cn (football statistics), {12}52pk.com (games), {13}fjqhdmw.com (now an adult dating site), and {15}whyyp.com (travel information). Their impact on Baidu Baike's content appears to be questionable; for example, many of the corresponding hotel-booking encyclopaedia articles contain only basic introductions to a single hotel. Again, comparing the numbers of N and N(CW), Chinese Wikipedia succeeds in filtering out these e-commerce or infomercial websites. Only two websites appear to be more-legitimate, Hong Kong-based media websites: {9}yahoo.com (online media) and {14}tvb.com (commercial TV media), each of which also has a substantial link presence in Chinese Wikipedia (over 5,500 links).

Chinese Wikipedia. Major Hong Kong-hosted websites are included in Chinese Wikipedia. Among those listed are not only the aforementioned popular Hong Kong-based media websites, {1}yahoo.com and {2}tvb.com, but also other media websites including {5}rthk.org.hk (the public broadcasting media Radio Television Hong Kong, or xiānggǎng diàntái 香港電台); {6}atnext.com (commercial media Next Media Limited, or yī chuánméi 壹傳媒), {12}on.cc (commercial print media Oriental Press Group, or dōngfāng bào yè jítuán 東方報業集團), and {13}nextmedia.com (commercial media Next Media Limited); and {3}wretch.cc (a community web site or wú míng xiǎo zhàn 無名小站) and {10}sina.com.hk (online media Sina Hong Kong or xiānggǎng xīnlàng 香港新浪). In addition, three governmental websites are listed: {4}info.gov.hk, {14}legco.gov.hk, and {15}lcsd.gov.hk. The remaining information websites are {7}hkfa.com (the non-profit Hong Kong Football Association), {9}hkex.com.hk (Hong Kong Exchanges), {8}mtr.com.hk, and {11}681busterminal.com (public transportation). Comparing the numbers of N and N(BB), each listed website has a significant number of links, suggesting that Chinese Wikipedia is more “encyclopaedic” for Hong Kong with a mixture of sources. The findings here not only reinforce the contrast of keeping versus excluding e-commerce websites, but also support the Mainland China versus outside (as opposed to China versus the

world) hypothesis because Hong Kong was returned to China in 1997.

Taiwan-hosted websites. Table 5-11 also shows that Baidu Baike has a much lower number of links to Taiwan-hosted websites.

Table 5-11

Top-linked Taiwan-hosted websites

	Rank- ing	Baidu Baike (BB)			Chinese Wikipedia(CW)		
		Website	N	N(CW)	Website	N	N(BB)
Taiwan (TW)	1	sinica.edu.tw	1493	< 4460	sinica.edu.tw	4460	> 1493
	2	npm.gov.tw	1394	> 71	libertytimes.com.tw	2983	> 5
	3	cnave.com.tw	527	> 0	udn.com	2048	> 30
	4	yahoo.com	241	< 1849	yahoo.com	1849	> 241
	5	wordpedia.com	208	> 15	chinatimes.com	1845	> 12
	6	pili.com.tw	169	> 87	yam.com	1686	> 54
	7	tc.edu.tw	99	< 754	tse.com.tw	1474	> 1
	8	pixnet.net	93	< 676	tp.edu.tw	1142	> 32
	9	kingnet.com.tw	92	> 59	tku.edu.tw	1071	> 16
	10	joypark.com.tw	82	> 50	nownews.com	977	> 16
	11	ttv.com.tw	71	< 499	ntu.edu.tw	858	> 42
	12	digitalarchives.tw	69	> 45	tc.edu.tw	754	> 99
	13	colorbird.com	54	> 0	pixnet.net	676	> 93
	14	yam.com	54	< 1686	ly.gov.tw	626	> 0
	15	woo.com.tw	51	< 292	nccu.edu.tw	579	> 16

Baidu Baike. Similar to or even worse than the Hong Kong-hosted findings, only the top 3 Taiwan-based websites have more than 250 links, while almost all the top 15 Hong Kong-based ones have over 250 links for Baidu Baike. Given the low numbers for Baidu Baike, only the top five are discussed here. The top two are {1}sinica.edu.tw (Academia Sinica) and {2}npm.gov.tw (the National Palace Museum, guólì gùgōng bówùyuàn, 國立故宮博物院), both of which host major digital archives and databases of an academic and knowledge-based nature. The remaining top five include {3}cnave.com.tw (an e-commerce website for Chinese music); {4}yahoo.com (a major portal website); and {5}wordpedia.com (an encyclopaedia website that provides paid online access to Encyclopaedia Britannica, “China Encyclopaedia”, and “Taiwan Encyclopaedia”). Baidu Baike does not include a substantial number of Taiwan-based websites as its sources.

Chinese Wikipedia. In contrast, Chinese Wikipedia includes many Taiwan-based sources, such as the aforementioned {1}sinica.edu.tw and {3}yahoo.com. Included also are several major Taiwan-based newspaper websites: {2}libertytimes.com.tw (the *Liberty Times*, or zìyóu shíbào 自由時報), {4}udn.com (the *United Daily News* or liánhé bào 聯合報) and {5}chinatimes.com (the *China Times* or zhōngguó shíbào 中國時報). Also listed are one television website, {15}tvbs.com.tw, and three online media websites, {6}yam.com, {10}nownews.com, and {12}pixnet.net. Similar to the findings for Hong Kong, Chinese Wikipedia's top websites also include the official website on stocks, {7}tse.com.tw (the Taiwan Stock Exchange). Listed educational websites include three university websites, {9}tku.edu.tw, {11}ntu.edu.tw, and {14}nccu.edu.tw, and a set of national school websites in Taipei, {8}tp.edu.tw. The only government website in the top 15 is the Legislative Yuan of Taiwan, {13}ly.gov.tw. Again, comparing the numbers of N and N(BB), Chinese Wikipedia covers more diverse sources from Taiwan, including major news, educational, and governmental websites that have little presence on Baidu Baike.

Macau-hosted websites. Table 5-12 shows the findings for Macau.

Table 5-12

Top-linked Macau-hosted websites

Rank- ing	Baidu Baike (BB)			Chinese Wikipedia(CW)		
	Website	N	N(CW)	Website	N	N(BB)
Macao (MO)	1 macaudata.com	182	< 294	macaudata.com	294	> 182
	2 icm.gov.mo	15	< 17	macaodaily.com	188	> 3
	3 cityguide.gov.mo	14	< 49	io.gov.mo	135	> 6
	4 iacm.gov.mo	11	< 44	tdm.com.mo	129	> 1
	5 umac.mo	6	< 20	ctm.net	88	> 1
	6 macau99.org.mo	6	< 10	gcs.gov.mo	80	> 0
	7 io.gov.mo	6	< 135	smg.gov.mo	79	> 3
	8 www.gov.mo	5	< 35	dsat.gov.mo	57	> 0
	9 dsec.gov.mo	4	< 25	macau.gov.mo	54	> 0
	10 namkwong.com.mo	3	> 0	cityguide.gov.mo	49	> 14
	11 smg.gov.mo	3	< 79	macauheritage.net	48	> 2
	12 al.gov.mo	3	< 13	iacm.gov.mo	44	> 11
	13 macaodaily.com	3	< 188	www.gov.mo	35	> 5
	14 macauheritage.net	2	< 48	saftp.gov.mo	34	> 2
	15 safp.gov.mo	2	< 34	macautourism.gov.mo	33	> 1

Baidu Baike. Similar to the findings for Hong Kong and Taiwan, Baidu Baike has far fewer links than Chinese Wikipedia. Some of the patterns found in Hong Kong and Taiwan also apply here. First, Baidu Baike has a much smaller number of links: for almost any of Baidu Baike's top 15 Macau-based websites, Chinese Wikipedia has more links, except for one corporation website, {10}namkwong.com.mo, providing additional evidence that Baidu Baike fails to filter out promotional content.

Chinese Wikipedia. In terms of both quantity and diversity, Chinese Wikipedia links Macau better. Comparing the numbers of N and N(BB) in the right half of Table 5-12, Baidu Baike fails to give expected links to Macau's major newspaper, TV, and online media, including {13}Macaudaily.com (*Macau Daily*), {4}tdm.com.mo (Macau Broadcasting TV), and {5}ctm.net (Macau's major online portal website). In addition to these media websites, Chinese Wikipedia's top 15 for Macau include other government information websites, ranging from {6}gcs.gov.mo (news) and {7}smg.gov.mo (weather) to {8}dsat.gov.mo (transportation), all of which receive few or no links from Baidu Baike.

5.2.4 Discussion. The results above have accounted for the factors of size and the institutions of external links, thereby providing greater detail that marks the differences between Baidu Baike and Chinese Wikipedia. Both geographic and linguistic findings have independently indicated that the most salient difference is not between China and the U.S., but rather between mainland China and the other three Chinese-speaking regions of Hong Kong, Taiwan, and Macau. Moreover, the U.S. findings challenge the binary notion of "Chinese" Baidu Baike versus "American" Chinese Wikipedia and support the observation of Chinese Wikipedia as more "encyclopaedic" and "international". The comparison of the top 15 most-linked websites suggests a strong contrast in relation to the institutions linked. Baidu Baike's external links are prone to e-commerce and infomercial links across regions, whereas Chinese Wikipedia's

cover the major news, governmental, and informational websites of Hong Kong, Taiwan, and Macau that Baidu Baike overlooks.

Thus, it can be concluded that Baidu Baike is subject to both the commercial interests and the geographic boundary of mainland China, whereas Chinese Wikipedia covers the major websites more substantially for four specific Chinese-speaking regions: mainland China, Hong Kong, Taiwan, and Macau. Note, as shown in Table 5-7, that the per-size normalized results indicate that Chinese Wikipedia cites the regions of Hong Kong, Taiwan, and Macau even more than the information-rich U.S., providing yet another indication that Chinese Wikipedia is “Chinese” in the sense that it engages knowledge sources from these Chinese-speaking regions. Thus, conceptualized in terms of “civic learning repertoires” for “information engagement” (Bennett & Wells, 2009) built by user-contributors to filter and make use of information and content that is deemed reliable, these external links reflect the location and content preferences of the “civic learning” involved.

The role of the U.S. is also addressed here: for both encyclopaedias, the U.S.-hosted websites reflect the dominant role of the U.S. as the world’s information hub instead of its American values. The fact that Baidu Baike has many links to Chinese-language websites in the U.S. (including its competitor, Chinese Wikipedia, as the top-linked website) suggests that the notion that U.S.-hosted websites may somehow decrease the level of Chineseness is problematic.

From these findings, different cultural-thickening patterns can be derived: in terms of regions, both encyclopaedias converge roughly on the online sources in mainland China but diverge on those in Hong Kong, Taiwan, Macau, and the U.S. For institutions, both converge on almost all major online sources in mainland China and some in the U.S., but they diverge on major news, government, and information websites in Hong Kong, Taiwan, and Macau. Hence Baidu Baike’s cultural-thickening patterns, as shown by its external links, can be described as “commercially dominant” and “Mainland-centric”, whereas

Chinese Wikipedia's can be described as more "encyclopaedic" and closer to a "transnational Chinese cultural sphere" (G. Yang, 2003). Regions of Hong Kong, Macau, and Taiwan and their traditional Chinese language scripts are shown to matter the most (and more so than the U.S. when the findings are normalized per size), suggesting that the difference between the two is mainland China versus the Chinese-speaking world combined, which confirms the findings comparing the geographic locations of the power users in Chapter 4. It is difficult to discern whether the omission of various sources here is the direct outcome of editorial decisions made by user-contributors, the indirect outcome of Mainland Chinese contributors who have little access to sources outside mainland China, or the outcome resulting from Baidu's employee censorship. Nevertheless, it is clear that Baidu Baike provides a cultural-thickening pattern that is limited to mainland China and is more open to spam information.

It should be noted that both seem to fail to engage information from Singapore and Malaysia, despite the fact that Chinese Wikipedia nominally recognizes Singapore and Malaysia as separate geolinguistic regions on par with the others. Thus, the "trans-regional" Chineseness of Chinese Wikipedia is limited. While further research needs to be conducted on this topic, for the purpose of the thesis, it suffices to point out that Chinese Wikipedia does have more links than Baidu Baike to websites in Singapore and Malaysia. Since Malaysia and Singapore also use simplified Chinese, as does mainland China, what contributes to the divergence between the two encyclopaedias is not simply Baidu Baike's exclusion of traditional Chinese content but mainly its undue concentration of online sources inside mainland China. In conclusion, Chinese Wikipedia is not only "Chinese" enough, in the sense that it includes major websites in mainland China, but more "Chinese" than Baidu Baike by having better "information engagement" (Bennett & Wells, 2009) with online sources in Hong Kong, Macau, and Taiwan.

The unit of analysis of the findings so far has proceeded from the level of encyclopaedia platforms to the level of regions. Thus, the analysis of cultural-thickening patterns may not apply to the level of individual entry articles. For instance, it would be more convincing if the cultural-thickening phenomenon were contrasted at the level of articles by examining whether different external links and/or sources are integrated. While the findings so far have the merit of being comprehensive, further citation and content analysis of specific articles will provide an additional check on the findings, which leads to the next subsection.

5.3 Defining and negotiating Chineseness

Though both are “Chinese” enough, Chinese Wikipedia and Baidu Baike each have a different focus and meaning, and thus the way they define and present Chineseness may be different. Thus content analysis of articles that are essential in defining Chineseness can be conducted to complement the findings so far. Indeed, this content analysis constitutes a test of editorial objectivity for any general-purpose encyclopaedia insofar as it defines the language, ethnicity, and worldview of its major target audience. For example, it may be more difficult for Encyclopædia Britannica (Latin for “British Encyclopaedia”) to maintain absolutely objective and neutral when defining the English language, the British people, and the idea of Commonwealth than to describe, for example, the German language, the German people, and “Weltanschauung” (German for worldview or philosophy of life). To compare how Baidu Baike and Chinese Wikipedia define “Chineseness” differently, an analysis of the articles on Chinese languages, people, and worldview will be useful. To do this, I have chosen to study three entry articles, the Han Chinese characters (hànzì 汉字 hereafter “Hanzi”), the Han Chinese ethnic group (hànzú 汉族 hereafter “Hanzu”), and the Han Chinese worldview (tiānxià 天下 hereafter “Tianxia”) because of their central roles in defining and presenting Chineseness. Although

by no means comprehensive, the selection contains several essential articles for anyone who is interested in examining how Chineseness is presented.

It might also be possible to compare entry articles that are even more culturally and politically controversial, such as “Mao Zedong” and “cultural revolution”, and my decision not to do so is based on the fact that these articles are often not stable or protected enough from editing for long periods of time, thereby making them poor choices for reliable analysis. I have collected the versions on August 20, 2011 and on the same date in 2010. Since these article pages in 2011 are found to be relatively stable (no major substantial changes for a year), the following paragraphs will detail the comparisons for the 2011 version. Although I have myself edited several Chinese Wikipedia articles, my edits were done at a time and on articles that do not affect the findings on the three selected articles.

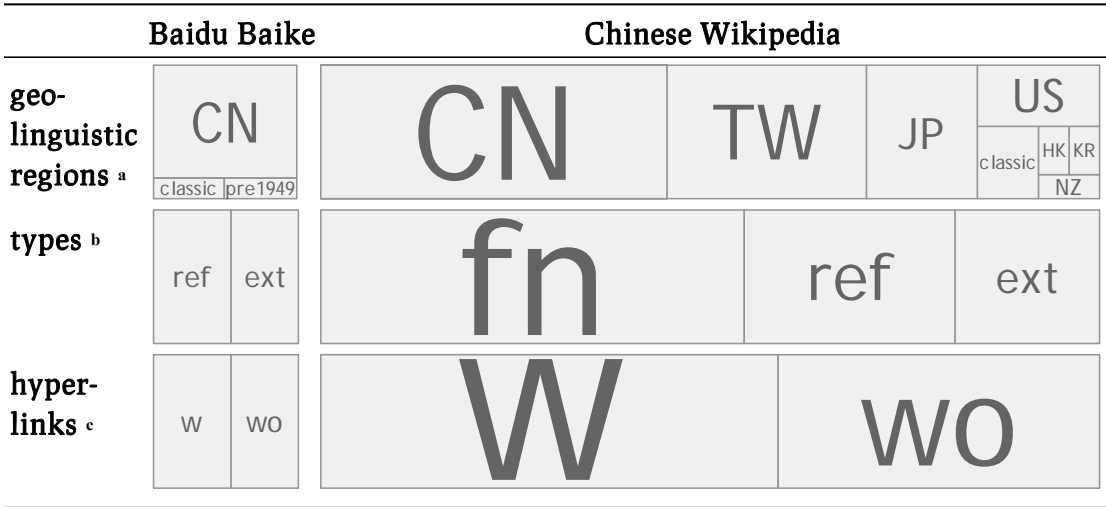
Still, my experiences in editing the more controversial article of “Zhonghuaminzhu” (Chinese nation) in Chinese Wikipedia can further illustrate the dynamics of keeping and removing content. All of this additional analysis will complement the citation and content findings in the first part of this chapter by testing the editorial judgement of concerned user-contributors.

5.3.1 Methods. I began the content analysis by examining the sources that are listed and cited as providing support for the content. After all, encyclopaedias are tertiary sources that depend mostly on secondary sources. I coded each of the sources using three different categorization schemes. First, based on how these sources are used in the articles, I coded them as three “types”: “footnotes with citations”, “references”, and “external links”. Second, I coded them based on their geolinguistic features. Third, I coded them based on whether a hyperlink is provided or not. The categorization outcomes have been visualized as treemaps for source comparison.

Further content comparison was conducted assisted by using difference-finding software tools (often called “diff” tools). Note that since copy-and-paste

activities are common on the Web, if two encyclopaedias share the same content, a further analysis of version history can provide some indication about which might be a copy of the other.

5.3.2 Findings: Citation and content analysis. Figure 5-3 shows the overall outcome of the three selected articles as a treemap. It shows the number of sources in proportion to the area size, and the treemap clearly indicates that Chinese Wikipedia has more than five times the sources that Baidu Baike has.



^a Country codes: CN (China), TW (Taiwan), HK (Hong Kong), JP (Japan), US (United States), KR (Korea), and NZ (New Zealand); also Chinese historical: classic (classic Chinese), and pre1949 (modern Chinese content published before 1949)

^b Types: fn (footnotes with citations), ref (references), and ext (external links),

^c Hyperlinks: w (with), and wo (without),

Figure 5-3. A treemap comparison of citations and external sources for “Hanzi”, “Hanzu” and “Tianxia”

Each row of Figure 5-3 shows the comparison based on different categorization schemes. First, in terms of geolinguistic regions, Chinese Wikipedia contains substantial sources from diverse regions and some classic Chinese sources. In contrast, Baidu Baike uses only one classic Chinese source and one pre-1949 Chinese source in addition to the majority of Mainland Chinese sources. This does not mean that Chinese Wikipedia is less Mainland Chinese

because, in absolute numbers, it has more mainland China sources than Baidu Baike has (see the area size of “CN”). Second, regarding the types of sources, Chinese Wikipedia has well-formatted footnotes and citations, whereas Baidu Baike lacks any footnotes. Third, in terms of the provision of hyperlinks, the two have a similar percentage of hyperlinked sources.

The findings above confirm what has been consistently established thus far: The content of Chinese Wikipedia is better supported with sources that are more diverse. A further breakdown for each of the three selected articles are provided as follows.

Tianxia. The historical editing evidence shows that Baidu Baike copied Chinese Wikipedia for the article on “Tianxia”. Based on the 2500-character-long Chinese Wikipedia version, Baidu Baike’s version removes approximately 750 characters and adds about 150 characters, keeping the 1600-character-long text exactly the same (about 60%). In terms of listed sources, Chinese Wikipedia’s eight references (all in Japanese) and six explanatory footnotes (without citations) are excluded from Baidu Baike’s version. Because the content is mostly the same, the differences in the listed sources are peculiar. This may be because the Japanese-language literature is not acceptable in Baidu Baike’s national framework or simply that copying citations takes extra effort in formatting. Regardless, knowing that the Chinese-language academic publication on “Tianxia” is enough for citations, I find it strange that this topic contains no citations in either Baidu Baike or Chinese Wikipedia. After examining the corresponding article’s edit history in the Japanese Wikipedia, I found that the Japanese version was created and expanded before the Chinese version by about two weeks, suggesting that the content is mostly translated originally from the Japanese Wikipedia for this topic. Chinese Wikipedia’s demand for high-quality, reliable sources to guarantee high-quality content is further exemplified by the discussion page on the article “Tianxia”. The article was designated as a “Good

Article” and passed by Chinese Wikipedians in 2006, but later, in 2012, the status of “Good Article” was revoked because of the lack of footnotes and references.

What Baidu Baike deletes and adds further shows a particular cultural-political stance in promoting an idealized historically consistent and unified China. For example, Baidu Baike’s version removes the historical detail that the Shang Dynasty did not have a concept of “Tianxia”, and adds a paragraph describing how “Tianxia” was unified under the Sui Dynasty. Probably in order to avoid phrases that sound too hegemonic or feudal, the internal links to other entry articles such as “China-centrism”, “Chinese emperor” and “Tian” (Heaven) were removed. In addition, the non-Han-Chinese “Tianxia” views, practiced by the Manchurian ruler of the Qing Dynasty, or by the ancient Korean ruler of Goguryeo, are all missing from Baidu Baike’s version. Chinese Wikipedia’s version, in contrast, retains the Japanese interpretation that the Manchurian rule of China transforms the once-Chinese “Tianxia” under barbarian rule, whereas Baidu Baike’s version presents the opposite: Chinese “Tianxia” persists despite the foreign rule by Manchus. This presents a case where Baidu Baike modifies the content that it copies from Chinese Wikipedia to fit the national framework of mainland China. What we can see here is the effect of knowledge-sourcing dynamics concerning the Han worldview (a contentious topic) across three encyclopaedia websites: Japanese Wikipedia, Chinese Wikipedia, and Baidu Baike.

Hanzi. Figure 5-4 shows the differences in the article on “Hanzi”. Again, Chinese Wikipedia contains more sources, and these sources are more diverse. Chinese Wikipedia’s article has 29 sources, whereas Baidu Baike has only six. With sources ranging from China to Taiwan to Korea, Chinese Wikipedia’s version draws upon wider sources, including six well-formatted footnotes with citations, ten references that include published academic sources, and thirteen external web links for “further readings”. In contrast, Baidu Baike’s article has only six items for “further readings”, all web pages hosted in mainland China.

	Baidu Baike	Chinese Wikipedia			
geo-linguistic regions ^a	CN	CN	TW	US	HK
					JP
					KR
types ^b	ext	ext	ref	fn	

^a Country codes: CN (China), TW (Taiwan), US (United States), HK (Hong Kong), JP (Japan), and KR (Korea)

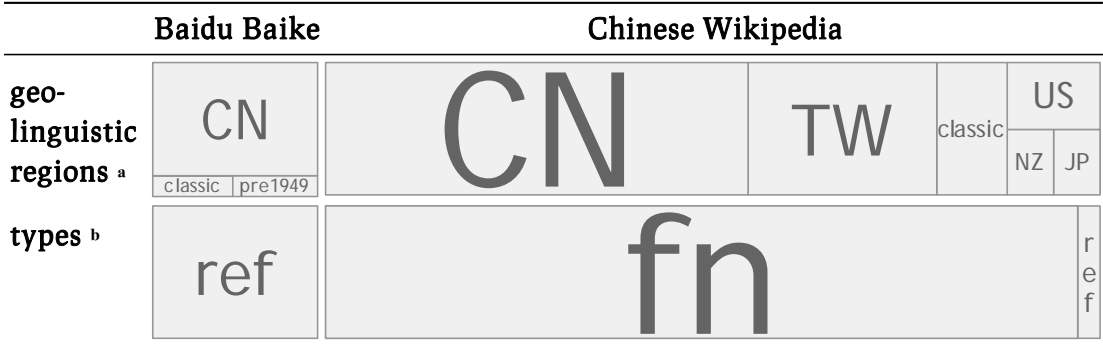
^b Types: ext (external links), ref (references), and fn (footnotes with citations),

Figure 5-4. A treemap comparison: “Hanzi”

Clearly, as regards Chinese characters, Baidu Baike reinforces a particular Mainland Chinese perspective whereas Chinese Wikipedia includes diverse perspectives. Further, the concept of the “Hanzi cultural sphere” (hànzì wénhuà quān 汉字文化圈) is not formally introduced in Baidu Baike’s version but receives a subsection of content (about a half-page long) in Chinese Wikipedia’s version, describing how Chinese characters have been historically used and appropriated by countries such as Japan, Korea, and Vietnam. Another feature is that the computing dimension of Chinese characters is addressed by a full section in Chinese Wikipedia, while it is briefly discussed in a subsection on the limitations of Chinese characters in Baidu Baike. Arguably, the most salient difference is the way the two versions address the differences between traditional Chinese and simplified Chinese characters. Baidu Baike never fully mentions the term “orthodox Chinese characters” (zhèngtǐ 正体) as proposed by Taiwan and cites only Mainland Chinese sources when describing the simplified versus traditional Chinese controversies. Chinese Wikipedia’s version, in contrast, presents more diverse (and thus arguably more balanced) views based on sources that are also more diverse. This cultural-political contrast is expected

because, by design, Baidu Baike as a linguistic platform already excludes contributions made in traditional Chinese characters (see Chapter 4).

Hanzu. Figure 5-5 shows the source differences of the article on “Hanzu”. Again, Chinese Wikipedia contains more sources in numbers and kinds: Chinese Wikipedia has 33 sources and Baidu Baike has only 7. Drawing on sources from China, Taiwan, classic Chinese texts, Japan, the U.S., and New Zealand, Chinese Wikipedia’s version addresses the subject matter of ethnic Han-Chinese with wider sourcing than Baidu Baike, whose seven listed sources include no specific footnotes but only seven references. Among the seven references, one provides a dead link to a page hosted by the Chinese Communist Party’s Youth Group’s old website; one provides a citation to a major book of history published in China before 1949; one provides another citation to an ancient history book written in classic Chinese; and one provides a citation to China’s 2005 population sample survey. The remaining three references are all China’s general history books published in mainland China by publishers closely affiliated with party and state organizations. In contrast, Chinese Wikipedia’s version lists 1 reference and 32 footnotes with citations, with materials ranging from Taiwan’s annual official survey of the overseas Chinese population worldwide - to an academic article that deconstructs the modern republican attempts to use the Yellow Emperor as the common ancestor of all Han-Chinese people. The sources in Chinese Wikipedia’s version are thus more diverse and balanced. Note that although having more diverse sources beyond just mainland Chinese ones, Chinese Wikipedia’s version also has more sources from mainland China than Baidu Baike’s version has.



^a Country codes: CN (China), TW (Taiwan), US (United States), NZ (New Zealand), and JP (Japan); Chinese historical: classic (classic Chinese), pre1949 (modern Chinese content published before 1949)

^b Types: fn (footnotes with citations), and ref (references)

Figure 5-5. A treemap comparison: “Hanzu”

The overall content difference shows (as with the other two articles) that, on the topic of Han Chinese people, Baidu Baike reinforces a particular perspective, whereas Chinese Wikipedia includes various perspectives. For example, when addressing the topic of linguistic diversity among the Han people, Baidu Baike’s version argues that, “the high level of unification in Chinese characters has strong cohesion effects within Han Chinese. Without any exaggeration we can say: without Hanzi, there are no Han Chinese people”. This statement is, at best, unencyclopaedic in tone and, at worst, misleading, since historically, Chinese characters are also used by literati-scholars in Korea, Japan, and Vietnam (Woodside, 2006), and they could function, “like medieval European Latin, as a common means between intellectuals of the region . . . well into the twentieth century” (Delanty, 2006, p. 33). In contrast, the political dimension of Han Chinese people is addressed in Chinese Wikipedia’s version; not only is the concept of the “Hanzi cultural sphere” addressed on various occasions but also the cultural-political tensions between “Hua” (Han Chinese culture or people) and “Yi” (non-Han Chinese culture or people) are discussed. As another example, on the issue regarding how the historical concept of Han Chinese people shifts to the modern concept of the Chinese nation, or

“Zhonghua minzu” (zhōnghuá mínzú 中华民族), Baidu Baike’s version presents a viewpoint that “Zhonghua minzu” will continue unification and merging with other ethnic groups as a continuity with what the Han Chinese people have historically achieved. In contrast, Chinese Wikipedia’s version provides a brief description on how and when “Zhonghua minzu” differs from “Hanzu”.

Altogether, the content and citation comparison of Baidu Baike and Chinese Wikipedia on three, key Chineseness-defining articles presents a clear contrast. The findings show that Baidu Baike avoids such tensions by removing or excluding them to fit a national framework of mainland China; sources and content that describe or explain such tensions, particularly published or sourced in Hong Kong or Taiwan, are excluded. In contrast, Chinese Wikipedia presents such tensions with more, and more-diverse, well-formatted sources and citations. As source citations indicate editorial judgement of knowledge authority, the contrast that has been found can thus be understood in terms of different cultural-political expressions of online collaborative projects. The two encyclopaedias, at least for the three Chineseness-defining articles, exercise different editorial judgements regarding what kind of knowledge sources can be considered as reliable and relevant. While the three concepts chosen are not representative, they relate to the contentious issue of how Chineseness and Chinese worldviews are related to Chinese languages, an obvious topic of concern in terms of the geo-linguistic approach used in this thesis.

Another example that demonstrates the contentious editorial dynamics related to Chineseness is my own editing experience after mid-2012 on the article about “Zhonghua minzu” in Chinese Wikipedia and some interesting outcomes in Baidu Baike, to which we can now turn.

Modern Chineseness: Zhonghua minzu. Because modern China was born out of a Chinese empire (Qing dynasty) ruled by a non-Han Chinese ethnic group of Manchus, modern Chinese national identity surrounding the term “Zhonghua

minzu” (中华民族 Chinese nation or Chinese race) has been a controversial topic (Fitzgerald, 1995; Gries, 2004; Unger & Barme, 1996; G. Wang, 1996a; Suisheng Zhao, 2004). In the early republican era, the republican view of the Chinese nation shifted from a Han-Chinese nation to a multi-ethnic nation that comprises five main ethnic groups (with Han-Chinese as the majority group). The term “Zhonghua minzu”, once a term proposed by constitutionalists in the late Qing period to resolve the cultural-political tensions between Manchus and Han Chinese, has gained wider acceptance and political legitimacy since then. Under the Communist state, the number of recognized ethnic groups was later increased from five to 56.

To test whether Chinese Wikipedia is indeed more inclusive, I undertook a kind of action research by editing the article on “Zhonghua minzu” myself. First, I identified several major scholars who have made authoritative and definitive contributions to the subject matter of “Zhonghua minzu”, including Xiaotong Fei from mainland China and Cho-yun Hsu from Taiwan (Fei, 1991; Xu, 2009). My plan was to see what could be kept and what might be deleted in two encyclopaedias by making several new edits based on these sources. I began the majority of my edits starting in mid-2012 and encountered several edits reverting my edits by various Chinese Wikipedians, including one who declared himself a Han-Chinese supremacist on his user page. By providing sources and explaining why my edits were necessary to improve the article, all recorded and documented in the edit history and talk page (zhWP, 2013a), most of my edits were slowly accepted and came through from October 26 to November 23, 2012 (zhWP, 2013b). At that time, I decided to wait a few weeks in order to make sure that my version was stable enough for other Chinese Wikipedia contributors before I took the next step, which was to contribute the new edits to Baidu Baike to see what would happen.

The situation changed rapidly when User:Shizhao, arguably the most famous Chinese Wikipedia contributor and administrator, decided to revert my

and others' edits on February 11, 2013, all the way back to an older version dated December 13, 2010, effectively removing 285 edits made by 74 different users (User:Shizhao, 2013). Based in Beijing, User:Shizhao has been known for his long-time maintenance service to Chinese Wikipedia, his escape from a series of recall elections against his administrative power, and as an administrator who upsets some users, especially those from Hong Kong, concerning specific articles (Mingpao, 2010a, 2010b, 2010c). Despite my efforts to explain the new edits, he seemed unconvinced, reverting to the old 2010 version with no further explanation except for a short statement that he thinks the old version is better (zhWP, 2013b). After a few back-and-forth edits, mostly between the two of us, another administrator (from mainland China) intervened to "protect" the older version. In Wikipedia projects, when an article is protected, no new edits are possible without administrative power. Though the "protect" mechanism aims to cool down editing controversies by stopping further editing activities and thus does not endorse the "protected" version, there is evidence that some Wikipedia editors exploit the mechanism to protect their preferred versions, so that other editors can make no further modifications.

With all the new edits after December 2010 that were effectively removed by Chinese Wikipedia, I decided to proceed to make similar contributions to Baidu Baike instead. To my surprise, without my action, Baidu Baike's article had already included some of my edits from the Chinese Wikipedia version (User:Wǒnǎiyěyún hè(我乃野云鹤), 2012) dated November 13, 2012. In the introduction section, one sentence (but not the references) describing the conceptual contribution by academics such as Xiaotong Fei and Cho-yun Hsu had been copied character by character from my edits in Chinese Wikipedia and had survived the editing of at least 11 contributors for four months ("Zhonghua minzu (zhōnghuá mínzú 中华民族)," 2013). Of course, as we have already seen, a lot of the content of Baidu Baike is simply copied from Chinese Wikipedia.

My editing experiences with “Zhonghua minzu” seemingly contradict the rest of the findings in this chapter. Removed from Chinese Wikipedia, my edits mysteriously appeared in Baidu Baike. The contradiction mainly indicates the dynamics of constant user contributions. In relation to a particularly controversial article, local dynamics may develop in a way that goes against the general editorial patterns. Researchers must recognize that a small number of user-contributors can potentially shift the dynamics of a certain encyclopaedia article, and thus they should avoid the over-generalization of findings based on a single article. Researchers may need to consider Yang’s (2003) guerrilla ethnography approach and embed themselves more into the dynamics across more online spaces. Alternatively, researchers can apply webometrics analysis that offers a more comprehensive picture that complements findings based on a single or a few articles.

Nevertheless, my editing experiences with the “Zhonghua minzu” article provided some important observations that are consistent with the findings thus far. First, Baidu Baike does copy from Chinese Wikipedia. Second, during the copying process, footnotes and citations appear to be dropped, probably because more efforts are required to keep them. Third, through removing and adding content copied from Chinese Wikipedia, Baidu Baike can better fit the new edits under the national framework of China. Fourth, cultural-political tensions are not automatically resolved in Chinese Wikipedia simply because it is more integrative than Baidu Baike is. For example, my editing experience shows that the editorial opinion of one Mainland Chinese administrator may prevail for a period without substantial challenges. Thus, the apparent contradiction does not challenge the findings so far but rather highlights the changing dynamics *within* and *between* the two major Chinese-language user-encyclopaedias.

5.4 Chapter conclusions

Citing and linking information sources requires active engagement with information sources on the part of user-contributors. As tertiary sources,

encyclopaedias must rely on secondary sources for reliable information. Thus, citing and linking patterns manifest themselves as sharing activities that gather sources from different regions. Boundaries are reintroduced when cross-regional linking and citing is missing or even discouraged. By not linking or citing enough sources from Macau, Hong Kong, and Taiwan, Baidu Baike effectively reinforces the boundary between mainland China and the other Chinese-speaking regions. In contrast, Chinese Wikipedia includes more geolinguistically diverse sources, resulting in a cross-regional repertoire for Chinese-speaking users. Because of Baidu Baike's failure to include important sources from these regions, the difference between the two, i.e. Mainland China versus the Chinese-speaking world combined, further confirms the similar observations of Chapter 4 that Baidu Baike has failed to overcome existing boundaries while Chinese Wikipedia has done so.

To conclude, we can first explicate these findings using the notion of cultural thickening. Two distinct cultural-thickening patterns show how the two websites engage the world's knowledge sources differently. As summarized in Table 5-13, the cultural patterns indicated by linking and citing sources suggest that Baidu Baike mostly thickens information sources within mainland China, whereas Chinese Wikipedia integrates sources across diverse regions. In addition, using the notion of "civic learning repertoires" for "information engagement" (Bennett & Wells, 2009), encyclopaedia content-sourcing can be seen as (1) building shared knowledge resources and (2) negotiating judgements about information sources. As a result, different cultural-thickening processes correspond to different engagement patterns. Thus, it is reasonable to suggest that Baidu Baike exemplified a limited version of civic engagement among Mainland Chinese users, whereas Chinese Wikipedia embodied civic learning repertoires and information engagement among users across Chinese-speaking regions. Furthermore, in dealing with spam and geolinguistic factors, the comparative analysis of content and citations in this chapter also confirm the

findings in Chapter 4. Altogether, the patterns of practice and discourse show two general patterns: One puts (mainland) China at the information centre of the world, whereas the other draws upon additional other Chinese information outside mainland China. One allows spam links while the other filters spam links.

Table 5-13

Comparing distribution of citing and content patterns

Patterns	Baidu Baike	Chinese Wikipedia
... of linking	Mostly sources from mainland Chinese in simplified Chinese	Diverse sources from around the world
... of citing	Mostly sources from mainland Chinese in simplified Chinese	Diverse sources from across Chinese-speaking regions
... of geo-linguistic features	Overwhelmingly mainland and simplified Chinese	Mixed
... of civic engagement	Very likely limited to mainland Chinese users	Very likely across Chinese-speaking regions
... of discourse	(Mainland) China as the centre of the world	(Mainland) China as part of the (Chinese-speaking) world
... of practice	Free to spam	Free from spam

Pre-existing cultural-political boundaries have been reintroduced or overcome accordingly: Chinese Wikipedia's geolinguistic extent of information engagement (mostly the Chinese-speaking world combined) is thus shown to be more extensive than that of Baidu Baike (relatively "selective" for Mainland China). Baidu Baike has failed to bring together content and sources from regions outside mainland China, including regions claimed by Beijing as inseparable parts of China. According to Table 5-7, in fact, Chinese Wikipedia still provides due and adequate citations to Mainland Chinese sources, despite having been blocked by Beijing for a long period of time (see also Chapter 4), reflecting the fact that it has a sizable proportion of contributors from mainland

China. The quantitative findings thus offer insights on the intensity level of cross-regional interactions and boundary dynamics. The lack (disconnections) or small amount of information engagement (relatively insignificant connections) signals information barriers or voids, suggesting the existence of boundaries being reinforced.

The geolinguistic analysis, both generally and specifically related to Chineseness, shows that the citations and thus the content in traditional Chinese from Hong Kong, Macau, and Taiwan distinguish Chinese Wikipedia from Baidu Baike. The difference provides important insights into the civic engagement with the two encyclopaedias. The geolinguistic analysis has proved useful in showing the reach or focus of civic engagement online.

The issues of geolinguistic affinity and barriers are often overlooked in the literature about civic engagement, leaving questions about Internet users of different world languages. How, for example, do geographic and linguistic factors shape the Arabic-language Internet? What is the role of the U.S. in hosting Arabic-written sources? Are there major emerging hosting websites outside the U.S. such as Yandex and Baidu? Special attention is needed to explore these linguistic and geographic aspects of online practices, especially with non-Latin writing systems such as Cyrillic, Arabic, and Chinese that are major world languages (Petzold et al., 2012). Thus, the study presented here can be considered as a case that has wider import for Internet that is increasingly more multilingual, rather than just pertaining to the special “Chinese case”. The reach or focus of civic engagement online can be better understood with more research using geographic and linguistic information online, loosely called “geolinguistic research” here.

It should be noted, however, that this study is just one example of research that examines the geographic and/or linguistic features and boundaries or gaps in public knowledge production (as discussed in Chapter 2; e.g. Thelwall & Smith, 2002; Liao, 2008; Hecht & Gergle, 2009, 2010; Graham et al., 2011; Liao

& Petzold, 2010; Lowe, 2011; Paolo Massa, 2011; Bao et al., 2012; P. Massa & Scrinzi, 2012; Petzold et al., 2012; Warncke-Wang et al., 2012). Particularly for research on online civic engagement, different geographic and linguistic factors are often historically and politically linked to certain specific cultural-political situations, as showcased in this chapter. Thus, a geolinguistic research agenda is useful in bridging at least two research communities that consider geographic and linguistic factors: On one hand, the area studies community has a deep interdisciplinary understanding of the factors of languages and geographical regions. On the other, the Internet research community has begun to develop a number of ways of processing and theorizing geographic and/or linguistic factors online. This chapter has merely presented one example (but arguably among the first for Chinese civic engagement online). It can be expected that useful research will be produced if the geolinguistic approach attracts more effort from both communities.

In the Chinese context, the reach and focus of civic engagement has important qualitative dimensions because of the different political and media regulation systems across Chinese-speaking regions. Online civic engagement, therefore, entails the negotiation of civic cultures by users with diverse civic experiences. I contend that the difference between Baidu Baike and Chinese Wikipedia reflects a more general condition of the Chinese-language Internet in the cultural patterns of citing, linking, filtering, and thinking about information sources. In this sense that this chapter has advanced upon Yang's (2003) notion of Chinese cultural sphere by highlighting the cultural-political tensions of online civic engagement that may be expressed in geolinguistic differences. Baidu Baike's reach and focus on mainland China has a qualitative dimension in avoiding engagement or contacts with civic cultures outside mainland China. Chinese Wikipedia's integrative ideal, in contrast, has to confront the cultural-political tensions that are bound to occur when civic cultures from different Chinese-speaking regions meet.

Some obvious limitations exist for research about understanding Chinese civic engagement online. While this chapter has provided one approach, the geolinguistic findings in this chapter alone cannot explain why such different outcomes are produced, which is better explained by the findings in Chapter 4. Furthermore, direct evidence about the geographic and linguistic profiles of contributors could help to confirm or reject the findings presented here (but there are obvious constraints to obtaining this information reliably and comprehensively). Despite these limitations, it is important that researchers should take geographic and linguistic factors into account when considering civic engagement and civic learning practices. For Chinese civic engagement online, the limitations of the study indicate a need to examine how different Chinese “civics” can be identified or even predicted by considering geographic and linguistic factors and how users from different Chinese-speaking regions may engage one another differently because of such differences. Beyond the Chinese context, a systematic review of geolinguistic methods and approaches would be useful in the future.

Analysis of geographic and linguistic information expands not only our view of what online engagement looks like, but also of who engages with what. The study of two major Chinese-language websites points to the relevance of analysing online civic engagement beyond the usual universal or national frameworks that are often uncritically assumed. Geographic and linguistic information can help researchers to be specific about the actual dynamics inside the so-called Twitter universe, Arabic online world, or Chinese cyberspace, opening up spaces for richer contextualization and interpretations. As more tools and datasets become available, research can examine various cosmopolitan, national, and local developments based on the systematic collection and analysis of geographic and linguistic information. Although this study presented only one specific geolinguistic approach for two websites, the findings nonetheless demonstrate how this can be done to enrich our understanding of Chinese civic

engagement online. Research such as this can be expected to have implications for the larger and on-going discussions about the Internet and cosmopolitanism (e.g. Ess, 1998, 2002; Jeffres et al., 2004; Zuckerman, 2013). Because geographic and linguistic factors remain too crucial to be overlooked, more systematic geolinguistic methods and research are needed.

The findings extend theories about the relationship between Internet connectivity and geolinguistic features by showing where and why Internet connectivity may result in different kinds of cultural thickening of online content and practices. This also includes using webometric data to identify “cultural thickening”, as described in the previous two chapters. Effectively, the patterns of connections and disconnections help to show the changing boundaries between “them” and “us” or of “technological communities” (Meyrowitz, 1997). The findings in this chapter show the connections (or lack thereof) of two online encyclopaedias to the world and Chinese-speaking regions. In relation to Chapter 4, the findings add empirical evidence based on content and citation to indicate different patterns of cultural thickening (connections) or lack thereof (disconnections). Such distinct patterns of cultural thickening can be expected to influence the way users may use (or not) and perceive the two encyclopaedias, a question that remains to be explored in the next chapter. So far, the evidence about a bounded sense of “them” versus “us” does not show two well-defined and clearly bounded communities. The citation and content findings do, however, show that the Chinese-speaking regions of Taiwan, Hong Kong, and Macau are excluded from the constituents of “us” in Baidu Baike, whereas these regions are included in Chinese Wikipedia. Differential citing, linking, and copying content lead to distinct patterns of reshaping boundaries. The next chapter will explore whether similar patterns can be identified from other major Web platforms such as search engines and social media.

It is worth repeating that both encyclopaedias, as part of an emerging phenomenon of online collaborative projects in the first decade of the 21st

century for Chinese-speaking Internet users, are major testing grounds for analysing change in overcoming or reinforcing existing communicative boundaries. Thus, the collection of external links not only represent the outcome of collective editorial exercises by Chinese-language user-contributors of the two encyclopaedias, but also position the two encyclopaedias differently in the overall structure of the Chinese-language Internet. These web links are at the intersection of content, users, and interactions between them, which could further engender the links and commentaries that occur elsewhere. In the next chapter, I will further compare how the two encyclopaedias are used, mentioned and perceived in other web platforms such as search engines and social media.

Chapter 6 Reception and use

This chapter is about the reception and use of the two encyclopaedias. Before going into the analysis, a brief personal anecdote can help to underscore how looking for encyclopaedic knowledge has changed dramatically. This is not intended to be taken as a representative case of users of Chinese-language encyclopaedias, but my own experience nevertheless exemplifies the contrast of encyclopaedia use before and after the widespread adoption of the Internet. During the early years of my elementary school, I was one of the participants in an encyclopaedia-looking-up competition where selected students of my age were given the same encyclopaedia to answer as many questions as possible. My winning strategy was to familiarize myself with the index system of that encyclopaedia in a public library. To do so, I needed to minimize the time spent in the following activities: decomposing Chinese characters (first by radicals and then by the number of strokes), locating the right knowledge topic in the index, and flipping pages to the entry article. Things have changed three decades later when search engines and online encyclopaedias have become popular; the literacy skills to find information has changed dramatically towards digital means. One only needs to type in the relevant terms to find the right web page that contains the information one is looking for. Since most of the weight-lifting tasks such as index-browsing and page-flipping are now handled by websites such as search engines, it is then important to examine how these machines may lead users to which websites.

In imagining how users across different Chinese-speaking regions use online encyclopaedias, we face a major research challenge, as described by Richard Rogers(2013): to conduct cultural analysis about the dominant websites which a group of users mainly use, especially the leading websites in a particular country and/or language. While users do have alternative choices on the Web, their use and reception of the Web is likely to be shaped by certain leading websites that dominate the local market, which is often delineated by certain language preference(s). Indeed, the standard by which I input Chinese characters is different

from most of the simplified Chinese users, and this may in turn result in different information outputs generated by various websites. Users across different regions may rely on different search engine systems that sort articles and topics. Thus, as the use of encyclopaedias shifts from offline to online, the information-seeking activities by users across regions are mediated by their respective (likely local versions of) websites, or by the locally dominant websites. Hence, it is necessary to step outside both encyclopaedia websites and evaluate how they are received and used, especially through the mediation of what are likely to be different major websites across regions. This chapter will examine two such types of websites to observe the reception and use of the two user-generated encyclopaedias: search engines and microblogs. The findings should add to what we have seen so far regarding content and information engagement.

To answer the questions of reception on other sites, I made several field trips in mainland China, Hong Kong and Taiwan⁷, asking questions about peoples' opinions about and experiences with online encyclopaedias. From these informal interviews, I noticed that few users in Hong Kong and Taiwan use Baidu Baike, mostly because the choice of language script and their choice of search engines. The experience of mainland Chinese users on the other hand is quite different. Some outright reject Baidu and Baidu Baike as a close alternative to Google and Wikipedia, while others believe that for issues internal to China, one should consult Baidu and Baidu Baike, but for issues external to China, Google and Wikipedia. This anecdotal evidence is consistent with some website query and traffic reports. For instance, in timelines comparing the popularity of different search terms (i.e. number of search queries), Google Trend shows that in Hong

⁷ For mainland China, I visited Xiamen, Quanzhou, Fuzhou, Changting, Nanchang from 19 to 30 August 2009. I visited Wuhan in August 2014. For Taiwan, I visited the Academia Sinica as a doctoral fellow from August 2009 to June 2010. For Hong Kong, I visited the City University of Hong Kong as a visiting doctoral student around the early half of the year 2013.

Kong and Taiwan, far more queries are submitted for finding Wikipedia than Baidu Baike.⁸ In mainland China, the search requests submitted to Google for the two were at the same level from 2009 to 2013, with Wikipedia having a slight overall advantage.⁹ Baidu Index (similar to Google Trend) also has the term Chinese Wikipedia on top from 2011 to 2014¹⁰. According to the 1:1000 sampled traffic report provided by the Wikimedia Foundation, for Chinese Wikipedia, around one-third of the visitors come from Taiwan and mainland China, around 15% from Hong Kong, and the rest mainly come from North America, Malaysia, and other regions where there are Chinese-speaking users. However, Baidu has not released any similar traffic reports about its Baike subdomain (i.e. baike.baidu.com), and thus it remains a research challenge to compare the two encyclopaedias on an equivalent basis.

Again, cross-boundary collaborative and networking possibilities seem to be bounded by various geographic, linguistic and cultural-political factors, suggesting a re-introduction of existing boundaries back onto the Web. Because the use and reception of an encyclopaedia appears to be linked to that of search engines, search engines are bound to shape which encyclopaedia is likely to be shown in the search results and hence visited. Thus, the first part of this chapter will discuss and

⁸ More search queries for Wikipedia than for Baidu Baike in Taiwan and Hong Kong. See Google Trend results in Taiwan: <http://www.google.com/trends/explore#q=%E7%B6%AD%E5%9F%BA%E7%99%BE%E7%A7%91%2C%20%E7%99%BE%E5%BA%A6%E7%99%BE%E7%A7%91&geo=TW> and in Hong Kong: <http://www.google.com/trends/explore#q=%E7%B6%AD%E5%9F%BA%E7%99%BE%E7%A7%91%2C%20%E7%99%BE%E5%BA%A6%E7%99%BE%E7%A7%91&geo=HK>

⁹ Search queries for Wikipedia is on par with those for Baidu Baike for Google China. See Google Trend results in mainland China for the search term Wikipedia and Baidu Baike: <http://www.google.com/trends/explore#q=%E7%BB%B4%E5%9F%BA%E7%99%BE%E7%A7%91%2C%20%E7%99%BE%E5%BA%A6%E7%99%BE%E7%A7%91&geo=CN>

¹⁰ Search queries for Wikipedia are more than those for Baidu Baike for Baidu Search. See Baidu Index results: <http://index.baidu.com/?tpl=trend&word=%CE%AC%BB%F9%B0%D9%BF%C6%2C+%B0%D9%B6%C8%B0%D9%BF%C6>

analyse a systematic method to “read” the search engine result pages (hereafter SERP) to see which encyclopaedia is visible or not, using an industry model of click-through-rates as the approximation of likely traffic directed by different search engines. The second part of the chapter will then focus on the opinions and experiences shared on the Sina Weibo and Twitter platforms. By covering several major online platforms that have garnered substantial visits by users across Chinese-speaking regions, this Chapter will show how the use and reception of the two encyclopaedias differs.

6.1 Search engine result pages (SERPs)

Using search engines is among the most popular online activity for users in the US (Fallows, 2008) and mainland China (CIC, 2009; CNNIC, 2009), and has been among the driving forces of fast-growing online advertising platforms (Varian, 2007; SEMPO, 2011; IDATE, 2011; PricewaterhouseCoopers, 2011). It has been reported that (and there is speculation why) the global search engine leader Google has consistently favoured the global leader of user-generated encyclopaedias Wikipedia by showing the relevant pages frequently and prominently in the SERP (Charlton, 2012; Čuhalev, 2006; Gray, 2007; Jones, 2007; Silverwood-Cope, 2012). Independent market research by Nielsen Online and Hitwise Intelligence has demonstrated that Wikipedia not only dominates online visits for encyclopaedia content, but also does so mainly because of the traffic directed by major Web search engines (Hopkins, 2009; Nielsen Online, 2008). Even the Wikimedia Foundation has acknowledged this (Google drives traffic to Wikipedia), but nonetheless argued that half of its readers want to look for Wikipedia content (Khanna, 2011). Thus, as major websites that dominate traffic and user attention, Google and Wikipedia seem to be central in guiding users where to look.

However, most research about search engines is limited to or predominantly focused on the English-language context (Battelle, 2005; Bermejo, 2009; Couvering, 2004, 2008; Dahlberg, 2005; Hargittai, 2007; Segev, 2008). Little effort has been made to understand whether phenomena that are specific to Google/Wikipedia

can be found for other major search engines and user-generated encyclopaedias. One study has been conducted analysing SERPs inside mainland China, with 316 Chinese-language search queries of politically controversial internet incidents in mainland China, indicating that indeed Baidu Baike and Chinese Wikipedia are ranked high among the SERPs (Jiang & Akhtar, 2011). However, and the findings are limited to simplified Chinese users in mainland China and a selected sample of search queries. .

For the larger scope of Chinese-language internet, there are many localized versions provided by several major search engines, including examples such as Yahoo China, Google Hong Kong, Google Taiwan, etc. I call them local search engine variants (hereafter SEv). Do different SEvs guide users from various Chinese-speaking regions to see the same websites regardless of which search engine they chose? Do the SERP diverge? Comparing Baidu Baike and Chinese Wikipedia across various popular Chinese-language search engines will fill the gap of considering both non-Google search engines and non-English versions of user-generated encyclopaedias. This part of thesis puts forward a method to quantify the SERP rankings and presents the most comprehensive findings on the Chinese-language SERP about online encyclopaedias to date. The findings should answer whether and how Baidu Baike and Chinese Wikipedia are visible across different local SEv and for what types of search queries. The results are based on 3,000 mainly-Chinese-language search queries across four Chinese-speaking regions (mainland China, Singapore, Hong Kong and Taiwan) can demonstrate that major user-generated encyclopaedias are indeed among the most visible and that localization factors matter. Thus, encyclopaedia websites such as Baidu Baike and Chinese Wikipedia provide leading indicators about how the Chinese-language search environment is structured.

For SERP, there is still a question about information control and linguistic boundaries, with national “borders” having been reintroduced in many technological and legal arrangements (Goldsmith & Wu, 2008). In particular,

Google's first collaboration with (or accommodation of) the needs of the Chinese government and its later exit from mainland China have demonstrated the intricate political and cultural dimensions of "localization" of search engine services (Vaughan & Zhang, 2007; Einhorn, 2010). Thus, there is a research gap on the effects of localization on SERPs and non-English Wikipedia, including prominent cases of Chinese-language and Arabic-language internet users whose recent rise in the new internet world has attracted much attention (Dutta et al., 2011). In particular, to answer how search engines and/or user-generated encyclopaedias reintroduce or shape national or social boundaries, more work on localization effects is needed. Localization is also discussed as a contributing factor to "internationalization mechanisms" of online gatekeeping (Barzilai-Nahon, 2008), holding the key to understanding the nationalization or internationalization dynamics of the Web.

The high likelihood for users to visit encyclopaedia websites via search engines indicates that their respective web spheres are indeed porous to one another. Nonetheless, questions remain as to which encyclopaedia websites are more porous to which SEVs in the Chinese-language Internet. Thus, cross-regional SERP research will provide answers about how cultural thickening patterns are produced that may overcome, reinforce or ignore existing cultural and political boundaries.

6.1.1 Methods. Using SERP data as evidence of search engines' recommendations for users and thus as a window onto different groups of users, the research method includes a generic straightforward visibility test. The test produces different scores that represent the level of visibility given by different search engines to different websites based on their positions in the SERPs. When applied to various choices that are most popular across different Chinese-language search engine markets, the data collected can thus be clustered to identify the underlying structure of the relationship between SEVs and websites. The patterns identified can then be interpreted as different patterns of cultural thickening. I will

also explain the data collection strategy and how the search terms were chosen for the SERP of 3000 search queries.

A straightforward visibility test. Since people often browse the SERPs from the top to the bottom, various marketing studies (Enquiro, 2007) as well as social science research (Bar-Ilan, 2006; Dunleavy, Margetts, Bastow, Pearce, & Tinkler, 2007; Margetts & Escher, 2006; Vaughan & Thelwall, 2004) and industry practices (Slingshot SEO, 2011) have measured the level of online visibility based on webometric data such as the positions in SERPs. Generally, a website is more visible if it is presented nearer the top of the SERPs, as shown in measurements constructed for keyword search advertising (Brettel & Spilker-Attig, 2010; J. Chen, 2008; B. J. Jansen, Brown, & Resnick, 2007; B. J. Jansen & Mullen, 2008; J. Jansen, 2011; Jung, 2008; Malaga, 2008; Spindler, 2010). The goal of advertising is to boost the ranking of a website for a target set of search terms (or search keywords). The higher the ranking, the higher amount of traffic.

At least five industry sources¹¹ have released different click-through rates¹² (hereafter CTR) for the SERPs based on traffic data (Hearne, 2006; Jones, 2007; Young, 2011). Such data helps analysts to convert the ranking number into measurements of traffic. For example, analysts can capture and compare the difference between the 1st and 2nd and that between the 9th and 10th based on traffic data, instead of assuming their difference to be the same (i.e. $10-9=2-1$). In what follows, I provide a means of assessing which items of the SERP are clicked through the most. Based on the average value of the five industry sources¹³, I derive a mathematical function that translates ranking position into different scores that I call “visibility scores”. Figure 6-1 shows that the function of $y = 0.2889x^{-1.078}$ (see the

¹¹ They are Chiticka, Slingshot, Optify, Enquiro, and AOL.

¹² Used in online advertising beyond search engines, the CTR measures the number of clicks on a web link divided by the number of times it is shown to the users (i.e. clicks/impressions).

¹³ See footnote 11.

trend line in the figure) fits the average industry data points (see data marked “x” in the figure) well (see the high R^2 value of 0.9934). With this function, a SERP ranking x can be translated into corresponding visibility score y . Effectively, visibility scores provide a weighting mechanism. Figure 6-1 contrasts the weighted values (see data marked “x” in the figure) and the unweighted ones (see data marked “.” in the figure) that divide the visibility scores equally¹⁴. The contrast clearly shows the importance of the top three positions of the SERP which produce higher than average results. Hence, this method allows researchers to quantify the difference made by the SERP ranking.

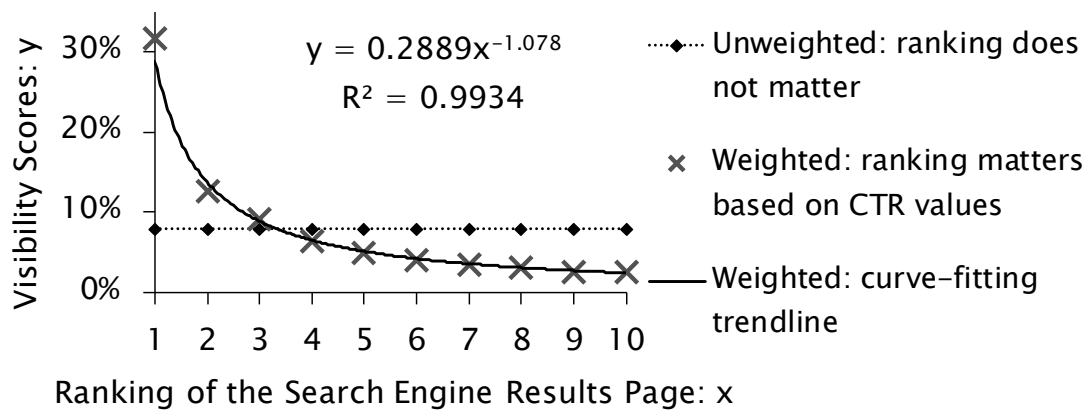


Figure 6-1. Higher SERP rankings x produce higher click-through rates and thus higher visibility scores

The visibility scores are by no means universal for all search engines or for any groups of users. Nevertheless, they serve as a useful instrument to analyse the SERP data in a way that is closer to known traffic trends as determined by the ranking position. Translating ranking positions into visibility scores, a visibility test

¹⁴ Since the first SERP usually contains the top-10 results, the proposed curve-fitting function suggests that about 79% of visibility scores are already assigned to them, thereby putting the unweighted value at 7.9% assuming the visibility scores of the first SERP are distributed equally among the top-10.

can aggregate results based on a selection of search queries submitted to different search engines. (One cannot add ranking numbers for comparison, but one can add visibility scores.) Thus, researchers can choose a relevant set of search queries and/or SEVs that cover the scope and focus of the research. The resulting SERPs dataset then represents the aggregated outcome, effectively, for different visibility scores contributed by different SEVs to different listed websites over certain keywords. Visibility scores allow for aggregation, comparison and further analysis once researchers have justified their selection of SEVs and keywords.

Chinese-language search engine markets. To select relevant the SEVs for this research, I first review the market shares across Chinese-speaking regions as follows.

I selected nine SEVs that cover over 97% of each regional market as of 2012. Table 6-1 lists the top-5 search engines for respective Chinese-speaking regions, based on StatCounter's data for March 2012. Google and Yahoo were among the top two for almost all regions except for mainland China where Baidu was dominant. The selection of Baidu, Google and Yahoo thus already covers over 97% of the respective market shares.

Table 6-1

Search engine market shares: top 5 according to StatCounter

Rank- ing	China (CN)		Singapore (SG)		Hong Kong (HK)		Taiwan (TW)	
1	Baidu	57.98%	Google	90.32%	Google	67.10%	Google	58.28%
2	Google	36.93%	Yahoo!	7.63%	Yahoo!	32.37%	Yahoo!	40.60%
3	Yahoo!	2.28%	Bing	1.78%	Baidu	0.29%	Bing	0.95%
4	Bing	2.16%	Ask Jeeves	0.11%	Bing	0.13%	Baidu	0.08%
5	Yandex Ru	0.17%	Baidu	0.10%	Ask Jeeves	0.06%	Ask Jeeves	0.05%

Because the market share data differs and varies according to time, researchers may need to consider time variations. For this research, Figure 6-2

shows the trend lines for each market of major Chinese-speaking regions, all of which indicate that the proposed selection of nine search engines remain valid from around 2009 to 2012. It can be noted that, after Google moved its mainland operations to Hong Kong (BBC, 2011), Baidu continued to enjoy its lead in mainland China with Google in second place. In Hong Kong and Taiwan around 2010 to 2011, Google has overtaken Yahoo's leading position while maintaining its top position in Singapore (StatCounter, 2011).

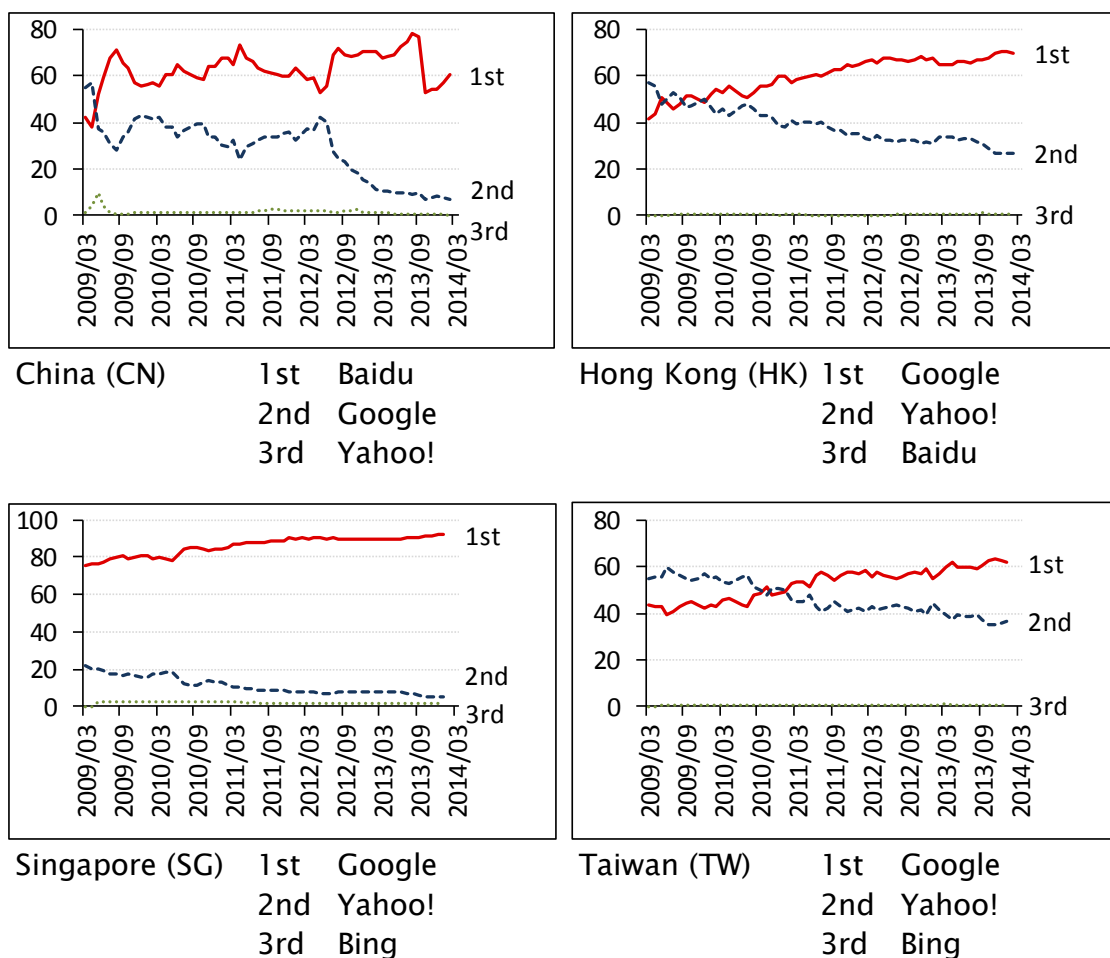


Figure 6-2. Search engine market shares: from 2009 to 2014

Nine SEVs can accordingly be selected for this research, which are hereafter abbreviated as Baidu_CN, Google_CN, Yahoo_CN, Google_SG, Yahoo_SG, Google_HK, Yahoo_HK, Google_TW, and Yahoo_TW. The basis for the selection of the three major search engine providers across four regions (mainland China,

Singapore, Hong Kong, and Taiwan) is also consistent with various other survey, market and traffic reports from both inside and outside mainland China (CIC, 2009; CNNIC, 2006, 2007; Nguyen, 2011; Russell, 2011).

For each variant, Baidu Baike and Chinese Wikipedia, along with other websites, can be expected to show up in the respective SERPs and receive different visibility scores. Based on the data, analysts can then not only answer questions about which SEv ranks which encyclopaedia higher, but also observe patterns of relationships between variants and websites (not just encyclopaedia websites). In other words, by observing specific variants, the research design allows researchers to examine rather than assume which SEvs display which websites. No prior assumptions are made about how factors of platform, geography or language will matter more. This also allow analysts to identify the structure underlying such relationships if certain data clustering is conducted.

Data clustering strategy. To identify the structure underlying the “which SEvs display/cite which websites” relationship, I use blockmodelling analysis (Doreian, Batagelj, & Ferligoj, 2004) of two-mode networks. Indeed, the relationship observed directly from the SERPs provided by different SEvs constitutes a one-way directional two-mode network where a SEv on one side effectively “cites” a websites on the other. The visibility network is one-way directional because its only relationship involves citing or showing, which is one path from SEvs to websites. The network is two-mode because it has only two separate sets of nodes: SEvs and websites.

Figure 6-3 presents a hypothetical example of how blockmodelling can be used to reveal the underlying structure of a network relationship. The same two-mode networks is presented in the form of a matrix (left) and a graph (right) respectively, showing how the nine nodes of the alphabet (K, L, M, etc.) link to another set of nine nodes of numbers (①, ②, ③, etc.). A black-coloured cell in the matrix and a directed arc in the graph represent the existence of a relationship. For

this research, the nine nodes of the alphabet can represent different SEVs and the numbers can represent a selection of visible websites.

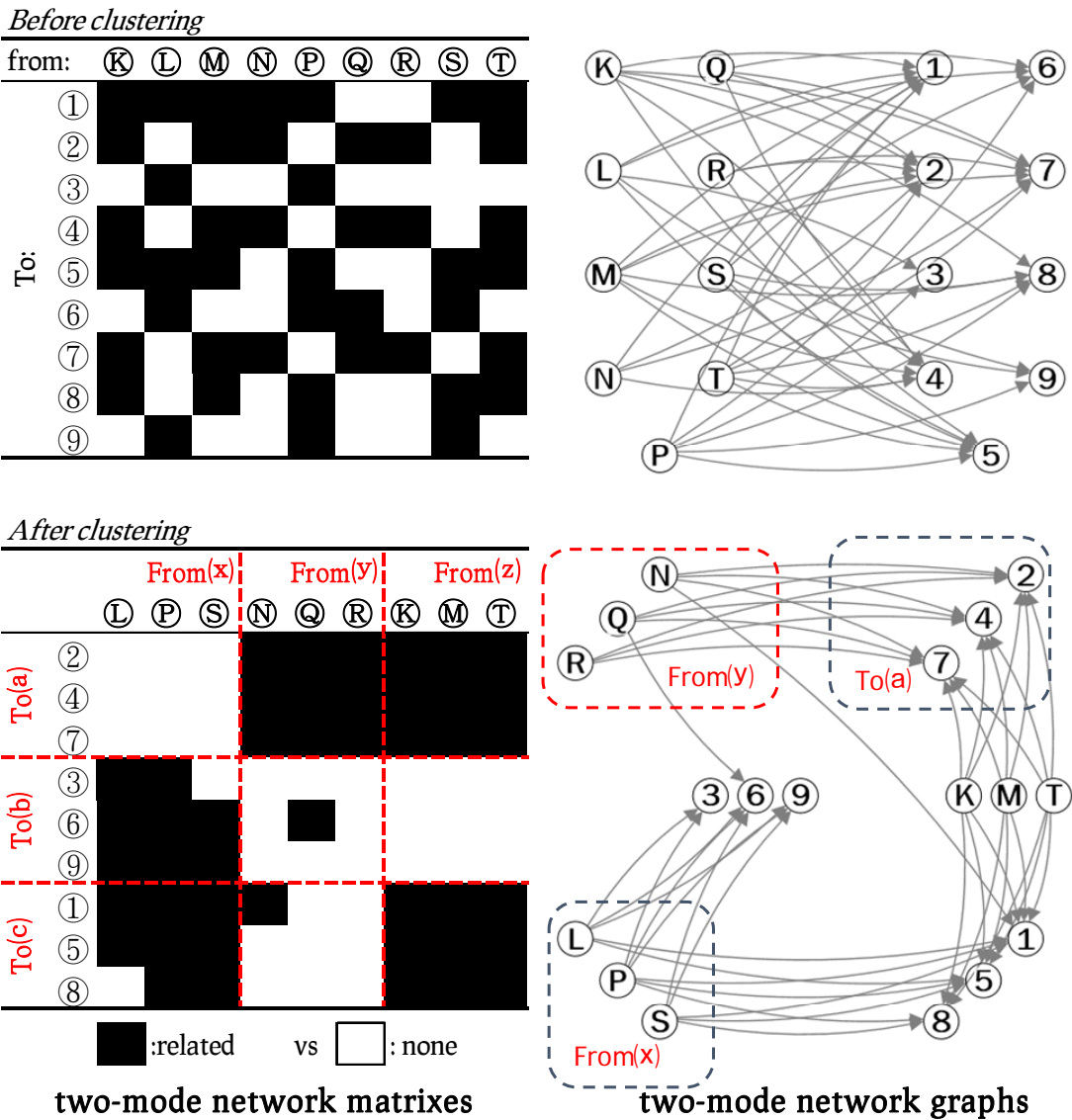


Figure 6-3. A 2-mode network before and after blockmodelling

The top and bottom half of Figure 6-3 shows the contrast before and after blockmodelling. Before blockmodelling, it is difficult to identify, based on the relationship shown, how the nine alphabet nodes can be clustered among themselves, and how the nine number nodes can be clustered among themselves. After blockmodelling, note how the three alphabet nodes of L, P, and S are found to be clustered, showing their common features in having no relationship to

number nodes of ②, ④, and ⑦, and having almost complete relationship to the remaining number nodes. The main benefit of doing blockmodelling is to identify the underlying structure of relationships through clustering. The contrast of the two matrix in Figure 6-3 shows how the 9x9 cells of relationships are clustered into 3x3 blocks, effectively categorizing each of the nine nodes into three categories. One includes From(x), From(y), and From(z), another, To(a), To(b), and To(c). The outcome of such clustering and categorization can be further shown by the network graph at the bottom right of Figure 6-3. It clearly shows that any node in From(x), i.e. ①, ②, or ③, has no link to any of the nodes in To(a). In contrast, any node in From(y), i.e. ④, ⑤, or ⑥ has links to every node in To(b). After clustering through blockmodelling, the network graph is effectively simplified to six categories of nodes: From(x), From(y), From(z), To(a), To(b), and To(c). Thus, by contrasting the relationship (and lack thereof) among categories of nodes, Figure 6-3 demonstrates the difference made by blockmodelling in identifying the underlying structure.

In addition, blockmodelling errors indicate which data points do not fit into the block model, which are illustrated by the label “e1” and “e2” shown in top of Figure 6-4. Label “e1” indicates two data points (⑤-⑥ and ④-①) where the block model expects them to be null (i.e. lack a relationship) but actual relationships do exist (thus non-Null). For example, the relationship ⑤-⑥ is the unfit exception compared to eight other data points in the block of From(y)-To(b). Similarly but indicating the opposite error, label “e2” shows two data points (③-③ and ①-⑧) where the block model expects the existence of relationship but in fact no relationship exists. The bottom of Figure 6-4 shows the ideal block model that the actual network tries to fit. Thus, by summing all the data points that do not fit, the fitting error can be said to be four. Computer algorithms can be implemented to find mathematical solutions that minimize the number of fitting errors, and thus researchers can make sure that the clustering outcome is the optimal outcome (or one such outcome if multiple solutions exist). Thus, running a blockmodelling

analysis should not only minimize (and indicate) the number of unfit data points, but also show the underlying structure of the network, showing the blocks of “complete” (i.e. all relationships present) and “null” (i.e. all relationships absent).

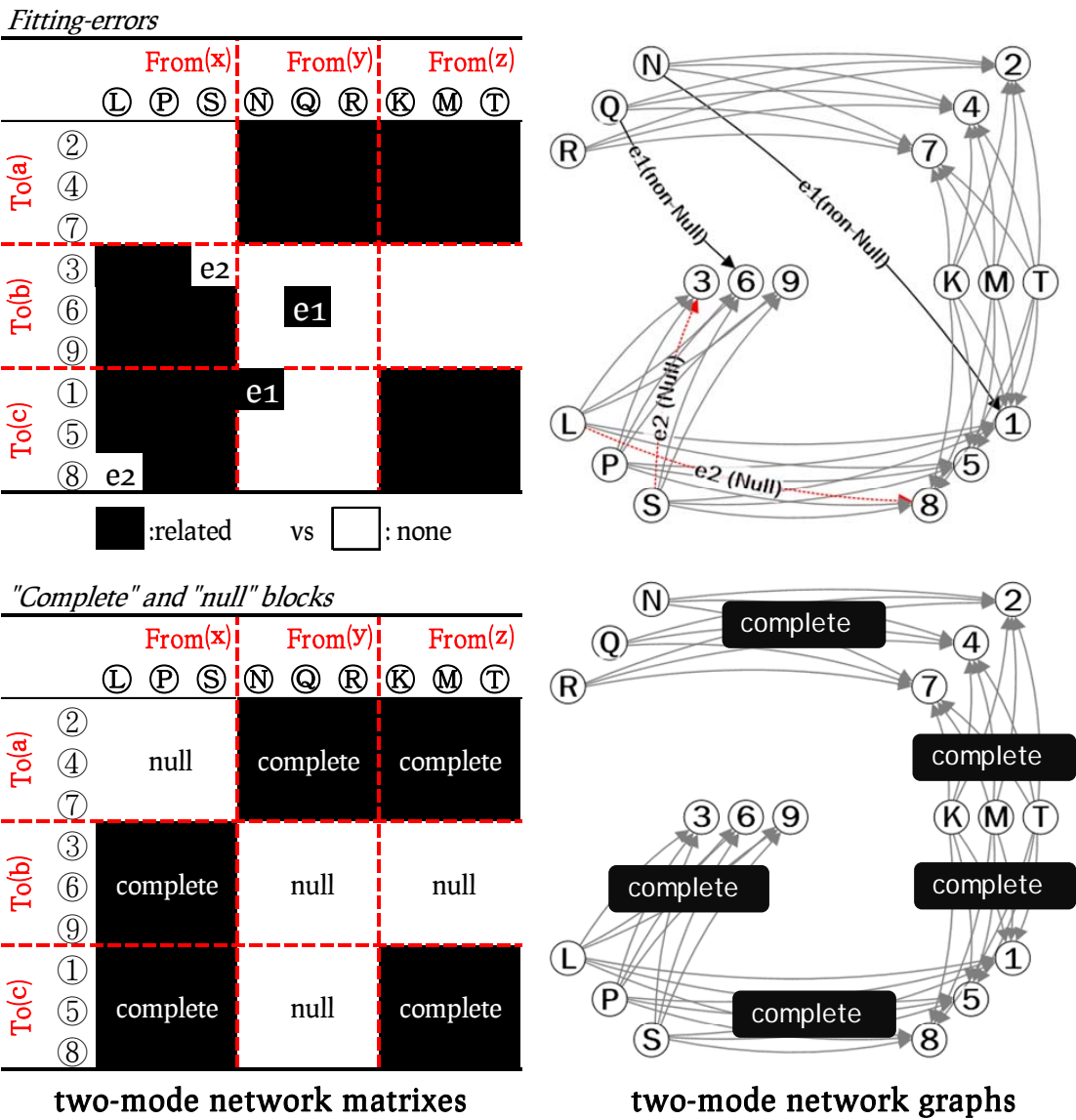


Figure 6-4. A blockmodelling example showing unfit data points

Blockmodelling analysis is thus expected to be useful in showing how “birds of a feather flock together” based on the relationships found in a network. For this research, conducting a blockmodelling analysis will help to cluster the overall visibility networks between SEVs and websites into a manageable number of blocks,

thereby grouping SEVs and websites. It can provide answers to the question regarding “which group of SEVs cite/display which group of websites”.

Data collection strategy. The dataset was collected based on the top-10 results of the first SERP of 3000 search queries that encompass a diverse range of topics, as explained below.

For search queries, I selected approximately 3000 keywords that are relevant to Chinese cultural and/or political topics. Table 6-2 shows that the selection includes all 990 entries in "The Cambridge Encyclopedia of China" (Hook & Twitchett, 1991), the top 10 search terms provided by Baidu and Google (including mainland China, Hong Kong and Taiwan versions) of various categories, major popular cultural references, the names of notable people and some other culturally and politically "sensitive" keywords.

Table 6-2

Sources and numbers of search queries

Categories	Numbers
The Cambridge Encyclopedia of China	990
Top 10 Search Terms (Google and Baidu)	387
Best Film/Popular Music (China, Hong Kong, Taiwan)	364
Modern Concepts (shared with modern Japanese)	171
Notable People	476
—Nobel Prize Winners of Chinese origin	11
—Major Chinese Politicians	187
—Rich People (China, Hong Kong, Taiwan)	82
—100 Contemporary Intellectuals (China)	100
—Major Fugitives From Taiwan	17
—Victims of White Terror in Taiwan	79
Potentially Sensitive Terms	112
—Japanese AV porn stars	48
—Prosecuted and Sentenced Corrupted Chinese Officials	14
—Documented Filtered Words by Great Firewall	50
Fortune500	500
Total	3000

Although other selections or combinations are possible, this selection provides a broad range for this research in relation to what is likely to be prominent in user-generated encyclopaedias across Chinese-speaking regions.

Next, search queries were constructed by transliterating the 3000 keywords according to the respective Chinese orthographic preferences (simplified Chinese for mainland China and Singapore; traditional Chinese for Hong Kong and Taiwan). Using these search queries, corresponding SERPs were collected, parsed and processed by the visibility test described above whereby higher visibility scores are assigned to higher-ranking websites.

As a result, around 270,000 web links were extracted from the SERPs based on the outcome of 3000 search queries submitted across nine search engine variations in 2012. Those links that correspond to the same website were aggregated (e.g., the website of sohu.com aggregates money.sohu.com and women.sohu.com). All education and government websites were further aggregated into respective top-level domain names, such as edu.tw, edu.cn, gov.cn and gov.hk.

6.1.2 Results. To show whether and how the SERP reflect the preferences of hypothetical users across SEVs, the results were tabulated, blockmodelled, visualized, unpacked and analysed as follows, showing significant geolinguistic differences where Baidu Baike and Chinese Wikipedia are the most visible websites and thus leading examples of a range of Chinese-language phenomena.

Tabulating visibility scores. As expected, the visibility scores concentrated on a few highly visible websites, and user-generated encyclopaedias are among the most visible from the dataset, on which the following paragraphs will discuss.

The visibility scores are concentrated, as evidenced by Figure 6-6 which shows that the top-100 websites have over 70% of the visibility scores. On average, the top-10 websites already have over half of the total scores; also, the degree of concentration differs from category to category, ranging from the most

concentrated Nobel Prize winners to the least concentrated Fortune 500 companies. In other words, the Nobel Prize winners are most highly concentrated in the top search results, while the Fortune 500 companies are least so.

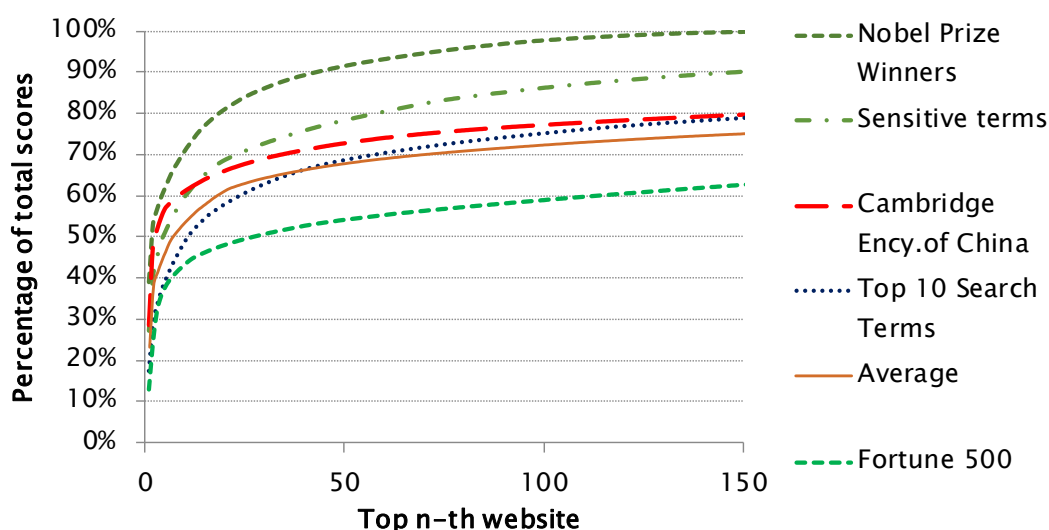


Figure 6-5. Concentrated visibility scores

Table 6-3 lists the scores of the top five websites within the collected dataset and their respective top five domains, indicating concentration of the visibility scores over three encyclopaedia websites. The last column of percentage numbers shows that Chinese Wikipedia (zh.wikipedia.org), Baidu Baike (baike.baidu.org) and Hudong Baike are the most visible sub-domains. Other sub-domains include various versions of Wikipedia (zh-yue: Cantonese, en: English, zh-classical: classical Chinese, ja: Japanese), other Baidu's services (zhidao: question answering, tieba: online forums, wenku: document sharing, and image: image sharing). The fourth most visible website Yahoo is due mostly to, also user-generated, question answering websites for Hong Kong and Taiwan users (tw.knowledge and hk.knowledge) and a blog hosting platform (tw.myblog). Similarly, the visibility of the fifth website, a major Chinese-language portal site Sina.com, is mostly contributed by its blog service (blog), entertainment portal (ent), news, finance portal (finance) and entertainment database (data.ent). Since the subdomain

Chinese Wikipedia on average makes up 95% of Wikipedia and Baidu Baike makes up 87% of Baidu.com, it is reasonable to use website statistics as the approximation of the subdomain encyclopaedia websites for the purpose of this research.

Table 6-3

Visibility scores: the top five websites and their top five domains

Websites and their subdomains	Visibility scores	% of the scores (grand total)	% of the scores (website)
wikipedia.org	4383.52	20.57%	
zh.wikipedia.org	4,171.61	19.58%	95.17%
zh-yue.wikipedia.org	93.14	0.44%	2.12%
en.wikipedia.org	50.40	0.24%	1.15%
zh-classical.wikipedia.org	37.07	0.17%	0.85%
ja.wikipedia.org	31.30	0.15%	0.71%
baidu.com	3513.87	16.49%	
baike.baidu.com	3,066.32	14.39%	87.26%
zhidao.baidu.com	165.84	0.78%	4.72%
tieba.baidu.com	137.56	0.65%	3.91%
wenku.baidu.com	74.19	0.35%	2.11%
image.baidu.com	69.95	0.33%	1.99%
hudong.com	707.40	3.32%	
www.hudong.com	693.38	3.25%	98.02%
tupian.hudong.com	12.77	0.06%	1.81%
so.hudong.com	0.72	0.00%	0.10%
fenlei.hudong.com	0.32	0.00%	0.05%
w.hudong.com	0.21	0.00%	0.03%
yahoo.com	398.94	1.87%	
tw.knowledge.yahoo.com	167.03	0.78%	41.87%
tw.myblog.yahoo.com	121.82	0.57%	30.54%
hk.knowledge.yahoo.com	67.43	0.32%	16.90%
tw.movie.yahoo.com	22.59	0.11%	5.66%
tw.news.yahoo.com	20.06	0.09%	5.03%
sina.com.cn	350.17	1.64%	
blog.sina.com.cn	154.93	0.73%	44.24%
ent.sina.com.cn	69.87	0.33%	19.95%
news.sina.com.cn	52.60	0.25%	15.02%
finance.sina.com.cn	41.76	0.20%	11.93%
data.ent.sina.com.cn	31.01	0.15%	8.86%
Top-5 total	8954.96	43.90%	
Grand total	21306.70	100.00%	

Note that the aggregated visibility scores cannot be converted back to ranking numbers because websites may appear more than once in a single SERP. Still, if we assume that, for most of the time, a website shows up only once in a SERP, then the visibility scores of Chinese Wikipedia (4383.52) and Baidu Baike (3153.87) are equivalent to the visibility scores of 16.24% and 13.01% respectively (by dividing the scores over 3000 search queries and nine SEVs). These scores indicate that both have the average ranking roughly between the first and second in a SERP. The proportion of visibility scores gives a rough indication of the likelihood that a website will be clicked on in the SERPs. The share of visibility scores can be used to compare approximately the level of visibility or the proportion of potential visits directed by search engines. Thus, if the visibility scores correctly reflect the distribution of likely clicks, then the top-5 websites together are expected to get 43.90% (see the last row in Table 6-3) of the traffic. In addition, the average level of visibility is about the same for Chinese Wikipedia and Baidu Baike.

To examine how visible major user-generated encyclopaedia websites are compared to other websites, I calculated the visibility scores of the top-10 websites, and then visualized how much each encyclopaedia got among the top-10. Table 6-4 shows the outcomes for each SEV and for each category of queries. The first stacked bar shows the outcomes based on the 990 search terms of the Cambridge Encyclopedia. For Baidu China, around 80% goes to Baidu Baike; about 7% and 2% for Wikipedia and Hudong Baike respectively. Across different SEV and query categories, the four encyclopaedia websites together account for at least 65% of the visibility scores. Thus, as a genre, encyclopaedia websites are among the most visible. Using the aggregated sum of the visibility scores of the top 10, the ranking of each top-10 website can also be derived for each SEV and each query category, as shown in Table 6-5. Clearly, Baidu Baike (baidu.com), Chinese Wikipedia (wikipedia.org), Hudong Baike (hudong.com) and a business-domain-specific user-generated encyclopaedia called MBAlib (mbalib.com) are among the most visible.

Table 6-4
Percentage of visibility scores: encyclopedia sites among the top-10

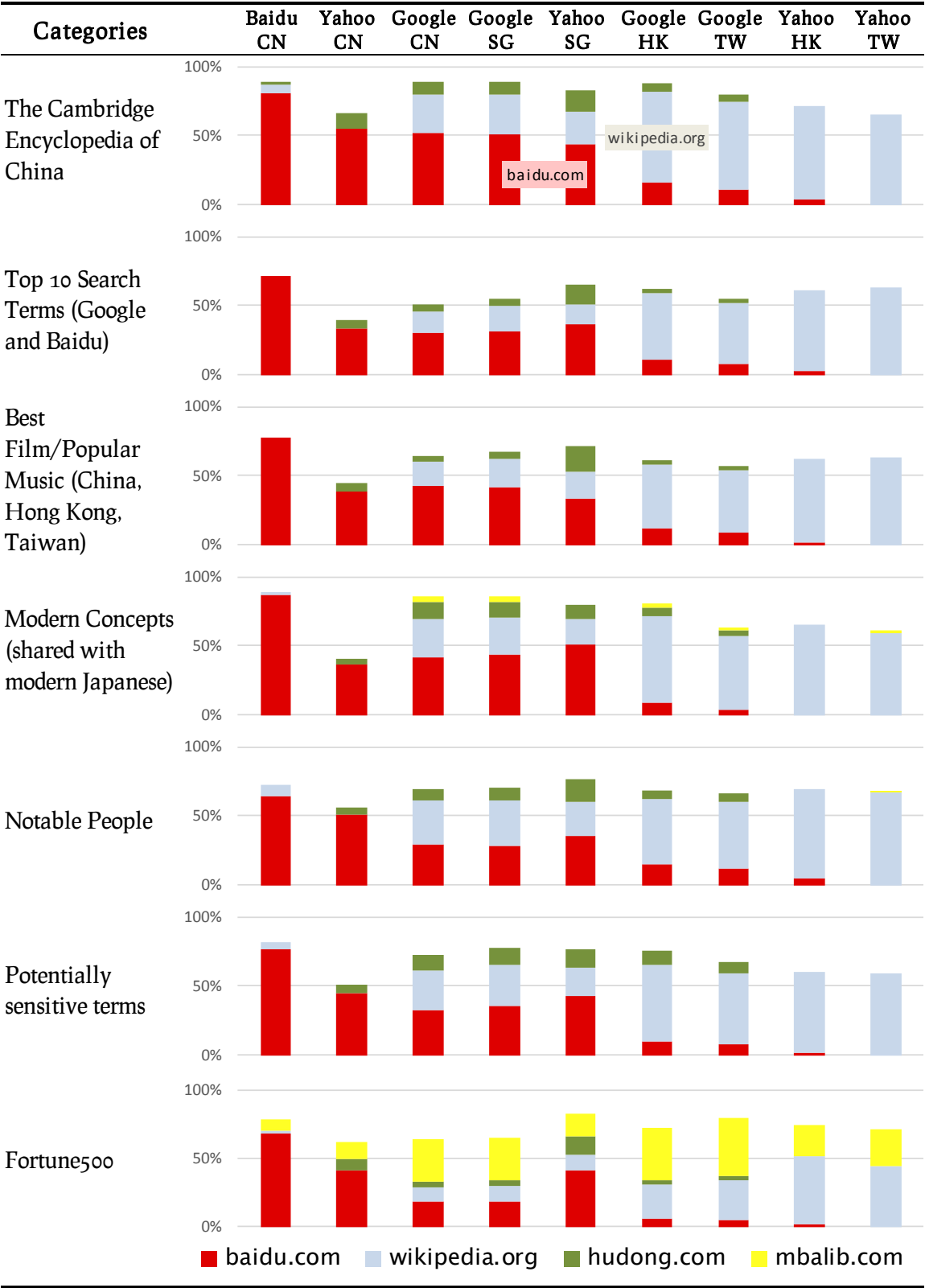
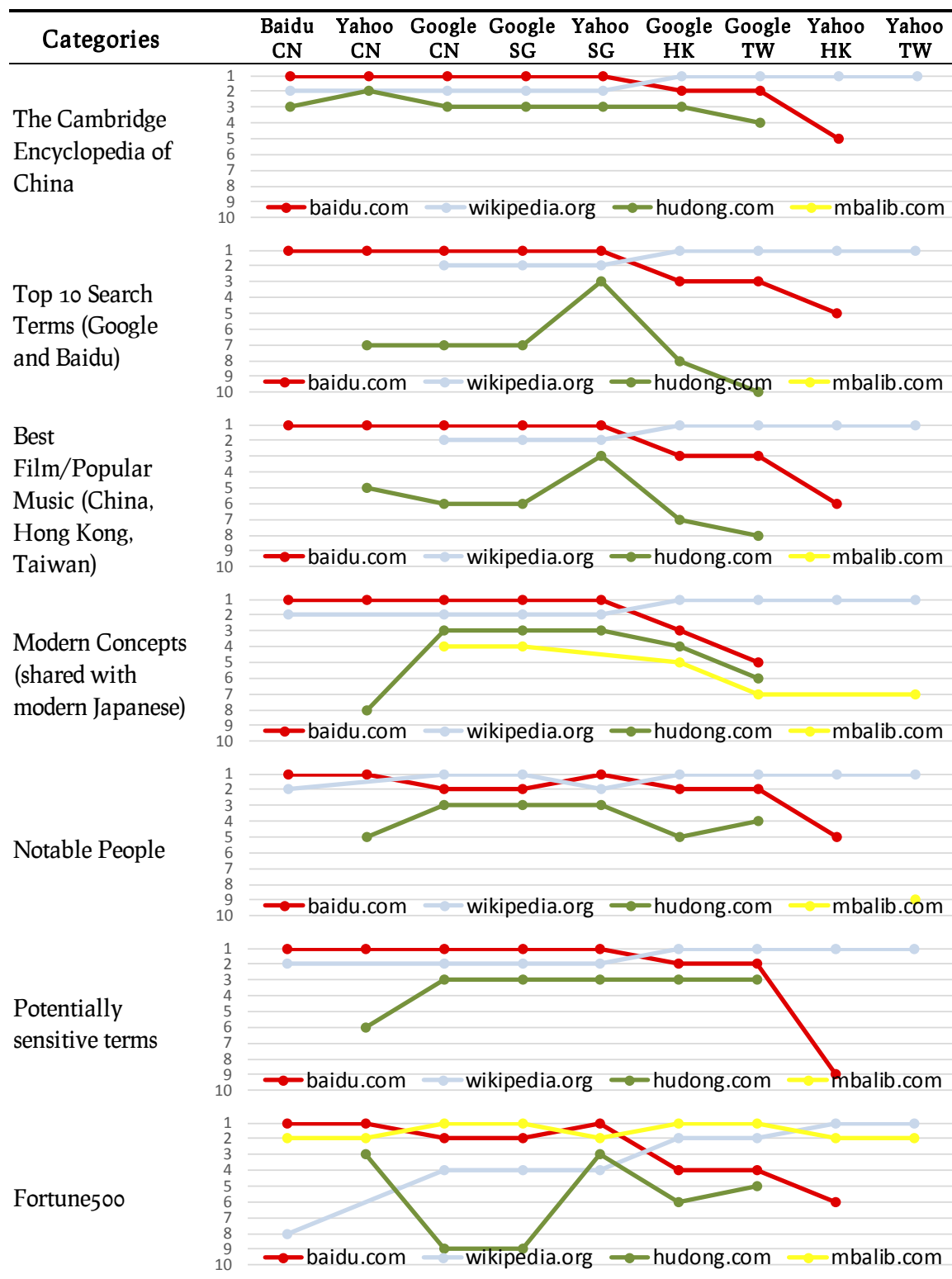


Table 6-5

Ranking of visibility scores: encyclopedia sites among the top-10

Among the most visible websites, Chinese Wikipedia and Baidu Baike are often the top-two, as shown in Table 6-5. In particular, for the query category of Cambridge Encyclopedia (the first row), Chinese Wikipedia is consistently the first or second most visible website, and Baidu is the other second or first most visible website except for Yahoo's Hong Kong and Taiwan SEv. It is also clear that, Hudong.com (the green dots and lines), despite having the self-proclaimed largest number of entry articles, is much less visible than the other two. As expected, the business-domain-specific MBAlib dominates the query category for the Fortune 500 query terms.

These results suggest a link between the choice of SEv and visible encyclopaedia websites. For example, for Baidu_CN (the first column), Baidu Baike has been the most visible, suggesting some level of “self-favouritism” in favouring its own website. Baidu Baike becomes much less visible for SEvs for Hong Kong and Taiwan, where Chinese Wikipedia is more visible. In addition, contrasting Hudong Baike's visibility scores for Baidu_CN, Google_CN and Google_HK, the fact that Hudong Baike is much less visible in Baidu's SEv seems to confirm the unfair competition accusation made by Hudong's CEO against Baidu (Y. Yang, 2011). Depending on the types of search queries, Google_HK and Google_CN rank it from third to ninth position. In contrast, for Baidu_CN, Hudong is not even among the top 20 for many categories of the sampled queries. Indeed, if Google's SERP can serve as an independent third party in the competition between Baidu Baike and Hudong Baike, Google does not render Hudong almost as invisible as Baidu does, although Google does not favour Hudong over Wikipedia.

Such a pattern cannot be readily observed from the pre-blockmodelling tables shown earlier in Table 6-4 and Table 6-5. It can be argued the patterns indicate also how user preferences can influence search engine results, and how each results page is tailored to the pages preferred by respective geolinguistic groups of users. Thus, the various SEvs construct different web spheres that shape user experiences and choices; and in turn, user activities can shape the web spheres.

(Recall the two-way interactions between user experiences and web spheres as demonstrated in Figure 3-1.)

Note that the SEVs in Table 6-4 and Table 6-5, from left to right, are ordered mainly by geolinguistic regions CN, SG, HK and TW. The last four however, are ordered first by platforms (Google first then Yahoo), and then ordered by geolinguistic regions (HK first then TW). This order is not accidental. The mathematical foundation behind this ordering is based on the blockmodelling analysis described above. The order of SEVs provided by the blockmodelling analysis allows us to see the contrasting pattern of visibility between Baidu Baike and Chinese Wikipedia, where Baidu Baike dominates much more from the left SEVs to the right where Chinese Wikipedia dominates. The next section will proceed to explain how the underlying structure of the relationship between SEV and websites can be revealed.

Identifying the underlying structure of citing/displaying. As explained earlier, the mention of a website by a SEV in the SERPs can be considered as a citing/displaying relationship. A blockmodelling analysis of such relationship can thus reveal the underlying structure of citing/displaying made by different SEVs to different search websites. Thus, researchers can transform the visibility scores into the strength of relationships between SEV-website pairs by differentiating between strong versus weak relationships (often using a threshold value).

After several iterations of blockmodelling of the top-100 websites (done with a social network analysis tool called Pajek), the resulting three-by-three blockmodel, as shown in Table 6-6, not only identifies the underlying structure but also clusters SEVs and websites into three groups. For each cell, the colour represents strong (dark) or weak (white) ties. The blockmodelling results have 80 data points (or cells in the Table) that do not fit the model, which amount to 9.67% of the data considered, suggesting that the produced blockmodel can explain over 90% of the data as to their underlying relationship structure. The structure is represented by the three-by-three blocks of cells, as partitioned by the red lines in

Table 6-6. The top-left and bottom-right blocks contain the cells with null relationships, which means SEVs there hardly display the corresponding websites. The remaining seven blocks contain the complete relationship, which means (almost) all SEVs there display the corresponding websites.

Table 6-6

Clusters identified by blockmodeling

Websites (Aggregated)	Baidu CN	Yahoo CN	Google CN	Google SG	Yahoo SG	Google HK	Google TW	Yahoo HK	Yahoo TW
1 wikipedia.org	67.76	0.99	318.95	344.78	325.16	702.52	678.75	1005.58	958.46
4 yahoo.com	1.95	1.54	6.35	6.67	35.76	25.88	33.98	185.31	216.68
8 youtube.com	0.29	0.00	13.72	14.82	3.69	82.43	85.12	62.39	13.41
9 edu.tw	1.88	0.58	5.62	6.48	16.24	15.80	66.77	33.78	82.53
13 facebook.com	0.29	0.00	3.57	3.75	10.02	12.45	29.03	94.65	39.65
18 epochtimes.com	0.00	0.00	2.10	2.68	2.61	25.42	31.23	38.28	29.42
21 gov.tw	0.19	0.25	6.83	6.55	5.30	10.07	35.11	11.19	35.77
... and other 28 websites (The total number for this category of websites is 35)									
6 mbalib.com	15.64	21.98	53.89	54.57	39.31	72.43	72.19	54.17	64.74
10 people.com.cn	13.58	37.05	26.66	27.35	12.42	23.15	27.82	17.43	21.28
12 ifeng.com	23.62	31.41	36.02	36.94	15.38	21.93	23.36	7.10	5.89
... and other 13 websites (The total number for this category of websites is 16)									
2 baidu.com	1156.63	552.07	528.29	540.60	658.81	170.50	124.03	48.38	7.74
3 hudong.com	14.15	95.90	102.50	107.28	252.70	68.80	62.70	3.84	0.02
5 sina.com.cn	49.75	79.62	91.45	91.90	76.43	48.82	44.25	6.31	3.04
7 qq.com	46.50	86.54	46.66	45.84	24.62	14.44	12.33	2.21	1.42
11 youku.com	44.15	75.23	23.77	16.23	25.17	8.37	7.36	1.66	1.02
14 soso.com	17.73	23.85	9.30	8.28	123.12	1.02	1.33	0.89	0.06
15 xinhuanet.com	16.97	17.38	28.01	27.63	12.22	37.14	38.89	1.79	0.39
16 sohu.com	21.02	40.94	34.44	30.22	21.92	10.35	6.84	3.23	1.18
17 163.com	19.65	38.44	37.56	35.58	14.74	10.59	8.29	2.30	1.09
19 douban.com	19.22	24.33	21.81	21.29	10.25	9.82	8.10	1.00	0.97
... and other 39 websites (The total number for this category of websites is 49)									

The underlying structure has effectively clustered the nine SEVs into 3 groups (from left to right, 2, 5, and 2 SEVs for each group) and the 100 most visible

websites into three sets (from top to bottom, 35, 16, and 49 each). The blocks are also visualized by extending the partitioning red lines into the row and column headings, a pattern of relationship structure that is revealed through blockmodelling.

The nine blocks of the data points in Table 6-6 can thus be simplified into Table 6-7, with two blocks of complete null and another seven blocks of complete relationships. According to the Pajek results, the block model in Table 6-7 can explain more than 90% of the relationship of a relatively complex network consisting of nine SEVs and 100 top visible websites in Table 6-6.

Table 6-7

Blockmodeling result matrix

		SEv(x)		SEv(y)					SEv(z)	
		Baidu CN	Yahoo CN	Google CN	Google SG	Yahoo SG	Google HK	Google TW	Yahoo HK	Yahoo TW
Sites(a)	①	null		complete					complete	
	...									
Sites(b)	⑥	complete		complete					complete	
	...									
Sites(c)	②	complete		complete					null	
	...									

two-mode network matrixes

The top-left null block means that the SEv group SEv(x), i.e. Baidu_CN and Yahoo_CN, have little to no display of 35 websites in the set of Sites(a) in their SERPs. The bottom-right null block suggests a similar null relationship between 2 variants in SEv(z) and 49 websites in Sites(c). Both null blocks entail a “missing” display relationship. Table 6-7 thus shows the underlying structure of the complex

network of the top 100 sites and nine SEVs based on missing relationships. Except for the five SEVs in SEV(y) and the 16 sites in Sites(b) which have complete relationships with the rest, the other two groups of variants and two sets of websites are defined by missing relationships (the null blocks).

Figure 6-6 shows the same results in network graph form. Note how the findings indicate the unfilled display patterns. Baidu China and Yahoo China missed Sites(a) and Yahoo Hong Kong and Yahoo Taiwan missed Sites(c).

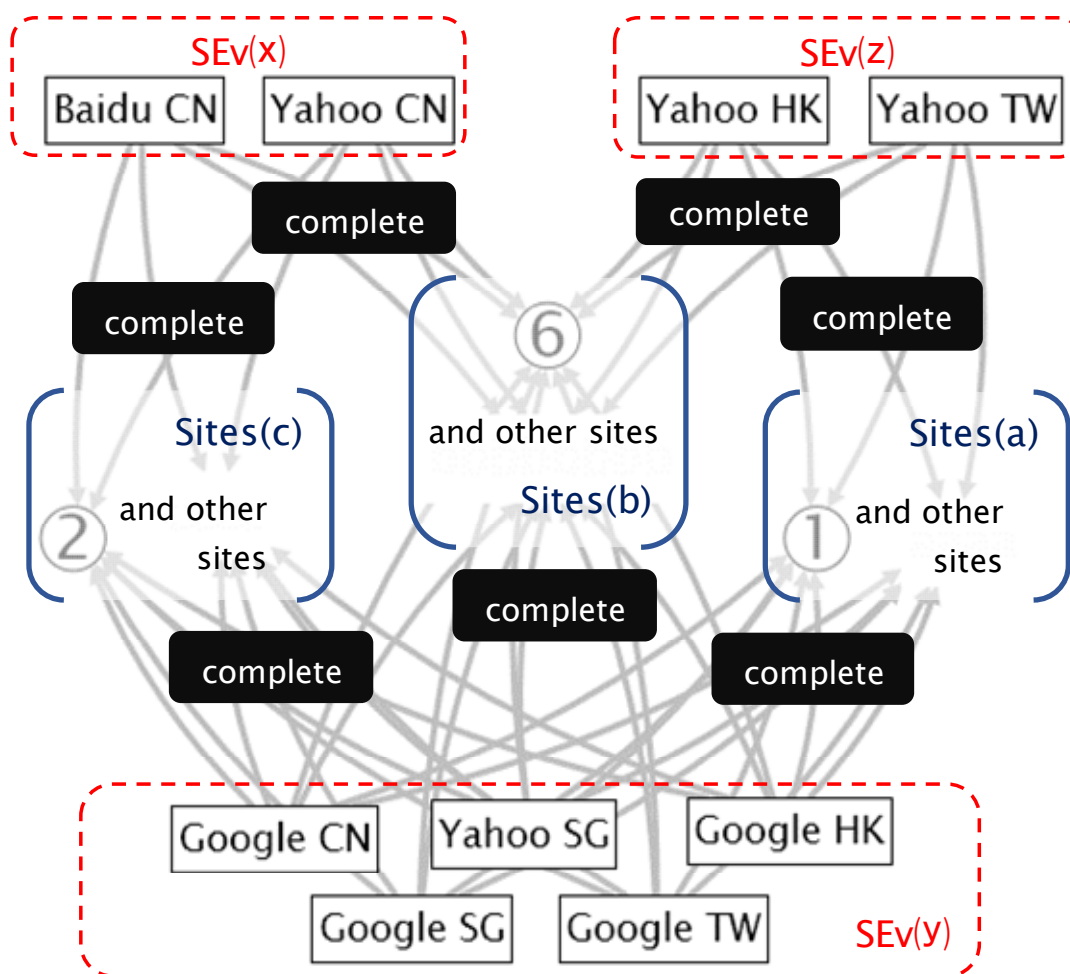


Figure 6-6. Blockmodelling result network visualization

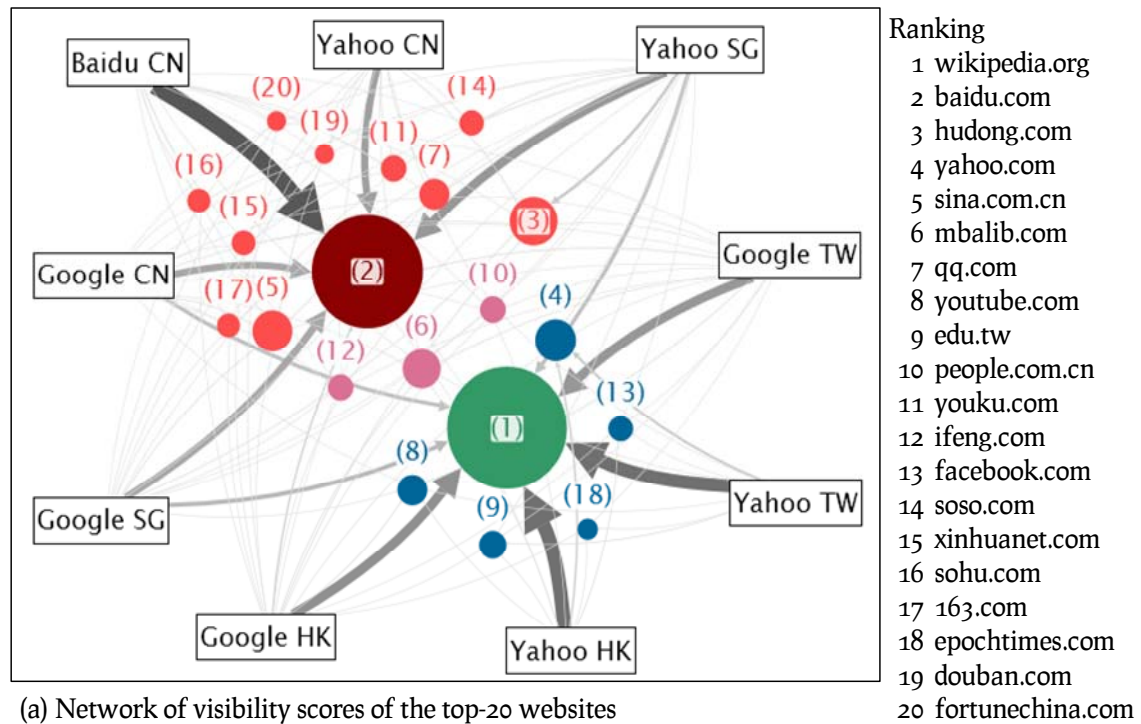
The blockmodelling results suggest some distancing pattern among SEVs and websites. In the middle, SEV(y) largely consists of Google's variants across regions. The remaining two groups differ the most from one another: SEV(z) has two non-

mainland Chinese variants provided by Yahoo; SEv(x) has two mainland Chinese variants. Based on the blockmodelling results, the distancing order between SEvs can be found: first SEv(x) that contains Baidu China and Yahoo China, and then SEv(y) that includes all Google variants and Yahoo Singapore in the middle, and finally SEv(z) that consists of Yahoo Hong Kong and Yahoo Taiwan.

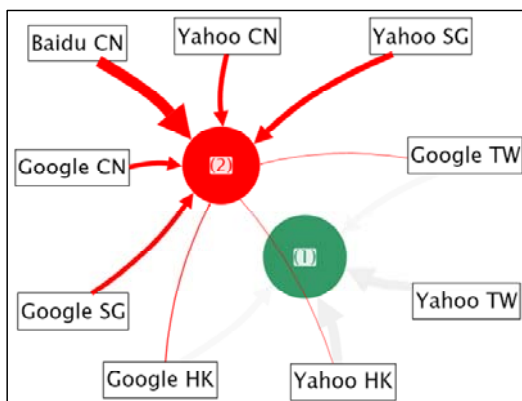
Hence, the blockmodelling outcome (as shown in Table 6-6 and Figure 6-7) explains the underlying relationship structure in terms of missing relationships. What Baidu China and Yahoo China miss are sites such as Chinese Wikipedia, Youtube, Facebook, Taiwanese governmental and educational websites, Falun Gong's newspaper Epoch Times, and the like - most of them hosted outside mainland China. What Yahoo Hong Kong and Yahoo Taiwan miss are largely websites hosted in mainland China, including encyclopaedia websites such as Baidu Baike and Hudong Baike. The missing relationship means that certain websites will not be as visible for certain local SEvs, indicating unrealized or underdeveloped connections. Full connections across all local SEvs are achieved only thinly on the 16 websites in Sites(b), which can be said to provide cultural thickening across all SEvs.

Visualizing the contrast in context. To contrast Baidu Baike and Chinese Wikipedia in the context of different visibility patterns, a network visualization graph is provided in Figure 6-8. The most distancing groups of SEvs are placed at the top left and bottom right respectively, with the top-20 most visible websites represented as disks of different sizes in the middle. The size of the nodes is proportional to the total visibility scores, and the width of the arrow is proportional to the respective visibility score contributed by a SEv to a website. The most visible website Chinese Wikipedia, i.e. node(1) is shown to be grouped with other five top-20 websites, which are among the sites that are largely missed by SEvs Baidu_CN and Yahoo_CN. The second most visible website Baidu Baike, i.e. node(2) is shown to be grouped with other 10 websites, which are among the sites that are largely missed by variants Yahoo_TW and Yahoo_HK. In the middle,

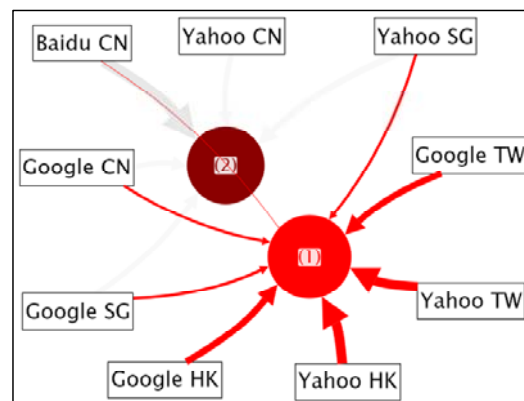
only the nodes (6), (10) and (12) belong to the middle set of websites Sites(b). The bottom two subgraphs (b) and (c) in Figure 6-8 highlight their visibility networks in contrast. Chinese Wikipedia is rendered almost invisible by Baidu_CN and Yahoo_CN, whereas Baidu Baike has low visibility scores for the Hong Kong and Taiwan variants.



(a) Network of visibility scores of the top-20 websites



(b) Baidu Baike's visibility network



(c) Chinese Wikipedia's visibility network

Figure 6-7. Contrasting Baidu Baike's and Chinese Wikipedia's visibility

Contrasting geolinguistic features. The distancing effect can be confirmed by another type of analysis of the same dataset of all the SERPs. Geo-linguistic analysis of the web links, as similarly conducted in Chapter 4, can indicate the

geolinguistic profile of the websites recommended in the SERPs. Figure 6-9 and Figure 6-10 show a geographic contrast between mainland China and Singapore (left) and Hong Kong and Taiwan (right), suggesting a localization effect of the SERPs. Baidu China and Yahoo China give more recommendations to websites with the domain name “cn” (see Figure 6-9) and/or hosted in China (see CN in Figure 6-10). They barely recommend any websites with the domain name “tw” (see Figure 6-9) and/or hosted in Taiwan (see CN in Figure 6-10). In contrast, Google Hong Kong and Google Taiwan provides more than 30% of recommended websites hosted in China (see CN in Figure 6-10).

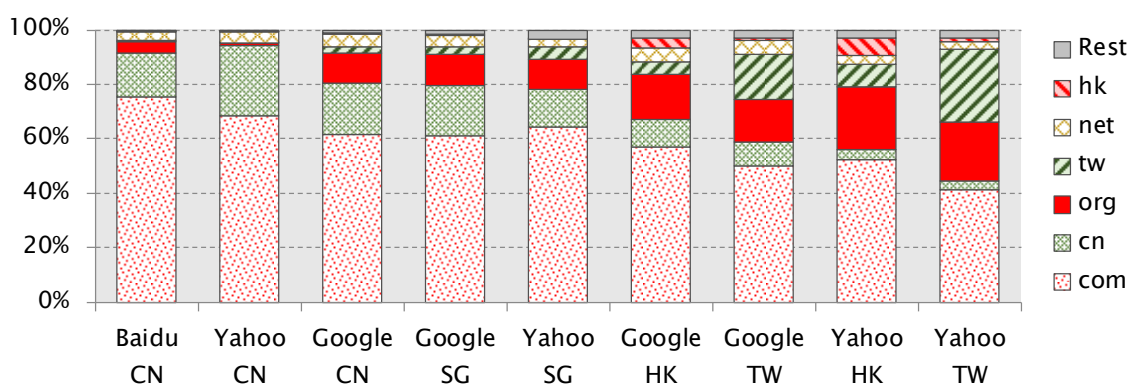


Figure 6-8. Contrast of top-level domain names (TLD)

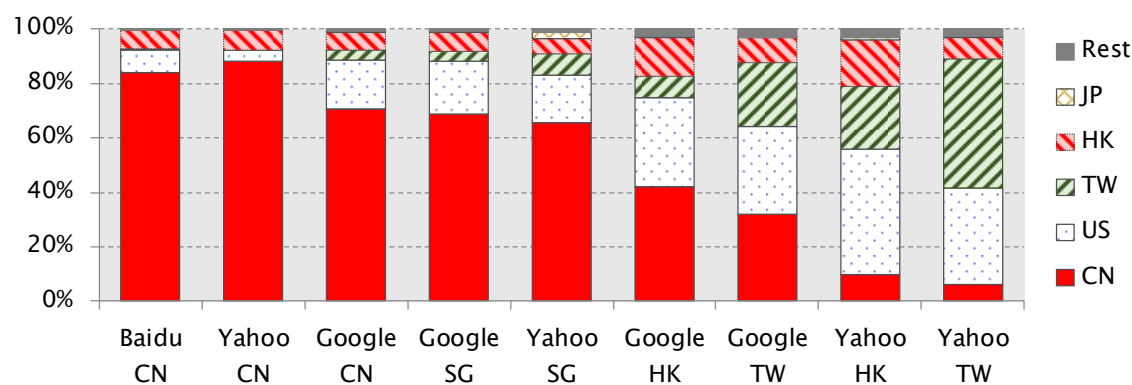


Figure 6-9. Contrast of geoIP locations

Figure 6-11 shows a linguistic contrast between simplified Chinese (left) and traditional Chinese (right), suggesting a similar localization effect. Baidu China and Yahoo China rarely recommends websites written in traditional Chinese, whereas

Google Hong Kong and Google Taiwan provides around 40% of the recommended websites written in simplified Chinese.

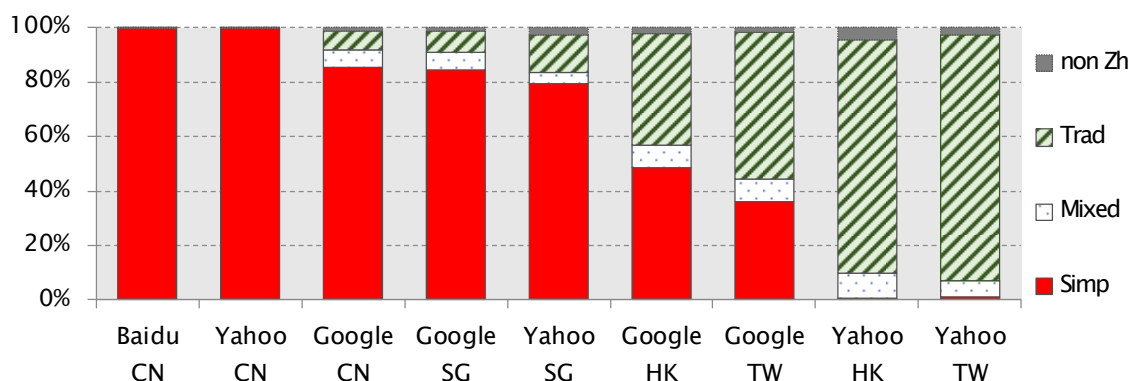


Figure 6-10. Contrast of scripts among the largely Chinese-language links

Despite the above geolinguistic differences, further examination of the datasets shows almost all (97.5%) of the cited/linked websites are written in Chinese. It confirms the expectation that Chinese search queries, when submitted to Chinese-language search engine interfaces, will produce largely Chinese-language results. Sharing a language writing system matters as the foundation for information mediation and exchange on the Web.

6.1.3 Discussion. The above analysis confirms that user-generated encyclopaedia websites are also among the most visible websites for Chinese SERPs, and demonstrates a clear geolinguistic difference within the Chinese-language Internet. The dataset is based on 3000 keywords that cover not only encyclopedic topics for China, but also modern popular, cultural, or political terms, which constitute by far the most comprehensive attempt to data for Chinese-Internet research. Thus, although the findings may differ for other knowledge subjects or domains, the collected dataset and the analysis provide important insights to both the underlying structure and geolinguistic makeup of the Chinese-language Internet. Baidu Baike and Chinese Wikipedia are shown to be the most visible websites and thus leading examples of variegated Chinese-language

phenomena. Since at least one of the encyclopaedias is visible to a specific SEv, any casual users of Chinese-language search engines must have seen and likely clicked through to an encyclopaedia page provided by Baidu Baike or Chinese Wikipedia. I argue that, in view of the fact that search engines are among the major Internet applications for Chinese-language users, the choice of SEvs is likely to lead users to different choices of user-generated encyclopaedias.

With the help of blockmodelling analysis, co-visible search engine results are identified. Chinese-language users across different regions are more likely to visit the same website if they use Google, and for them Chinese Wikipedia will be more visible. Those from mainland China are more likely to be kept away from certain websites including Chinese Wikipedia if they use Baidu or Yahoo China.

The findings here, supported by blockmodelling analysis, have the following implications: Baidu Baike is shown to be visible across seven SEvs (except for Yahoo_TW and Yahoo_HK), and Chinese Wikipedia a somewhat different set of seven variants (except for Baidu_CN and Yahoo_CN). Thus, it is likely that users who use Baidu search (which happens to be the majority of mainland users) or Yahoo China will likely miss Chinese Wikipedia and other websites that belong to the same category in the blockmodelling results, including Epoch Times, YouTube, Facebook, etc. that have been reported as being blocked by mainland China. They also include the websites of government and education institutions in Taiwan and Hong Kong. Arguably, compounded by the fact that Baidu Baike does not include as much and as diverse information sources as Chinese Wikipedia does (see Chapter 4), the use of Baidu Search reinforces the information boundary between inside and outside mainland China.

In contrast, since the majority of users from Taiwan and Hong Kong use Google, the information results provided by both Google and Wikipedia will expose them to mainland Chinese information sources. The use of Google and Wikipedia by Taiwan and Hong Kong users partly overcome the information boundary because mainland Chinese sources, including Baidu Baike are visible.

SERPs effectively recommend websites to users, and different sets of recommendations suggest different sets of websites. Google and Chinese Wikipedia are shown to have more boundary-overcoming cultural thickening for their mixing of Chinese-language content. This contrasts with Baidu and Yahoo, both of which provide more locally confined content.

Thus, different website recommendation patterns vary by the choice of platforms, and geolinguistic regions reflect the cultural political complications of the Chinese-language Internet. In particular, while the offline boundary between Hong Kong and Taiwan seems to be overcome (also mostly by Google and Wikipedia), that between mainland China and Hong Kong seems to be reinforced. If not using Google China, mainland Chinese users are likely to experience a specific kind of “nationalization” (i.e. limited to the mainland segment of China’s Internet). In contrast, Hong Kong and Taiwan users are likely encounter a greater “trans-nationalization” phenomenon, especially when they use Google and/or Chinese Wikipedia. It is particularly intriguing that all the Hong Kong variants share more similar results with the Taiwanese variants and much less so with mainland Chinese variants, while Hong Kong is much closer to mainland China geographically, politically and administratively.

There are of course obvious limitations for the findings presented here. First, the selection of search query, while significant larger than previous social scientific research on Chinese-language search engines (Jiang & Akhtar, 2011), is still a limited sample. Second, due to limitations of space, this chapter has not been able to detail the different findings for different categories of search queries. Third, only standard Mandarin Chinese terms were used for this research, overlooking other possibilities of written Cantonese queries that are likely to be used by Hong Kong users (Chau, Fang, & Yang, 2007). Fourth but not last, only the default setting for each localized search engine was analysed.

While the findings may be limited in scope, I have demonstrated the usefulness of using visibility tests for a wide-ranging analysis. For instance, this

type of analysis can help online linguistics research by analysing different SERP outcomes for regions that use a shared writing system but with regional variants, such as the difference between Egyptian Arabic and Maghrebi Arabic. Both the data collection and analysis strategies can be systematically extended for other contexts. Various SEVs can be chosen for research for almost all the other languages in the world, including languages that are transnational such as Arabic, Hindu, Tamil, English, Spanish, Portuguese, etc. Researchers can thus further interpret the merging and diverging SERP outcomes for research questions that are relevant for global, transnational or inter-cultural communications on the one hand, and another set of questions for human-computer interaction and information systems on the other. Also, the focus on examining geo-linguistic factors as important variables for understanding search engines can contribute to the development of geo-linguistic analysis of the Web (Liao & Petzold, 2010; Petzold & Liao, 2011). This type of analysis can also be adopted for market and industry applications when geo-linguistic identifiers are central (DePalma, 2002; Dunne, 2006).

I will now use a different method to examine the visibility of the two encyclopaedias that complements and extends these findings from search engine visibility; namely, how the content of the two encyclopaedias travels through microblogs.

6.2 Microblog posts

The microblogging services Sina Weibo¹⁵ and Tencent Weibo¹⁶ were launched in 2009 and 2010 respectively and have contributed to the rapid growth of

¹⁵ Sina Weibo (Weibo), similar to Twitter, is one of the most popular microblogging social networking sites in the Chinese-speaking communities globally.

online advertising markets in China (CNNIC, 2013; iResearch, 2013). The term “Weibo” (wēi bó 微博) is the transliteration of the term “microblog” in Mandarin Chinese. Weibo allows users to publish short messages like tweets with a 140-character limit per post. Like other major companies in China and elsewhere whose business is user-generated content (hereafter UGC), the amount of content and traffic a platform can garner is crucial for its survival.

Microblogs can be seen as part of the historical development of the rise of user-generated content since 2005 (Liao, 2014). A recent survey has shown that user-generated content is more popular in Asia than in the West (Bolsover et al., 2013). By the first half year of 2010, the content produced by amateur Chinese Internet users surpassed that produced by professional websites (Xiaoji Qiang, 2010). According to a survey report based on 2012 data, around 66.1% of mainland Chinese Internet users used a blog/personal space, 54.7% used microblogs, and 48.8% used social networking websites (CNNIC, 2013). One US market report found that 47% of Chinese broadband users had contributed to UGC in China, including online review sites, forums, blogs, etc., and that this UGC content had influenced 58% of purchase decisions in China. Both numbers are much higher than the corresponding statistics in the US (Netpop Research, 2007).

The growth of UGC in mainland China may have contributed to the creation of a so-called “public opinion monitoring sector” (yúqíng jiāncè fúwù shìchǎng 輿情監測服務市場), where managers (mostly from the public sector) buy up-to-date information about the latest online public opinion in mainland China. The online public opinion data relies heavily on processing and mining UGC, especially from microblogs. One of the most successful example is the Online Public Opinion Monitoring and Measuring Unit of the People's Net (rénmín wǎng yúqíng jiāncè shì 人民網輿情監測室), which alone has an annual revenue of close

¹⁶ Tencent Weibo is a major competitor for Sina Weibo. Tencent Weibo has a user base from its instant messaging service QQ.

to 200 million RMB (T. Luo & Li, 2010; Y. F. Tan, 2012), or about £20 million. It is noted that the Chinese government and corporations have been developing and using “Internet public opinion monitoring systems,” which involve intensive collection and analysis of web data (T. Luo & Li, 2010; Y. F. Tan, 2012). As of 2014, there is little Internet research in English about this development of public opinion monitoring, except for a panel discussion that took place during the 2013 Chinese Internet Research Conference (which I helped organize and took part in). Further work will be needed to compare the political and commercial uses of microblog data in China and elsewhere.

As expected, microblog platforms that are hosted in mainland China are subject to Beijing’s filtering and censorship regime. Several recent research papers have reverse-engineered the censorship mechanism, mostly based on the selection of keywords to understand how censorship works on the Weibo platforms. Analysing sensitive terms on Sina Weibo and Twitter, researchers found initial evidence that content is filtered or censored on Sina Weibo (Bamman, O’Connor, & Smith, 2012). Researchers at Hong Kong University who systematically collect Weibo data on a website called Weiboscope have also conducted censorship research (Fu, Chan, & Chau, 2013; Fu & Chau, 2013). Another group of researchers found that individual criticisms of the government are permitted but that collective expression (which might generate protest against the government) is censored (King, Pan, & Roberts, 2012). Yet another study tracked and quantified sensitive topics on Weibo and found that the posts were often short-lived and confined to a small core of users who posted sensitive content, suggesting that multiple layers of filtering produced effective censorship to contain the spread of such messages (T. Zhu, Phipps, Pridgen, Crandall, & Wallach, 2012).

According to an industry report, in 2012, about 88% of Internet users in mainland China are Sina Weibo users (DCCI, 2012). Weibo platforms are thus important sites of observation to see how user-generated encyclopaedias like Chinese Wikipedia, Baidu Baike and Hudong Baike (Liao, 2014) are mentioned and

used. (Baike means “encyclopedia” in Chinese. Baidu is the name of China’s major search engine company. Hudong means “interactive” or “interaction” in Chinese.) First, the microblog posts are likely to contain the views and experiences of Chinese Internet users with the encyclopaedias. Second, since microblog platforms are also user-generated content platforms, it is possible that the users of microblogs are more likely to use user-generated encyclopaedias. Third, researchers can observe how Beijing’s censorship influences the use or perception of user-generated encyclopaedias. Especially in the case of Sina Weibo, which reaches a large enough number of mainland Chinese users, the posts can provide important indications about how mainland Chinese users are receptive to the two online encyclopaedias. Fourth, users are likely to use microblogs (both reading and writing), which can be regarded as one amongst several online information engagement practices (Bennett & Wells, 2009) in conjunction with search engines and user-generated encyclopaedias. Researchers can thus observe how users use different platforms in combination.

The concept of “information engagement” is used in two disciplines with different audience and research concerns. In the context of civic engagement including youth engagement, the focus is to engage citizens for more and better political participation, as proposed by W. Lance Bennett and Chris Wells to build “civic learning repertoires” for “information engagement” (Bennett & Wells, 2009). In the context of warfare, including information warfare, the practical concern is to provide actionable guidelines to inform foreign and domestic audiences (US Army, 2009), which has been discussed as part of the redefinition of information warfare (Porche, Paul, York, Serena, & Sollinger, 2013). In the Chinese context, the Internet is central for civic engagement (Damm, 2008; Yongnian Zheng, 2007) and the state’s cyberwarfare doctrine (Hagestad II, 2012), and the study of the use and reception of user-generated encyclopaedias expressed on microblogs can be conducted as a case study of information engagement in both senses.

This research aims to explore what Chinese-language microblog users do with major user-generated encyclopedias such as Baidu Baike (Baidu's encyclopedia website) and Chinese Wikipedia (Liao, 2014) by analysing the content of their posts. As user-generated encyclopaedias are expected to solve information problems for users, microblog content about users' experiences and perceptions should indicate the status (and thus also the gaps) of information engagement experiences with the two major online Chinese encyclopaedias as expressed in a major type of platform that constitute a major part of Chinese public opinion.

6.2.1 Methods. Computer-assisted quantitative methods (including several Chinese natural language processing and text-mining techniques) were first used to identify patterns and posts relevant to the research question, and then qualitative methods were used to interpret the findings. The approach effectively analysed the public discussions that mentioned each of the two encyclopaedias in microblog posts.

Similar to the ways in which Zhu, Phipps, Pridgen, Crandall, & Wallach (2012) tracked and quantified censorship by selecting certain topics for research, I developed a means of filtering relevant posts that mention Baidu Baike or Chinese Wikipedia, which produced a dataset that combines data from Sina Weibo and Twitter.

This exploratory method mimics the processes that are probably employed for public opinion monitoring. Relevant posts are first screened from designated online spaces using keywords based on the analyst's judgement. Although there are no guidelines for how many and which terms yield sufficiently comprehensive results, it can be expected a few iterations will yield certain heuristics for gauging the selected posts. Using computer-assisted quantitative methods, analysts then categorize or code the posts in such a way to find relevant patterns for further content and contextual analysis. The combination of qualitative and quantitative methods is bound to be further refined in the future as the analysis of microblogs

advances. I now examine how the Chinese term ‘wiki’ (in a number of variants) is mentioned in microblog posts.

Data selection and data sets. For data collection, the Chinese term (both simplified and traditional characters) “wiki” was used instead of the more specific term “Wikipedia”, in addition to specific Chinese terms such as “Baidu Baike” and “Chinese Wikipedia” (see Table 6-8 for specific Chinese characters). Two unrelated entities were expected to yield false positives since they contain exactly the same characters of popular Chinese term of “wiki”: One is “WikiLeaks” with its variant Chinese terms and another is a Hong Kong businessperson Ricky Wong, whose first name happens to share the exact same Chinese characters as “wiki” (see the bottom row in Table 6-8).

Table 6-8

Wikipedia-related and Wiki-similar Chinese terms

	Simplified Chinese	Traditional Chinese	Notes
Baidu Baike	百度百科	百度百科	unambiguous term
Chinese Wikipedia	维基百科	維基百科	unambiguous term
Hudong Baike	互动百科	互動百科	unambiguous term
wiki	维基	維基	ambiguous term
wikileak	维基解密	維基解密	
wikileak	维基揭密	維基揭密	alternative name
wikileak	维基泄密	維基泄密	alternative name
"Wong Wiki"	王维基	王維基	Ricky Wong, unrelated to Wiki

Two unrelated entities were expected to yield false positives since they contain exactly the same characters of popular Chinese term of “wiki”: One is “WikiLeaks” with its variant Chinese terms and another is a Hong Kong businessperson Ricky Wong, whose first name happens to share the exact same Chinese characters as “wiki” (see the bottom row in Table 6-8). I am aware that

there are other Chinese translations of the term “wiki”, especially during the time when the term was introduced in the Chinese-language context. Nonetheless, by the time of data collection around 2011, the translation used here had become the dominant one.¹⁷

The keyword filter used here can introduce other false positive outcomes since the Chinese term “wiki” can also refer to general wiki technologies and practices, yet in this case, the resulting dataset can be used to see how dominant Wikipedia or wiki-implemented encyclopaedias are on the subject of wiki.

Based on the keyword list, two datasets were retrieved from two data intermediaries. The WeiboScope dataset, provided by Dr. Fu at the Hong Kong University, randomly sampled results from Sina Weibo between January 2011 and April 2012 (see Table 6-9 for details). Second, the DiscoverText dataset contains publicly available microblog posts from Sina Weibo and Twitter for a time period of 23 days in 2012. Table 6-9 summarizes the number (N) and proportion (percentage) of posts from the two platforms and the two intermediary sources, showing a disproportionately large dataset from Weibo (81%).

¹⁷ In fact, according to the comparison of search queries from Google Trend, the translation “维基/維基” has outnumbered alternative translation such as “维客/維客” or “围记/圍紀” as early as 2005: See urls:(1) <http://www.google.com/trends/explore#q=%E7%BB%B4%E5%9F%BA%2C%20%E7%B6%AD%E5%9F%BA%2C%20%E7%BB%B4%E5%AE%A2%2C%20%E7%B6%AD%E5%AE%A2%2C%20%E5%9B%B4%E8%AE%Bo&cmpt=q> or (2) <http://www.google.com/trends/explore#q=%E7%BB%B4%E5%AE%A2%2C%20%E5%9B%B4%E8%AE%Bo%2C%20%E5%9C%8D%E7%B4%80%2C%20%E5%85%B1%E7%AD%86%2C%20%E5%85%B1%E7%AC%94&cmpt=q>

Table 6-9

Posts collected: including the expected false positives

Microblog platform	WeiboScope		DiscoverText		Total	
	N	%	N	%	N	%
Twitter	0	0%	11640	63%	11640	19%
Weibo	42992	100%	6865	37%	49857	81%
Total	42992		18505		61497	
Timestamps						
Earliest	03 Jan 2011 00:21		20 Mar 2012 18:08		03 Jan 2011 00:21	
Latest	05 Apr 2012 13:05		12 Apr 2012 12:09		12 Apr 2012 12:09	
Time Duration	458 days		23 days		465 days	

The first step then was to remove the false positive data (unrelated to the research topic) included in Table 6-9, and then also to remove another set of “jamming” posts targeted at Wen Yunchao, resulting in a reduced dataset as summarized in Table 6-10. Wen Yunchao is a mainland Chinese activist who launched a series of online campaigns against Internet censorship in China, through writing messages on blogs and other social media platforms.

Table 6-10

Posts collected: false positives removed

Microblog platform	WeiboScope		DiscoverText		Total	
	N	%	N	%	N	%
Twitter	0	0%	1152	15%	1152	3%
Weibo	37185	100%	6330	85%	43515	97%
Total	37185		7482		44667	
Timestamps						
Earliest	03 Jan 2011 00:21		20 Mar 2012 18:08		03 Jan 2011 00:21	
Latest	05 Apr 2012 13:05		12 Apr 2012 12:09		12 Apr 2012 12:09	
Time Duration	458 days		23 days		465 days	

The removal of posts targeted at Wen Yunchao resulted in a notable reduction (more than 10,000 posts) in the Twitter dataset; so, it is worth explaining why these were removed. I examined these posts and noticed that almost all of them followed the format: “@wenyunchao Some Entry or Web Page Title – Baidu Baike or Wikipedia” (for example, “@wenyunchao 李明博 - 维基百科” and “@wenyunchao 山海关_百度百科”). These posts were often written by users with ID names that are likely to be randomly generated by machines (e.g. 5sx63cyd, bd8wwkp). All these machine-generated posts were posted to Twitter within the 23 days covered by DiscoverText as shown in the first row of Table 6-11. These posts mostly cited Baidu Baike, but a few posts cited Chinese Wikipedia. The two posts containing “wenyunchao” from the WeiboScope dataset, in contrast, were sent by human users. As these posts would skew any content analysis, I decide not to include them, but rather to consider, separately, how the cyber-attacks against Wen Yunchao relate to the research here.

Table 6-11

Posts that mention wenyunchao (the set "WYC")

Microblog platform	Dataset	BB	CW	HD	"wiki"	wikileak	"Wong"
Twitter	DiscoverText	8755	1426	0	1524	10	4
Weibo	HKU	0	0	0	2	2	0
Total		8755	1426	0	1526	12	4

The remaining 44667 posts are then processed for Chinese-language word-tokenization, a necessary procedure to analyse Chinese-language texts because, unlike languages such as English, Chinese is written without space.

From overview to detailed analysis. To compare the use and reception of the two encyclopaedias as reflected in the dataset, I conducted both an overview and a more specific detailed analysis as follows.

I first ran a word frequency analysis to obtain an overview about hot topics and frequently associated words. Then I compared the posts that mention the two

encyclopaedias individually. In the process, I grouped and differentiated microblog posts using computer-assisted techniques (including Latent Semantic Analysis) so as to break down the large number of posts into manageable clusters. Originally developed for word-document relationship analysis in the field of natural language processing and information retrieval, these techniques proved suitable for the following reasons. First, they can process a large amount of documents, assisting researchers in clustering and differentiating documents and words for further analysis. Second, they mimic or even simulate the way the online texts are processed by search engines and social media analysts (including China's online public opinion monitors). Third, techniques such as Latent Semantic Analysis has been applied to Chinese-language texts with noticeable success (B. Chen, Yeh, Huang, & Chen, 2006; Yulian Yang & Xie, 2008), solving effectively the matching problems caused by synonymy or polysemy (Yu, Fan, Guo, & Geng, 2006). Finally, by repurposing the quantitative techniques for cultural research, researchers can ground their analyses not only on specific relevant words, but also on the latent semantic relationships amongst them that are not immediately apparent to casual readers.

For this study, I implemented such analysis using a combination of NLTK (Natural Language Toolkit), scikit-learn (a set of machine learning tools) and gensim ("generate similar" documents). After the posts were clustered, I categorized the relevant posts under the two themes (filtering/censorship regime and information engagement) to elicit how the two encyclopaedias were discussed.

6.2.2 Results. The findings are now presented, proceeding from overall patterns to specific individual posts. Table 6-12 lists the most frequently occurring words. The most mentioned term is "encyclopaedia" (bǎikē 百科), after which the Chinese names of the three encyclopaedias follow in this order: Baidu Baike, Wikipedia and Hudong Baike. The thesis focuses on the posts mentioning the first two of these.

Table 6-12

Most frequently-occurring words

Ranking		Word	Counts	Ranking	Word	Counts	Ranking	Word	Counts	Ranking	Word	Counts	Ranking	Word	Counts				
1	Encyclopedia	百科	39951	31	Company	1601	61	Sprinkle over	撒过	1346	91	Support	支持	1143	121	Content	内容	833	
2		百度	22672	32	Harm	危害	1600	62	Test	检验	1344	92	Wikipedia	维基百科	1096	122	Event	事件	816
3		Wiki	21495	33	(English)	wikipedia	1570	63	Initiation	引发	1342	93	Raiders	攻略	1066	123	Address	地址	814
4	Baidu	Baike	18623	34	Addition	增加	1569	64	Kai-fu Lee	李开复	1338	94	Data	资料	1060	124	Like	喜欢	808
5	Wikipedia	维基百科	10780	35	Ha ha ha	哈哈	1561	65	Nerve	神经	1331	95	Come from	来自	1050	125	Name	名字	775
6	Interactive	互动	6948	36	Cause	导致	1536	66	Times	万倍	1324	96	English	英文	1046	126	Enterprise	企业	773
7	Hudong	Baike	6583	37	Reference	参考	1522	67	Notes	笔记	1312	97	Full attack	全攻	1033	127	Friend	朋友	771
8	China	中国	4428	38	Life	生命	1518	68	Jokes	恶搞	1308	98	Raiders	全攻略	1033	128	This is	这是	760
9	Floating	Spirit	3954	39	The book	全书	1513	69	100,000 times	十万倍	1307	99	Networking	联网	1031	129	Release	发布	756
10	Share	分享	3291	40	One kind	一种	1495	70	Bad	不良	1290	100	Chinese	中文	1025	130	Yuan Hong	袁弘	753
11	Reply	回复	3023	41	Not work	不行	1474	71	Necessary	就要	1286	101	Japan	日本	1022	131	Work	工作	751
12	Activity	活动	2732	42	Network	网络	1453	72	Search	搜索	1283	102	Interconnected	互联	1016	132	Hope	希望	750
13	Excessive	超标	2682	43	Occur	发生	1444	73	Scientific name	学名	1268	103	Internet	互联网	1003	133	Stuff	东西	738
14	Wallet	钱包	2608	44	Community	社会	1427	74	Serious harm	严重危害	1265	104	News	新闻	996	134	Global	全球	732
15	Wiki	维基	2310	45	In vivo	体内	1422	75	Ridicule	调侃	1264	105	Malignant	恶性	995	135	Doctrine	主义	721
16	Microblogging	微博	2290	46	Similar	类似	1421	76	Notebook	笔记本	1263	106	Culture	文化	988	136	Editor	编辑	720
17	United States	美国	2239	47	University	大学	1418	77	Syndrome	综合症	1247	107	User	用户	986	137	Economy	经济	700
18	Recommend	推荐	2222	48	Knowledge	知识	1411	78	Sitting	坐姿	1241	108	Attention	关注	971	138	News	消息	684
19	Discovery	发现	2217	49	Oxygenated	含氧	1400	79	(English)	syndrome	1235	109	Hong Kong	香港	969	139	Movie	电影	675
20	World	世界	2181	50	Oxygen	含氧量	1399	80	(English)	piriformis	1232	110	Mobile	手机	963	140	Children	孩子	674
21	Introduction	介绍	2130	51	Temporarily	暂时	1399	81	Ischial	坐骨	1225	111	Tumor	肿瘤	959	141	Service	服务	665
22	Website	网站	2008	52	Oxygen	氧量	1399	82	Intelligence	智力	1225	112	Government	政府	956	142	Quote	引用	657
23	Forwarding	转发	2008	53	Canadian fish	加鱼	1395	83	Sciatic	坐骨神经	1222	113	Cancer	恶性肿瘤	946	143	Golden Egg	金蛋	655
24	Free	自由	1938	54	Originally	原本	1380	84	Country	国家	1211	114	People	人民	933	144	Google	谷歌	653
25	Health	健康	1795	55	Complex	综合	1367	85	History	历史	1208	115	Media	媒体	910	145	Science	科学	651
26	First	第一	1752	56	Time	时间	1366	86	Information	信息	1197	116	Research	研究	886	146	Data	数据	647
27	Shanghai	上海	1714	57	Soon	眼看	1365	87	Beijing	北京	1193	117	google	google	877	147	Include	包括	634
28	Click	点击	1701	58	Encyclopedia	百科全书	1353	88	Face	正视	1191	118	Little	小小	877	148	System	系统	630
29	Entry	词条	1669	59	Hypoxia	缺氧	1352	89	Because fish	因鱼	1181	119	Life	生活	850	149	Computer	电脑	622
30	Call	称为	1642	60	Extend	延长	1350	90	Really	真的	1177	120	Explanation	解释	835	150	Education	教育	611

When mentioning either. Table 6-13 provides an overview for each of the two portions of posts mentioning the two encyclopaedias, and ranks the most frequently co-occurring words. Instead of commenting on the two encyclopaedias per se, many posts *use* or *cite* the encyclopaedias to comment on certain popular incidents. The left-hand side of the table shows the predominance of the Yúfúlíng (鱼浮灵) incident²¹ for posts mentioning Baidu Baike, whereas the right-hand side shows the predominance of the ‘Wallet’ incident²² for posts mentioning Chinese Wikipedia. The findings mainly demonstrate the topicality of microblogs in responding to the latest events. Baidu Baike was mentioned because of the background information it provided concerning an online rumour that a chemical used to revive dying fish called Yúfúlíng is carcinogenic. Chinese Wikipedia was referred to for its content on piriformis syndrome, a topic around which the ex-Google China Chief Lee Kaifu was ridiculed by many and defended by some for recommending that people not sit with their wallets in their rear pockets. To assess the impact of these events on the datasets, I used computer-assisted clustering analysis and manual keyword search independently, and found about 1,400 posts for the Yúfúlíng incident and about 1,200 posts for the ‘Wallet’ incident. I also identified the words predominantly associated with the incidents (see the “Memo” columns in Table 6-13), which included ‘Shanghai.’ The findings highlight the potential of using how online encyclopaedias are referenced and linked in topical microblog public discussions when authoritative information is needed.

²¹ See People’s Daily’s coverage of this incident: http://paper.people.com.cn/jksb/html/2011-07/21/content_876434.htm

²² See China Economy Net’s coverage of this incident: http://www.ce.cn/xwzx/gnsz/gdxw/201201/04/t20120104_22972476.shtml

Table 6-13

Most frequently-occurring words when mentioning ...

Baidu Baike				Chinese Wikipedia			
	Word	Counts	Memo		Word	Counts	Memo
1	Floating Spirit 浮灵	39951	Yúfúling		Wallet 钱包	2600	Wallet
2	Excessive 超标	22672	Yúfúling		Introduction 介绍	1451	Wallet
3	Baidu 百度	21495			Reference 参考	1317	Wallet
4	Share 分享	18623			Call 称为	1307	Wallet
5	China 中国	10780			Kai-fu Lee 李开复	1297	Wallet
6	Discovery 发现	6948	Yúfúling		Health 健康	1293	Wallet
7	Addition 增加	6583	Yúfúling		Similar 类似	1292	Wallet
8	Life 生命	4428	Yúfúling		Kuso 恶搞	1258	Wallet
9	In vivo 体内	3954	Yúfúling		Cause 引发	1251	Wallet
10	Oxygen level 含氧量	3291	Yúfúling		Scientific name 学名	1239	Wallet
11	Canadian fish 加鱼	2732	Yúfúling		Notebook 笔记本	1238	Wallet
12	Temporarily 暂时	2682	Yúfúling		Sitting 坐姿	1237	Wallet
13	Hypoxia 缺氧	2608	Yúfúling		piriformis piriformis	1232	Wallet
14	Extend 延长	2321	Yúfúling		syndrome syndrome	1228	Wallet
15	Sprinkle over 撒过	2310	Yúfúling		Ridicule 调侃	1224	Wallet
16	Originally 原本	2290	Yúfúling		Sciatic 坐骨神经	1222	Wallet
17	Test 检验	2239	Yúfúling		Syndrome 综合症	1217	Wallet
18	Shanghai 上海	2222	Yúfúling		Face 正视	1184	Wallet
19	Serious harm 严重危害	2217	Yúfúling		Wikipedia 维基百科	1096	
20	Necessary 就要	2181	Yúfúling		Website 网站	768	

Note that some word segmentation errors are present in Table 6-13: The term Yúfúling (鱼浮灵) was segmented into two separate words: fish (Yú 鱼) and floating spirit (fúling 浮灵). Also, the first two characters from the Chinese term for using Yúfúling were wrongly segmented into “Canadian fish” (jiā yú 加鱼). This reveals the challenges in dealing with Chinese texts containing seldom-used terms. At the same time, these can be used to detect new terms.

While the above findings indicate the “hot topics” on which microblogs *cite* the two encyclopaedias, they say little about the comments *about* the encyclopaedias. Thus, the posts about the two incidents are removed, producing the results listed in Table 6-14. Some of the words most commonly associated with Baidu Baike included the verbs “share” and “quote,” the nouns “China,” “entry,” and “Weibo,” and some geographic entities such as the “United States,” “Beijing,”

and “Japan.” With Chinese Wikipedia, the list of words included “website,” “free(dom),” “China,” the “United States,” “wiki,” “Google” (both in Chinese and in English), “a bit,” “Baidu,” and “Weibo.” As will be discussed later, the term “a bit” (yìxià 一下) points to the phenomena where the term “wiki” and other wiki-based encyclopaedias become verbs that describe the action of using online encyclopaedias (e.g. to wiki a bit”).

Table 6-14

Most frequently-appeared words when mentioning Baidu Baide or Chinese Wikipedia (one major incident removed)

Baidu Baike				Chinese Wikipedia			
		Word	Counts			Word	Counts
1		Baidu 百度	2552			Website 网站	768
2		Share 分享	2418			Free(dom) 自由	669
3		China 中国	1784			China 中国	663
4		“a bit” 一下	1135			United States 美国	660
5		Entry 词条	983			Forwarding 转发	608
6		Microblogging 微博	930			Encyclopedia 百科全书	533
7		Ency. 百科	683			Wiki 维基	498
8		One kind 一种	624			Google 谷歌	447
9		World 世界	618			“a bit” 一下	444
10		Quote 引用	587			Baidu 百度	430
11		United States 美国	567			Microblogging 微博	423
12		Forwarding 转发	562			Foundation 基金会	401
13		Company 公司	551			google google	394
14		Search 搜索	536			Event 事件	382
15		Beijing 北京	507			Community 社会	377
16		Support 支持	501			Entry 词条	374
17		Introduction 介绍	494			World 世界	374
18		Activity 活动	485			Knowledge 知识	371
19		Network 网络	453			English 英文	366
20		Website 网站	441			Data 资料	359

The same procedure that was used to extract the major discussion threads and “hot topics” could be deployed again for obtaining high-frequency words listed

in Table 6-14. The words on the left provide clues that lead to a series of prize-winning campaigns initiated by Baidu Baike to promote itself on Weibo, accounting for the high frequency of the term “share”. The words on the right, including the terms “website” and “free(dom),” point to Internet filtering and censorship in China.

The high frequency of the verb “share” for Baidu Baike is the direct result of a 5-year anniversary campaign by Baidu Baike²³ that started in early April 2011. More than 85% of the instances were simply sharing a certain entry article on Sina Weibo, with an explicit note: “shared from @Baidu Baike” [in Chinese]²⁵. Weibo users were incentivized to promote Baidu Baike in this way by a chance to win a prize in addition to gaining points in Baidu Baike. With the help of machine-assisted clustering of the texts, I also identified another set of more than 2600 posts that promoted Hudong Baike and Baidu Baike with mentions of awards. (Like Baidu Baike, Hudong Baike is another commercial user-generated encyclopedia website hosted in mainland China(Liao, 2014)). In fact, more than 1800 posts that came from Hudong Baike constituted more than 30 different prizes from smartphones to chocolate. This phenomenon also indicates certain forms of cross-spherical activities between the platforms of Weibo and those of other user-generated encyclopaedias.

The high frequency of the terms “free(dom)” and “websites” for Chinese Wikipedia, in contrast, is largely the outcome of the filtering and censorship experienced by Weibo users across several websites or topics. Using terms such as “sensitive” (mǐngǎn 敏感), “the Great Firewall” (GFW), “walled” (bèiqiáng 被墙), “climbed over the wall” (fān qiáng 翻墙), I found that many Weibo users share

²³ See the official website <http://baike.baidu.com/cms/s/5years/t.html>

²⁵ Derived from the query results (2072) of the dataset using the SQL statement: “SELECT count(*) FROM _all WHERE (WHAT_ LIKE “%分享自 @百度百科%” OR WHAT_ LIKE “%分享自 百度百科%”);”

their experiences on a range of websites and topics to gauge the boundaries of censorship/filtering.

Table 6-14 lists some of these posts. (Each ID number refers to a particular post; negative numbers signal that the post is collected from the WeiboScope; positive ones from the DiscoverText). Hereafter I refer to a post by using the format of ID₁₂₃ or ID-45.

Table 6-15

Selected posts that mentioned Chinese Wikipedia in relation to censorship/filtering

Category	ID	Main text translated
Sina as censors (blog or Weibo)	4414	• Sina blog deleted my post copied from Wikipedia; is Sina blog more sensitive than Sina Weibo??? The same post survive Weibo.
	15243	• [citing CW] ... after the Wang Lijun incident, the name "Jiang Weiping" was no longer blocked by Baidu or Sina Weibo.
	-18987	• Well, examples such as the "June Fourth" is blocked by Sina and Baidu as sensitive historical events, but Wikipedia dares to tell the truth on historical facts.
	-29805	• Wikipedia has been listed as sensitive words in Sina Weibo [emoticon: pity]
Access Issues	2667	• Can any one open Wikipedia websites? Why after several pages of visit, I saw "this page cannot be displayed". Nothing sensitive really ah.
	-5667	• The ban on Wikipedia is lifted, and the term "mosquito cells" [referring to China's Cultural Revolution] is also removed from the list of sensitive terms. But the recent ex-leader of Chongqing remains inaccessible.
	1303	• On Wikipedia, I visited a relatively sensitive topic, ended up with no access to all pages. Later I entered Baidu's address by accident. As a result, I can go online and search for information. Thank all-powerful "Baidu", my grateful tears of
Views on the issue of access	2609	• Just finished the exam on Marxism, and then you will not let me to visit wiki now?!! You crazy ah! I hate the Great Firewall!!
	2997	• After searching some politically sensitive terms as well as several major events using Wikipedia, it feel really good, and should be closer to the truth. I have a new perspective on the historical events. The same words on Baidu Search I find "Search results are not displayable because they may not comply with relevant laws, regulations and policies." Better use foreign websites, for they may be closer to the truth. My mourning for the innocent people who were persecuted.
	3270	• What triggers the sensitive nerve system as Wikipedia becomes inaccessible? Life is like a farce at times. Can never get the intention of the "superstructure" [high officials]. Better to live in the moment and avoid commenting. Time is the best encyclopaedia, older when we revisit, we'll realize other stories exist.

For instance, one Weibo user questioned why her/his post from Wikipedia to the Sina blog was deleted but survived in Sina Weibo (ID4414) and another user claimed, “Wikipedia has become classified as sensitive words in Sina Weibo” (ID-29805). For others, Wikipedia become one of the tests of censorship, with various access issues to its website and specific pages. In post ID2667, a Weibo user shared her/his experience of being blocked from visiting Wikipedia. Post ID-5667 described another users’ observation that the article on the Chinese Cultural Revolution was unblocked but the article on a Chinese official was not. Post ID1303 even suggested that, her/his experience of being blocked, probably because of several visits to sensitive articles on Wikipedia, could somehow be resolved by visiting Baidu.

Still other users voiced their opinions on their encounter with Internet filtering or censorship when using Wikipedia. Post ID2609 seemed to express a student’s angry response for not being able to access Wikipedia after a Marxism exam. Posts ID 2997 and ID-18987 expressed their belief that Wikipedia was more likely to be true for sensitive topics, whereas post ID3270 expressed a more passive attitude about the futility of commenting.

This way of using high frequency words to identify major discussion threads and “hot topics” can be iterated after removing found sets of posts. The procedure can be useful to identify major topics one by one. We can thus now move to the posts that explicitly mention both Baidu Baike and Chinese Wikipedia.

When mentioning both Baidu Baike and Chinese Wikipedia. To zero in on the posts that explicitly compare Baidu Baike and Chinese Wikipedia, we can now examine the posts that mention both Baidu Baike (BB) and Chinese Wikipedia (CW). Figure 6-11 shows that there are 281 posts mention both. Users’ comparisons are most likely to occur in these posts, which can highlight contrasting experiences and attitudes about the two encyclopaedias.

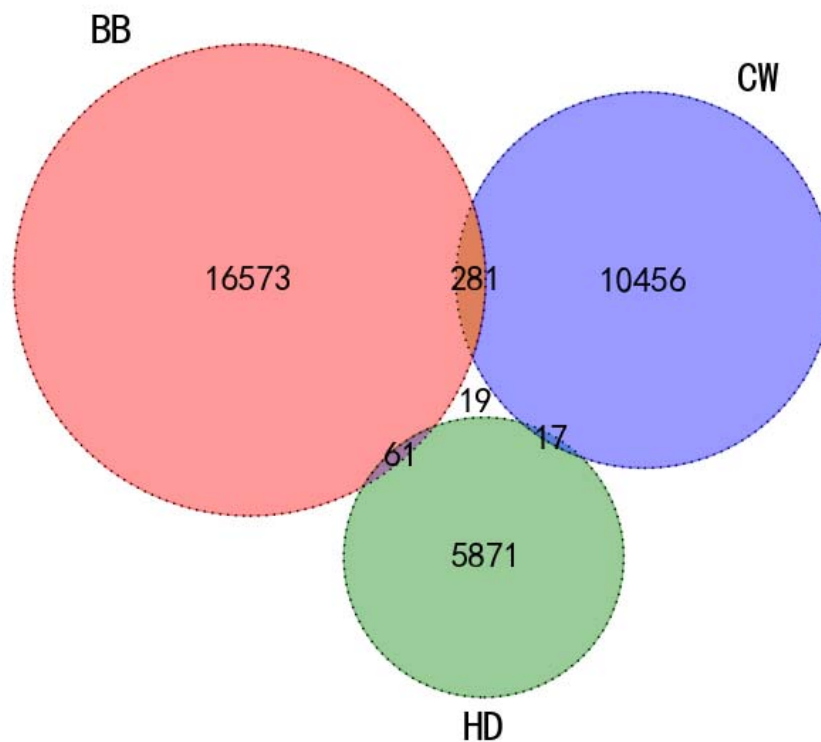


Figure 6-11. Venn diagram of the mentions of the three encyclopaedias

Thus, focusing on the 281 posts that mention both Baidu Baike and Chinese Wikipedia, two set of analysis can be conducted to see the relationship between the users and the censorship/filtering regime. Because it is likely that the censorship/filtering regime influences users, I used Chinese terms such as “wall,” “harmony,” and “river crabs” that refer to the regime to derive the first sample.²⁷ Then I manually coded the remaining posts for the second analysis.

At least 21 posts specifically mentioned the keywords referring to the censorship/filtering regime, of which 15 posts are listed in Table 6-16 (the unlisted

²⁷ The SQL statement is used “SELECT sn,WHAT_ FROM _all WHERE BB=1 AND CW=1 AND (WHAT_ LIKE “%GFW%” OR WHAT_ LIKE “%墙%” OR WHAT_ LIKE “%牆%” OR WHAT_ LIKE “%敏感%” OR WHAT_ LIKE “%和谐%” OR WHAT_ LIKE “%和諧%” OR WHAT_ LIKE “%河蟹%” OR WHAT_ LIKE “%河蟹%” OR WHAT_ LIKE “%上不了%” OR WHAT_ LIKE “%开不了%” OR WHAT_ LIKE “%看不了%” OR WHAT_ LIKE “%看不到%” OR WHAT_ LIKE “%防火%”);”

ones contained irrelevant content). From the top to the bottom, the first category of posts showed the minority opinions of the selected dataset: one claimed that Wikipedia's specific entry on "Socialism with Chinese characteristics" is nonsense when compared to the Baidu Baike's correct version (ID-13491). The other argued that Baidu Baike was enough for people of China, citing the risk of being "Walled" (ID-26194). The post in the second category shared a similar opinion that Baidu Baike is an alternative when Chinese Wikipedia is blocked. Of the majority of the posts that mention both encyclopaedias in relation to filtering/censorship, the third and most representative category consisted of posts arguing that Chinese Wikipedia is better because it is not censored: Wikipedia is "more fair and comprehensive" (ID4388), "the best" (ID-33992), "have much to see" (ID-33991) and can serve as "the standard test" of censorship (ID-40579), whereas Baidu Baike is censored (ID5323) or "harmonized" (ID4388, ID-33991 and ID-40579) or full of "lies" (ID-40570), repetition (ID2790) and errors (ID-3473). Among them, clear anger was also expressed in posts ID2790 and ID-24937 against the Internet censorship/filtering regime.

Table 6-16

Selected posts that mentioned both encyclopaedias (in relation to censorship/filtering)

Category	ID	Main text translated
Baidu Baike is better	-13491	• On "Socialism with Chinese characteristics", I responsibly vouch for the correct version in Baidu Baike, Wikipedia's entry is simply nonsense! How could such an unreliable site have a place in China? It should be thrown over the Wall.
	-26194	• For the people of China, visiting Baidu Baike is enough. Consider the risk of being "Walled" when visiting Wikipedia.
Baidu Baike as alternative	-28952	• Having no choice, Wikipedia is "Walled". Please use Baidu Baike as the foundation.
Chinese Wikipedia is better	4388	• Wikipedia is more fair and comprehensive! I advise everyone not to use that "harmonized" Baidu Baike!
	5323	• Some people died on foreign soils, but remained alive in the country. Some died in Google, but still alive in Baidu. Some died in Weibo, but still alive on the CCTV [Chinese Central TV]. Some died in Wikipedia, but still alive in Baidu Baike. Some are dead outside the Wall, but inside the Wall still alive.
	-5667	• Some people died in a foreign country, but still alive in the country. Some died in Google, but still alive in Baidu. Some died in Weibo, but still alive on the CCTV. Some died in Wikipedia, but still alive in Baidu Baike. Some is dead outside the Wall, but still alive inside.
	1303	• [On certain botany entry..] Opposite plants are treated as the same species in Baidu Baike. What a miracle in this harmonious society ah! While accumulating your wealth, Du-girl [feminized name for Baidu], would you please do something substantial a bit? Access to Wikipedia blocked, Baidu Baike cannot be trusted. No wonder many brain-damaged in this Heavenly Dynasty!
	2609	• Ever since on the other side of the Wall, I become used to using Google and clicking on Wikipedia.. [@other user]:To be honest, I have now subconsciously use Baidu because Google is often off due to some unknown forces. Besides, Baidu Baike, Baidu Zhidao and Baidu Wenku indeed make a great contribution to the diffusion of knowledge to the ignorant Netizens like me. [@yet another user] I choose to wait another ten minutes when Google is not accessible.
	2997	• Baidu Baike not only copied but also harmonized the content from Wikipedia. Much to see in Wikipedia once over the Wall.
	3270	• I always thought Baidu Baike is the best, because it understand Chinese language very well! I was wrong! Completely wrong! In fact, Wikipedia is the best! And truly the best when over the Wall!
	3301	• Better to visit Wikipedia over the wall . . . All lies in Baidu Baike.
	3403	• This "river-crabbed"[meaning censored] entry existed in Baidu Baike long time ago, which tells us nothing about the Great Firewall. Whether we can visit certain entries in Wikipedia is the standard test whether the "river-crab" (censorship) is loosening.

Note that post ID-32406 discussed the decisions users face when they are blocked. One user reflected on the gradual change to the use of Google and Wikipedia that occurred when living outside the Wall, and another comment reflected on the subconscious shift to Baidu because of the Wall, while another made the conscious decision to wait for ten minutes when “Walled”. These comments clearly suggest that not all users of mainland China are oblivious to the existence of Wikipedia and that their use and reception of the websites inside and outside the Wall, while diverse, is more or less conditioned by it. Concerning Baidu Baike and Chinese Wikipedia, users clearly expressed contrasting attitudes to the Great Firewall, which also provides additional user testimony on the impacts of the censorship/filtering regime imposed on the two: Baidu Baike is often censored and Chinese Wikipedia is blocked.

To compare how user opinions differ apart from in relation to the censorship/filtering regime, I examined the remaining posts for users’ evaluative terms about the two websites, and found 39 posts (listed in Table 6-17). Of these, the majority found Chinese Wikipedia to be better. Baidu Baike was reported to be less reliable (e.g. post ID3502, 5733, 5917, 6210, -7323, -7371), less objective (e.g. post ID3570, 6190, -1457, -10079, -13965, -19690, -25687, -34458) and prone to the Chinese government’s political influence or even censorship (e.g. postID5587, -19547, -19690, -21512, -25182, -25221). In contrast, Chinese Wikipedia was perceived to be more useful (e.g. post ID6306), better referenced (e.g. post ID5733, -3755) and to have an elegant interface and style (e.g. post ID3664). Nonetheless, some posts exist stated Baidu Baike was better. Post ID-14431, for example, claims that Baidu Baike is more detailed on R&B (rhythm and blues music). Post ID-26631 suggested that Baidu Baike was richer in content because it could “freely” copy content from other websites, and thus had more comprehensive content that suits the needs of the users in China. Post ID-19336 seems to be ironic by agreeing that Baidu Baike is thus “freerer” (I am not certain whether it is meant to be ironic).

Table 6-17

Selected posts that mentioned both (apart from those concerning censorship/filtering)

Category	ID	Main text translated
Baidu	-14431	• On R&B, Baidu Baike is more detailed than Wikipedia.
Baike is	-19336	• It turns out Baidu Baike is freerer than Baidu Baike [Haha]
better	-26631	• Users of China do not like Wikipedia. Why? For users who only read without editing, they do not care whether the content is copied or original. All they want is plenty and complete information. Because Wikipedia cannot copy but Baidu can copy Wikipedia and others, Baidu Baike's content is richer than Wikipedia's.
	-27322	• You are welcome to read the essay penned by "Between Men&God", an administrator of Chinese Wikipedia, published in Douban.com: "Why is (Chinese) Wikipedia encyclopaedia worse than Baidu Baike?" (Chinese Wikipedia indeed needs constant self-inspections. Wiki-girl approves this article [emotion:greivance])
	16123	• Chinese Wikipedia recently did something stupid. Competing with Vietnamese Wikipedia on the number of entries. It appears that, outnumbered by Baidu Baike and Hudong Baike, Chinese Wikipedia starts to enjoy itself by comparing with minor languages. After Chinese Wikipedia criticized its competitors' value of quantity over quality, a rather ironic development that it enters the number game.
Both good	-5667	• Occasionally Baidu Baike is useful and very comprehensive. Wikipedia is better when looking up English words.
Chinese	1303	• [On "Great Chinese Famine"] Better read more detailed Wikipedia. ... Just also checked Baidu Baike's, which is almost consistent with Wikipedia's. It is a rare situation where Baidu Baike shows some progress.
Wikipedia	2609	• [On Zhang Zhixin] Huge difference between Baidu Baike and Chinese Wikipedia. Which lies?
is better	2997	• The difference between Wikipedia and Baidu Baike? Can tell from one simple entry on Fang Zhouzi..
	3270	• Stronger than Baidu Baike: Wikipedia
	3301	• My foreign teacher asked us to consult Wikipedia and share found information in class. Finally realized that Baidu Baike is the epitome of the new online "Sakoku" (isolationism). Although having less entries, Chinese Wikipedia has absolute essentials, totally unlike Baidu Baike which copies and pastes stuff everywhere. Better learn English well to enjoy the fun reading English Wikipedia!
	3403	• Wikipedia better than Baidu Baike. Checked the history of Haifeng, Wikipedia has the same info as the county government website, but Baidu Baike is way off!
	3502	• Baidu Baike's content unreliable. Contributors often promote companies, books, and so on. The entries on the Internet of things, Smart Industries, Industry clusters, etc. mislead readers. I checked; none is comparable to English Wikipedia.
	3570	• For "Wukan incident", now we can visit Wikipedia's description, which is objective, very detailed, encyclopaedic, and thus much better than Baidu Baike's
	3664	• Wow the Web interface is elegant and language is stylish in Wikipedia. No more Baidu Baike for me from now on.
	4021	• Our history class assignment has the Red Guards as its theme, a bit heavy and difficult. Everyone is advised to take a look at Wikipedia's and Baidu Baike's entry.

(Table 6-17 continues)

(Table 6-17 continued)

Category	ID	Main text translated
Chinese Wikipedia is better	5587	• Very fun contrasting Wikipedia and Baidu Baike. Cannot open Wikipedia's page because it contains the phrases such as "... regarded as a dissident for her opposition to Mao's personality cult and extreme leftism". Can open it in Baidu Baike, because its description goes like this. ...
	5733	• Baidu Baike not reliable. It appears to have complete information but all of them unreferenced or self-referenced... Better use Wikipedia. At very least, each Wikipedia's entry has clear references, respecting original authors' copyright.
	5917	• After reading the linked article, I search for information on "Pan Jinfu". With the help of @hischild, I found this article. Baidu Baike's entry proved to be "watery"... Wikipedia is more reliable. Of course, what is the most reliable is our habits to look for original sources, which can help understand the facts.
	6190	• The description on Milton Friedman in Baidu Baike, in contrast to Chinese Wikipedia, removed the parts where he was challenged and questioned by others.
	6210	• Scholars with expertise have conducted research and found the precision of Wikipedia is higher than encyclopaedia Britannica, and thus can be treated as reliable sources. Noted: Do not ever use Baidu Baike and Hudong Baike, where false information is distributed, otherwise you will hurt yourself and others.
	6306	• Now I realized that Wikipedia is more useful than Baidu Baike.
	-1457	• Better use Baidu Baike less. Like Baidu search, it is castrated and inflicted by viruses. ...Wikipedia has verification and coordination mechanisms and Baidu Baike has not. With a political leaning, Baidu Baike is wildly unfair and chaotic.
	-3755	• Wikipedia first demands info sources, second requires footnotes in scholarly formats, third prohibits [editors'] original research, and fourth says no to vandals and ads. If Baidu Baike can do any of the above, it is no longer Baidu Baike
	-7323	• [Comments on the errors in Baidu Baike's prazepam entry...] Everyone please use Wikipedia when consulting entries in professional fields such as medicine, architecture, law, etc. Baidu Baike is far behind Wikipedia in terms of professionalism and precision.
	-7371	• [On making a mistake regarding the age of an actor.] I recognized my mistakes after googling. ... I realized this: Wikipedia is indeed better than Baidu Baike. It is not about the word count, but about correct information
	-7622	• Fortunately Wikipedia is not wrong on this. When can Baidu Baike correct itself to avoid repeating baseless assertions?
	-10079	• Today I bumped into a new term "neoliberalism". Not knowing what it is, I looked it up in Baidu Baike and Wikipedia. After comparison, the orientation of the two sites is clearly different. The former contains largely negative comments, with emphasis on the opposition of socialism and capitalism... The latter does not contain comments with ideological orientations, except for explaining all knowledge backgrounds. ... From now on wiki only Baidu you garbage.

(Table 6-17 continues)

(Table 6-17 continued)

Category	ID	Main text translated
Chinese Wikipedia is better	-13965	• Encyclopaedias are supposed to document facts. Have a look at the content in Wikipedia and Baidu Baike on TVXQ (a South Korean pop group) and its five members. I seldom visit Baidu, and today at a casual look I found it contains weird stuff in it. Speechless. An encyclopaedia is used as a fan site, with all its biased comments. Only Baidu is brain-damaged enough to allow such content. No chance for such content to survive in Wikipedia's standard.
	-16600	• In practice, Wikipedia is proved to be more cultured than Baidu Baike.
	-19547	• Wikipedia is better than Baidu Baike. Baidu is a dog for the government.
	-19690	• On many topics, the results out of using Wikipedia is more objective than Baidu Baike.
	-21512	• Both as encyclopaedias, why such a big different between the content from Baidu Baike and that from Wikipedia? ...[sh!] [emoticon:depressed]
	-25182	• Baidu Baike does not have and Chinese Wikipedia has it. Immediately I know why.
	-25221	• No such paragraph in Baidu Baike. So much content is different from Wikipedia.
	-25470	• Ha ha ha ha ha Said the teacher. Wikipedia is more authoritative than Baidu Baike!
	-25687	• Consult Wikipedia and Baidu Baike respectively on the topic of the "White Sea - Baltic Canal". Contrast and see the obvious evil of Baidu.
	-26296	• [Facts about the (morning) after pills] Wikipedia shows that Levonorgestrel is linked with ectopic pregnancy. Such a risk not mentioned in Hudong Baike and Baidu Baike.
	-34458	• A question that I have not yet the answer to: On the same topic, the content of Baidu Baike could be so different from that of Baidu Baike. Wikipedia is obviously more objective; Baidu Baike is frequently doped with large amount of false propaganda. Both are interactive encyclopaedias and everyone can participate. Why such a big gap?

Still, these comments do not contradict the observation that overall, Chinese Wikipedia values content quality and respects copyright more (see also Chapter 4). In fact, the essay mentioned by post ID-237322 is originally titled: "What kind of improvements does (Chinese) Wikipedia need when compared with Baidu Baike." In the essay mentioned, before summarizing his opinions that Chinese Wikipedia's content is too limited, the page editor too complex to use, and it is too unfriendly for newcomers, the writer (an administrator of Chinese Wikipedia) acknowledged that much has already been said about where Chinese Wikipedia is better.

Altogether, these posts represent a sample of opinions voiced in both Weibo and Twitter, suggesting that Chinese Wikipedia is perceived as more reliable and less censored than Baidu Baike is. In particular, posts ID3301 and ID-25470 use teachers' opinions, and the former shares her/his class experience in using Wikipedia as that "Baidu Baike is the epitome of the new online 'Sakoku'" (isolationism). The two posts are amongst many posts that mentioned Baidu Baike or Chinese Wikipedia as one of the important digital tools for improving one's understanding of the world. The idea that online encyclopaedias are important tools for gaining information is further supported by the fact that the Chinese term "wiki" has become a verb similar to the verb "to google".

When "wiki" becomes a verb. Indeed, in the dataset, like "google" in English, "Baidu" and the Chinese term "wiki" (wéijī 维基) are used as verbs that refer to visiting these encyclopaedias for information. I found nearly 200 posts for each of the Chinese-language phrases "Baidu a bit" (bǎidù yíxià 百度一下) and "Wiki a bit" (wéijī yíxià 维基一下). To find out what kinds of "verbification" are expressed, I categorize these posts according to the preceding words because they provide contexts for such verbification. Table 6-18 lists the categorized results. The four categories in Table 6-18 distinguish between when an action is done, will be done, can be done, or should be done. More than 20 posts refer to the posters' past actions (the category "Action done" in the table) in using one of the encyclopaedias for information, and there are more such posts for Chinese Wikipedia: e.g. "... just recently wikied a bit". This suggests that some users share their experiences using online encyclopaedias on microblogs. In addition, the categories "Action possible" and "Action better required" refer to the cases when users make the suggestion that online encyclopaedias should be consulted.

Table 6-18

Baidu/Wiki a bit

Category	"Baidu a bit" (百度一下)		"Wiki a bit" (维基一下)	
	Precedent	Word context	Precedent	Word context
Action done	刚(2)	...just (recently) Baidu a bit...	刚(11)	...just (recently) Wiki a bit...
			刚(3)	...just (recently) Wiki a bit...
			刚/刚才(2)	...just (recently) Wiki a bit...
	于是(4)	...then Baidu a bit...	顺手(2)	...casually Wiki a bit...
	特意(3)	...purposely Baidu a bit...	重新(2)	...again Wiki a bit...
			再(2)	...again Wiki a bit...
Action	去(17)	...go Baidu a bit...	去(10)	...go Wiki a bit...
Action possible			可(10)	...can Wiki a bit...
			能(10)	...can Wiki a bit...
	可以(8)	...can Baidu a bit...	可以(3)	...can Wiki a bit...
Action better required	先(6)	...first Baidu a bit...	不会(4)	...why not Wiki a bit?...
	要(2)	...must Baidu a bit...	为啥不(2)	...why not Wiki a bit?...
	只要(2)	...simply Baidu a bit...	最好是(1)	...better Wiki a bit...
	请(2)	...please Baidu a bit...		
Pronoun and noun	大家(4)	...everyone Baidu a bit...	大家(2)	...everyone Wiki a bit...
	我(3)	...I Baidu a bit...		
	你(3)	...you Baidu a bit...	你(3)	...you Wiki a bit...
	同学(2)	...classmate, Baidu a bit...	@(5)	...@[a Weibo and Twitter user]...
Promotion	都(23)	...all Baidu a bit...		
Action combined			百度(8)	...Baidu/Wiki a bit...
			或者(6)	...or Wiki a bit...
			和(3)	...Baidu and Wiki a bit...

There are eight posts that use the combined verb “Baidu/Wiki a bit” (bǎidù wéijī yíxià 百度维基一下), with an example post quoted below (ID-31873):

The masses should take advantage of this opportunity to popularize this universal knowledge ah! If one does not believe in experts, one can Baidu/Wiki a bit.

One instance that distinguishes Baidu Baike and Chinese Wikipedia is when the verb “Baidu a bit” is used by Baidu to promote itself in Weibo. At least 23 posts, most likely generated by some Weibo users who participated in promotion campaigns, followed the following format:

Got news? Brush [meaning “refresh” or “update” in English] Weibo. Got new knowledge? Brush Weibo too. Because what matters most for communication is brush something anew. It is most foreign flavour (meaning “exotically fashionable” in English) to all Baidu a bit on everything. What you do not know, @BaiduZhidaol let you know. I am @UserID [Weibo user] at Weibo.com. Got something to ask? Don't ask me. Ask @Baidu Baike.

What the above post demonstrates is exactly the crossover of communicative and information spaces, where the information spaces of Baidu Baike and Baidu Zhidao [a Baidu user-generated question-and-answer website,] can be used to update new information in the communicative spaces of Weibo. Although such cross-over or cross-spherical activities may in this case not be completely voluntary, there is ample evidence that Weibo or Twitter users share their experiences in using search engines, user-generated encyclopaedias and microblogs across platforms. Some of them have also noticed the relationship between the SERPs and UGEs:

...As matter of fact, you will know by simply Baidu a bit. Key in three-character “Yinbingshi” to search, the first item should be Baidu Baike's entry on Yinbingshi.... (ID-33642)

Google is sensitive....search “lie group” one can find Lie group-Wikipedia, the free encyclopaedia as the top result. (ID-23435)

...Simply Baidu a bit, Baidu Baike has a specific entry on him.... (ID-34850)

...Amazing that through Baidu I found the “sensitive”- “men” [i.e. Tiananmen] incident twenty years ago on a sensitive day and sensitive month in a sensitive year. The first SERP item is Wikipedia's entry page on this incident, and it can be opened. (ID-42437)

Thus in general, what is shared between “Baidu a bit” and “Wiki a bit” as a verb is the specific use of Baidu Baike and Chinese Wikipedia: looking for

information. I found no cases where the verb is used to refer to writing or contributing content to them.

Thus, in the Chinese-language context, as shown by the text analysed here, when used as verbs, the terms “Baidu a bit” and “Wiki a bit” refer to general information-seeking actions, not to the original meaning of wiki for collaborative writing. The posts clearly indicate the prominence of using Wikipedia and/or Baidu Baike as information practices on par with Google or Baidu.

6.2.3 Discussion. Ample evidence exists to show the existence of cross-platform activities. Weibo and Twitter users are often aware of both encyclopaedias. Here we can also see the porous nature of web spheres. Research must take such cross-platform activities into account, as suggested by some digital method researchers (R. Rogers, 2013). Such cross-platform activities also include promotion campaigns initiated by Baidu Baike and Hudong Baike, and even cyber-attacks against Wen Yunchao’s Twitter account by flooding tweets using Baidu Baike’s entry, as also testified by Wen Yunchao himself at a US Senate hearing:

....For about a year starting April, 2011, unidentified persons “tweet bombed” me on Twitter with trash information. Using a software [program] called Tween to filter the trash, I found the heaviest attack took place on April 25, 2012 – with a staggering 590,000 spam posts within 24 hours. Unidentified persons also posted viciously defaming information about me online at the rate of over 10,000 times per day. As far as I know, artist Ai Weiwei has been similarly attacked

Thus, the Twitter data collected for this research (from March 20, 2012 to April 12, 2012) contains the early phase of the spam post attack against Wen Yunchao on Twitter. Encyclopaedia pages’ titles and links, mostly from Baidu Baike, can be used to jam a Chinese activist’s account in Twitter that is outside Beijing’s filtering and censorship regime. This is evidence of information engagement in the sense of information warfare. The availability of user-generated

encyclopaedia data is repurposed to jam a microblog account by introducing noise to keep certain Chinese activists from engaging their intended audience.

Also, the systematic method of first clustering posts (by computers programs) and then categorizing them (by analysts) one by one through identifying and then removing the hot topics should also help detect other spamming or en-masse promoting activities, as demonstrated by the self-promotion campaign initiated by Baidu Baike and Hudong Baike. Although the approach of the study may be questioned on the grounds of accuracy and validity, it has several merits that traditional methods such as user interviews cannot rival. First, this approach is feasible and reproducible enough for constant monitoring. Dynamic reports could be quickly generated with a workflow consisting of computer programs and social media analysts. Second, this approach may be better to anticipate or even predict the hot topics of public discussions that use or comment on user-generated encyclopaedias. Third, the accuracy and validity of findings can be further enhanced through better programs and more experienced analysts after several iterations.

The use and reception of the two major user-generated encyclopaedias on Weibo and Twitter yielded several findings. First, it confirmed that some Weibo and Twitter users use Chinese Wikipedia and/or Baidu Baike, and that Beijing's filtering and censorship regime has different effects on these uses, thereby contributing to the existing research on online Chinese-language encyclopaedias (Liao, 2009b; Luo Z.-C. & Fu, 2008; F. Ma & Xia, 2008). Several users were also aware of the differences between the two encyclopaedias. In fact, several posts suggest the contrast between Baidu Baike and Chinese Wikipedia as a test for tracking what counts as a sensitive topic. While users did sometimes get frustrated when they were denied access to Chinese Wikipedia, many of them preferred Chinese Wikipedia precisely because it was not censored. Second, overall, users had positive experiences using online encyclopaedias, and more so in using Chinese Wikipedia than in using Baidu Baike. For example, Chinese Wikipedia was

perceived to value content quality and respect copyright more, in contrast to Baidu Baike, which was known for copying and pasting content from other websites. Explicit comparisons were also made by Weibo and Twitter users in short posts, and more users perceived Chinese Wikipedia to be of higher quality. Third, like the new English word “google” being used as verb, the Chinese terms “Baidu” and “Wiki” are also used as verbs to describe looking up content in Baidu Baike or Wikipedia. This suggests that the use of online user-generated encyclopaedias have become an important part of users’ online experience, including when users cite information to discuss topical issues, request that others find information, or verify information seen elsewhere. For these users, user-generated encyclopaedias were resources they relied on.

Moreover, the findings show that many users view the two resources as an established part of everyday information activities, like using the search engines Google and Baidu. These activities are often linked with information or digital literacy, as exemplified by comments from students using Wikipedia in classroom settings. Here we can see the expectation that user-generated encyclopaedias lead to the betterment of self and society through gaining knowledge, partly expressed by a post suggesting that “[t]he masses should take advantage of this opportunity to popularize this universal knowledge ah! If one does not believe in experts, one can Baidu/Wiki a bit” (ID-31873). The new experiences with information and communication spaces are increasingly associated with these websites, as are cultural and social norms about information access, knowledge sharing, and content filtering and censorship. The online practices of information engagement are thus expected to be influenced by users’ experiences with different platforms, which are of cultural-political significance for the Chinese-language Internet.

The findings also have implications for youth engagement in mainland China since Weibo users in mainland China are more likely to be younger. More than 50% of Sina Weibo users were 26-35 years old according to an industry report in 2012 (DCCI, 2012) and more than 80% of the visiting traffic came from 10-39 year

old users according to an annual report released by Chinese Academy of Social Sciences in 2014 (Tang, 2014). In light of the rough user profile, it can be argued that state and market players are expected to engage the younger generation in mainland China through microblog platforms, which may explain why the information warfare kind of information engagement is observed in the case of jamming microblog accounts of Chinese activists such as Wen Yunchao. It may also explain why the learning aspect of citing, using, and sharing experiences about user-generated encyclopaedias is also important for several microblog users.

In terms of information engagement, this study finds that microblog users cite major user-generated encyclopaedias to engage in information-critical or information-intensive public discussions. In a broadened sense of open collaboration, the open discussion on microblog platforms use information resources from user-generated encyclopaedia platforms.

The dynamics of the information engagement concerns the largest non-English segment of the world's Internet population, and this research has examined how it is practised and expressed on two major Chinese-language microblog platforms about two major Chinese-language encyclopaedias. The findings demonstrate the potentials and gaps towards more active and sound collaborative information engagement beyond mere information seeking and consumption. The microblog posts were still predominantly about using encyclopaedias (in a way contributing to better informed users and public discussions), not about contributing back to the content of user-generated encyclopaedias.

Still, both Baidu Baike and Chinese Wikipedia enjoy mentions from Twitter and Weibo users, and both have become verbs in the Chinese language just like the new English verb "google." The Chinese verbification of the two encyclopaedias thus indicates what has been achieved and what has yet to be achieved. More research and practice is needed to advance the understanding of the potentials and the challenges to improve information engagement on both platforms, possibly using microblog discussion to improve encyclopaedia platforms and their content

as well as other open collaboration activities that are more “egalitarian,” “meritocratic” and “self-organizing” (Riehle, 2012).

With regards to the recent efforts by Chinese Wikipedians to engage users on Sina Weibo (A. Wang, 2013), the findings point to a direction in which users engage in public discussions on microblog using and enhancing encyclopaedia articles. The research methods here can be repurposed and streamlined to build a public opinion monitoring system. First, robust jam- or spam-detection software can be developed and implemented to highlight the abuse and/or misuse of user-generated encyclopaedias. In addition, it might be possible to show (or even predict) which incidents may become hot discussion topics that require quality and authoritative information sources that may or may not be found in major user-generated encyclopaedias. Future work can be conducted to explain, predict, or even plan interventions to the information-related activities that can be publicly observed across the microblog platforms and user-generated encyclopaedia platforms. Overall, the development of certain event monitoring devices, information engagement rules and new encyclopaedia articles may result in better public discussion and better human knowledge. This may in turn translate into more kinds of civic information engagement and less information warfare.

A division of users and user experiences currently exists in mainland China characterized by the choice between Baidu Baike and Chinese Wikipedia. While the frequent occurrences of the verbs such as “Google a bit”, “Baidu a bit”, and “Wiki a bit” suggest that user-generated encyclopaedias are part of users’ overall information-seeking activities, the choice between Baidu Baike and Chinese Wikipedia entails different norms and attitudes. Baidu Baike was often associated with the Beijing-censored version of information sources and Beijing-sanctioned ways of information engagement. On the other hand, Chinese Wikipedia is not censored and thus was perceived by some to be more reliable and was associated with the freedom of information. Blocked access to Chinese Wikipedia was also perceived as signals of Beijing’s political “sensitivity” towards information.

The perspective of learning to find and cite information in order to participate in public discussions shows the potential of social media platforms as “civic learning repertoires” for “information engagement” (Bennett & Wells, 2009). For users in mainland China, at least two major learning strategies can be identified: one takes the censorship and filtering regime for granted, and the other seeks information and communication platforms that are outside the influence of the regime. It is likely that the “civic learning repertoires” for “information engagement” are limited to the bounds of the regime (i.e. mainland China) for those who use Baidu and Baidu Baike. Indeed, both Baidu Baike and Chinese Wikipedia become important reference points and even everyday communicative practices (e.g. “Baidu a bit” or “Wiki a bit”). Still, the different uses and receptions suggest two set of learning experiences and information choices. For contributors and readers, the choice of one over the other reflects the different experiences in civic learning made possible by the two websites. Since both platforms of microblogs and encyclopaedias depend on users and their contributed content, users learn to use and choose platforms while platforms learn from users. Future research and practice are needed, with better information engagement and content, to advance the quality of public discussions and public knowledge.

6.3 Chapter conclusions

To my knowledge, no research to date has systematically examined how variants of Chinese-language scripts provide different user experiences on specific online platforms. Such knowledge provides insights on the dynamics of the Chinese cultural sphere online, which represents the largest non-English segment of the world’s Internet population. This chapter has begun such an investigation by looking at how users view the two major Chinese-language encyclopaedias, as observed from major search engines and microblog sites. The findings not only complete the third aspect of collaborative websites - of participation, content, and readership - but also provide insights into the Chinese-language Internet generally.

Table 6-19 contrasts the use and reception of Baidu Baike and Chinese Wikipedia. First, both were visible for many of the sample queries. If this continues to be the case, we can infer that Chinese-language Internet users who use search engines will regularly will encounter at least one of the encyclopaedias. The centrality of the two encyclopaedia websites for the Chinese-language Internet is thus established by evidence gathered from major search engines and microblog platforms. The blockmodelling analysis further shows the choice of search engines matters as to which encyclopaedia website will be more visible. In most instances, Google users across regions will find Chinese Wikipedia more visible whereas Baidu or Yahoo China users will find Baidu Baike more visible. The findings also show strong differences in the geolinguistic features of the SERPs. Again, Baidu Baike is most visible in the SERPs that are overwhelmingly mainland and contain simplified Chinese, whereas Chinese Wikipedia is most visible in the SERPs that are geolinguistically mixed. Second, both enjoy mentions from Twitter and Weibo and both have become verbs in the Chinese language just like the new English verb “google”. Further analysis of the comments suggests that different ideas are associated with the two websites. Baidu Baike is associated with the Beijing-censored version of information sources. In contrast to Chinese Wikipedia, Baidu Baike is more accessible because it is not blocked as often as Chinese Wikipedia. On the other hand, Chinese Wikipedia is not censored and thus more reliable, and is associated with the freedom of information and its access being blocked by Beijing.

Table 6-19

Comparing use and reception patterns

Patterns	Baidu Baike	Chinese Wikipedia
... of visibility	Visible site for many queries	Visible site for many queries
... of co-visible variants	Mostly Baidu_CN and Yahoo_CN	Mostly Google, TW and HK variants
... of geo-linguistic features	Overwhelmingly mainland and simplified Chinese	Mixed
... of co-visible websites	Mostly sites with mainland Chinese focus	Mostly sites censored or blocked by Beijing
... of mentions	Both Twitter and Weibo	Both Twitter and Weibo
... of thinking	Likely censored	Uncensored (and reliable)
... of discourse	Easy access (not blocked)	Freedom of information
... of practice	"Baidu a bit"	"Wiki a bit"

Overall, the findings suggest that for users going from one website to another, different platforms generally constitute no barriers. Across major Chinese-language search engine platforms, users will notice and likely click on major user-generated encyclopaedias because their high visibility in the SERPs. Weibo and Twitter Chinese-language users mention Baidu Baike and Chinese Wikipedia, sometimes using them as verbs. Web spheres are indeed porous. Nonetheless, there is some distance between them in terms information access and cultural exchange (even if these are not impenetrable barriers). This distance exists due to the cultural-political factors, including geolinguistic differences, across Chinese-speaking regions and due to Beijing's filtering and censorship regime. Chinese Wikipedia's visibility across search engine platforms differs significantly from one set of geolinguistic regions to another. It is much less visible in Google_CN and Yahoo_CN. The pattern is largely reversed for Baidu Baike, but Baidu Baike is still visible for users in Hong Kong and Taiwan if they use Google rather than Yahoo by default. When considered as a whole, the findings from the

analysis of SERPs and microblogs show that the cultural-political boundaries of web spheres are not absolute but porous. However, this does not prevent web spheres from imposing distancing mechanisms by amplifying or diminishing information from different sources. Cultural-political boundaries are thus likely to be reproduced not only by the Internet filtering and censorship regime, as conventionally believed, but also by shaping information suggestions and choices.

The findings add to the cross-spherical analysis of web spaces proposed by Richard Rogers(2013). Because of the searching, citing and commenting activities, the web spaces are indeed porous between Chinese-language search engine and encyclopaedia platforms, and also between Chinese-language microblog and encyclopaedia platforms. On the other hand, the distinct choice between Baidu Baike and Chinese Wikipedia suggests not only different information engagement practices (including the preferred choice of platforms) because of the censorship/filtering regime, but also different civic learning repertoires. The censorship/filtering regime clearly has impacts on shaping web spaces and their information engagement activities.

The findings also update our understanding of the online “Chinese cultural sphere” of which Chinese-language resources are important (G. Yang, 2003). Chinese-language platforms such as search engines, microblogs and user-generated encyclopaedias constitute major instances where such resources can be indexed, searched, retrieved and viewed amongst mostly Chinese-language users. Such openly available resources permit potentially new information engagement and even open collaboration. However, such potentials may be limited by Internet censorship (e.g. Baidu Baike) or filtering (e.g. Chinese Wikipedia), shaping the engagement dynamics with different sets of norms and knowledge sources. Thus, when users become familiar with certain platforms hosted in a specific location and used by a certain groups of fellow users, this amounts to both a learning process and a cultural thickening process.

The cultural thickening perspective highlights these differences. The findings in the previous chapters and in this chapter have shown that Baidu Baike is not widely read or used by users in Hong Kong and Taiwan, partly because Baidu Baike does not offer a traditional Chinese reading and editing platform, as well as Baidu Baike's being subject to Beijing's filtering and censorship regime. The SERP data also shows how the choice of regional settings and language scripts shape which encyclopaedia website is likely to be seen (and thus clicked). As different online sources have been shown to be cited/linked in Chapter 5, different online sources of websites are now also shown in the case of the two encyclopaedias in the SERPs. The geolinguistic dimension of cultural thickening for the two encyclopaedias shows how Baidu Baike has produced the patterns mostly within mainland China, whereas Chinese Wikipedia is used and well-received by users from Taiwan, Hong Kong, and mainland China, if one can assume Weibo users who experienced the Great Firewall are largely from mainland China. Hence, Baidu Baike and Chinese Wikipedia present different content and scope for Chinese-language cultural resources online and, as they are processed and thus shared through other platforms like search engines and microblogs, this indicates two cultural thickening patterns distinct in their geolinguistic reach and cultural-political content.

Finally, the basic idea of promoting knowledge for the betterment of individuals and societies is linked to the use of online user-generated encyclopaedias. Here, the choice between Baidu Baike and Chinese Wikipedia, when considering the content and scope of civic learning and cultural thickening, has cultural-political consequences. The choice is of particular importance to mainland Chinese users, especially the new majority of Internet users (see Chapter 4). Future research may also consider how the concept of information engagement intersect with the official and popular discourse of the Chinese term "*suzhi*" (sùzhì 素质) of people, a discourse of human quality that demands individual competition, responsibility and self-improvement—including the ability to use technology to do

so. In the microblog dataset, several posts discussing the lack of “quality” information and people can be further analysed. Since the notion of information engagement is relatively unknown to the general public in China, the notion of *suzhi* can be used as potential bridging concept to discuss which kind of information engagement is better in promoting knowledge for the betterment of Chinese individuals and societies. Some questions can also be formulated based on the existing Chinese studies literature on “*suzhi*” for online platforms. For instance, *suzhi* discourse may have aimed to turn peasants into modern Chinese citizens (Murphy, 2004), but how does it work, say, to turn migrant workers into modern online Chinese citizens? If *suzhi* discourse justified social hierarchies (Kipnis, 2006), how is it related to the choices of information platforms and the practices of information engagement? More empirical research is thus needed to examine and compare across different information platforms and groups of users.

In this cultural-political context, the findings from the SERP and microblog datasets show important links between processability and information engagement. The concept of processability is taken in this thesis as an engine that mediates online information and communication activities and provides the enabling conditions of current media-language systems. From a technical perspective, how SERP and microblogs function as communicative and information spaces require machines to be equipped to process large amounts of information. From the perspective of users, these online platforms require users’ contributions and require users who are equipped with adequate skills and the necessary cultural contexts for participation. Search engine companies compete to provide information free for different groups of users, and in return, users’ information activities are processed so that companies provide the SERPs that users find relevant to keep them using their services. The same is true for microblogs. Data collected from both Weibo and Twitter shows how users learn and put into practice the information and communication skills that shape and are shaped by the cultural politics of the Chinese-language Internet. The fact that these platforms

dependent on user-generated activities and content thus demands a learning process on both sides whereby platforms learn from users and users learn to use and choose platforms.

Since knowledge has normative dimensions (the betterment of individuals and societies), the difference in cultural thickening through knowledge also has its normative implications. Especially for users in mainland China, the windfall of Internet population growth from 2005-2008 and the Internet filtering of Wikipedia from 2004-2008 suggest different learning experiences in adopting online user-generated encyclopaedias (see Chapter 4). Indeed, this chapter has shown a division of users and user experiences in mainland China, drawing from both SERP and microblog data. The choice between Baidu Baike and Chinese Wikipedia presents a choice similar to the one between Baidu and Google, or more generally the one between censored or uncensored. Hence for mainland Chinese users, while the frequent occurrences of the verbs such as “Google a bit”, “Baidu a bit”, and “Wiki a bit” suggests that user-generated encyclopaedias are part of users’ overall information-seeking activities, the choice between Baidu Baike and Chinese Wikipedia entails different norms and attitudes. The two user-generated encyclopaedias provide instances of different cultural thickening patterns, and the norms that go with them.

Chapter 7 Conclusion

This thesis has explored whether and how Internet connectivity can reshape geo-linguistic and geo-cultural boundaries in relation to nation-states based on a comparative study: How have the two major Chinese-written user-generated encyclopaedias, Baidu Baike and Chinese Wikipedia, overcome, reinforced, or ignored the existing boundaries within and beyond the Chinese-speaking world? These two encyclopaedia websites have been analysed in terms of the ‘thickening that occur[s] within an increasingly global connectivity’ (Couldry & Hepp, 2013, p. 255) and cultural patterns were identified: one reinforces the existing political boundaries (largely Baidu Baike, within mainland China) whereas the other overcomes boundaries by integrating them (largely Chinese Wikipedia, creating patterns across national boundaries). The two websites thus exemplify two different patterns of content-creation and readership of Chinese-language websites.

The first section of this chapter will summarize the findings for the field of Chinese Internet research, and the following section will discuss the implications for research more broadly and for policy. The final section will summarize the main concepts and how they bear on the geolinguistic dynamics of online knowledge.

7.1 Understanding the Chinese-language Internet

Briefly, Baidu Baike, as might be expected of a true Chinese ‘national’ website, did not overcome the communicative barriers between mainland China, Taiwan, and Hong Kong. Chinese Wikipedia, on the other hand, achieved some success in overcoming these barriers: its success consisted partly of enabling users from mainland China to make bridges beyond national borders and the mainland’s censorship/filtering regime in seeking information within a wider Chinese-speaking region.

This thesis is the first to map out these cross-boundary interactions of the Chinese-language Internet by means of combining a variety of datasets. While

Internet connections are expected to foster extensive cross-boundary interactions, they have also been found here to be subject to various cultural-political barriers among Chinese-speaking societies. These include not just the censorship/filtering regime, but also the choice of a simplified Chinese script versus the traditional one and other cultural-political differences. While this thesis has covered only two types of websites (user-generated encyclopaedias and, to a lesser extent, search engines), this research has analysed what are arguably the most influential websites that organize and gatekeep Chinese-language information and knowledge. By analysing how web spheres order information and knowledge for various groups of users, insights have been gained about the cross-boundary dynamics of Chinese-language regions, including the relationship between indigenization and blockage, different patterns of cultural thickening, and the cultural politics of information.

7.1.1 Ordering information and making Chineseness visible. Perhaps the most significant contribution of this research to Chinese Internet research is that the analysis of Chinese web spheres yields an understanding of the boundaries and norms of media-language systems. By examining the cultural thickening patterns of the two user-generated encyclopaedias, it is possible to identify the dynamics among and within Chinese societies. This approach demands considering the factors of language, geography, and platforms in producing different specific but empirically observable web spheres in order to decipher the boundaries of the media-language systems for particular groups of users. Several Chinese media-language systems were identified, including the major regions of mainland China (dominated mostly by Baidu and Baidu Baike), Taiwan, and Hong Kong (dominated mostly by Google and Chinese Wikipedia). Each region has its own media-language system, with intricate inter-relationships among them. In one sense (of socio-cultural heritage), a web sphere should ‘reflect’ even if it does not ‘represent’ the media-language system preferences of its corresponding regions. In another sense (of information system design), a web

sphere should ‘implement’ even if it does not ‘constitute’ a Chinese information order for its target users in a region. The mutual shaping relationship between information systems and societies remains a challenging theoretical question. Nonetheless, online materials provide strong materials for identifying Web spheres: for example, despite its relatively small population and online presence, Macau clearly has its own web sphere that is constructed by various providers for its users. In particular, because these web spheres yield evidence for how information is ordered, they show how the boundaries of information are overcome, reinforced, or shifted. These spheres are implemented as computer codes, localized interfaces, and meta-data, and produce different results. The concept of Chinese web spheres thus forms the basis for an understanding of the dynamics of Chinese societies on the Internet along with their media-language and political systems.

Comparing and analysing Chinese web spheres is complicated because of various ‘transnational’, ‘national’, and ‘local’ dimensions of ‘Chineseness’ (A. Ong, 1997; G. Wang, 1991, 1993). For instance, does one consider the connections between mainland China and, say, Hong Kong as transnational or national? Acknowledging this complexity, this research has drawn on various sources to identify Chinese web spheres. With data systematically collected from online sources such as encyclopaedia websites, search engines and microblogs, the thesis applied the concept of ‘cultural thickening’ (Hepp & Couldry, 2009; Löfgren, 1997). In addition, the thesis drew on concepts from the global TV studies and Chinese studies literature about the geocultural aspects and geolinguistics of Chinese contexts. The two user-generated encyclopaedias examined were shown to produce two distinct ‘cultural thickening’ patterns across different Chinese web spheres in a way that overcame or reinforced existing boundaries. The overall comparative analysis showed that Baidu Baike overwhelmingly focuses on mainland China, thereby producing a separation (from the rest of the Chinese-speaking world). In contrast, Chinese Wikipedia integrates diverse

sources, users and contributors, effectively producing cultural thickening across various web spheres and across geographical and across platform differences.

7.1.2 Indigenization and blockage. The thesis contributed to the debate about the relationship between Chinese indigenization of the Internet and the Chinese blockage of information flows. Patterns of integration and of separation were observed in comparing Baidu Baike and Chinese Wikipedia. Indeed, both the ‘Chineseness’ issue and the censorship/filtering regime were found to be intertwined with the historical trajectory of the two encyclopaedias (see Chapter 4). Thus, Baidu Baike and Chinese Wikipedia can be seen as important sites to assess the effects of Beijing’s censorship/filtering regime in relation to the indigenization of Chinese information and media systems. Two major views can be contrasted: isolation versus indigenization. The isolation view has been criticized for the lack of strong empirical evidence for the complete isolation of China from the rest of the world, while the indigenization view argues that the censorship/filtering regime has little impact on changing users’ information choices (Suo, 2007; Taneja & Wu, 2013), either because Baidu Baike suits Chinese culture better (Suo, 2007) or because such a censorship/filtering regime does not constitute an absolute barrier inside a so-called Chinese ‘culturally defined market’ (Taneja & Wu, 2013). The indigenization perspective views Beijing’s censorship/filtering regime as a form of cultural protectionism, along with geographical and linguistic factors that influence users’ preferred choice of websites. To put it bluntly, this view believes that filtering blockage is merely a weaker form of cultural protectionism when compared to more potent forces of geographical and linguistic barriers (especially those between Chinese and English). On this view, the regime only slightly adds to cultural protection for Chinese indigenous ‘informatization’, and does not isolate China from the world.

This thesis puts forward a third argument: the information blockage fostered two parallel cultural thickening patterns: Separation effects, or the marginalization of undesired information away from the majority of mainland

Chinese users, are the main outcome of indigenization of online spaces induced by the censorship/filtering regime. Integration effects, on the other hand, are among the main achievements of the web in overcoming the barriers that include Beijing's censorship/filtering regime. This argument allows us to move beyond the isolation and indigenization views towards a picture of distinct cultural thickening patterns. Based on the findings that have been presented, the mainland Chinese web sphere does distance some sources outside mainland China away from mainland Chinese users. While it is possible that some mainland Chinese users can and do use Chinese Wikipedia and Google, and are thereby more likely to be exposed to information sources outside China, these users are shown to be the minority. In addition, the Hong Kong web sphere is closer to the Taiwanese one, because their dominant online information platforms, such as Chinese Wikipedia and Google, mix more diverse Chinese information sources from around the world, including those hosted in the U.S. and mainland China but mostly in the Chinese written language. This argument is more consistent with Yang's 'domestication of the Chinese Internet' (G. Yang, 2012) where the complex interactions with diverse outcomes are produced by Beijing's blocking certain information outside mainland China and censoring information inside. Thus, instead of homogenous or evenly-connected Chinese 'culturally-defined markets', two wider Chinese web spheres are produced by the two parallel cultural thickening communication patterns. In relation to mainland China, Baidu Baike exemplified the internal pattern impacted by the censorship regime, whereas Chinese Wikipedia became marginalized or externalized by the censorship/filtering regime. The following discussion will explain how the studies of different web spheres helped to identify different realities facing various groups of Chinese users, and how the concept of cultural thickening helps to understand these.

7.1.3 Identifying cultural thickening. In identifying cultural thickening processes, this thesis has shown two distinct patterns of Chinese information

marginalization and mobilization based on the differences between Baidu Baike and Chinese Wikipedia: Baidu Baike focuses on the mainland web sphere without too much emphasis on other spheres, whereas Chinese Wikipedia acknowledges different spheres and seeks to integrate them to avoid region-centric bias. Thus, different sources of information are marginalized to the periphery and others are mobilized to the core, effectively reflecting different ways in which knowledge and information are ordered.

The webometric and content analysis further demonstrated these encyclopaedias’ differences in representing ‘Chineseness’ (Chapter 5): results were less revealing about the ‘level’ of Chineseness and more about the geographic and linguistic ‘extent’ of Chineseness (mainland China versus the rest of the Chinese-speaking world), suggesting strong cross-regional thickening patterns for Chinese Wikipedia but not for Baidu Baike. The search engine visibility tests, users’ micro-blog comment analysis, and user interviews together demonstrated how users from different web spheres might encounter different types of information. The analysis also compared users’ experiences with information providers such as Baidu, Wikipedia and Google showing the effects of information marginalization. The overall results amount to two clear cultural thickening patterns, as summarized in Table 7-1.

Table 7-1

Different gatekeeping lead to different cultural thickening patterns

	Baidu Baike	Chinese Wikipedia
is (relatively more) open to	copyright-dubious, self-promoting, or advertisement-based content	reliable information sources across Chinese-speaking regions
is (relatively more) defensive against	topics politically sensitive to Beijing and information sources outside mainland China	copyright-dubious, self-promoting, or advertisement-based content
and thus produces cultural thickening patterns of ...	commercially and politically correct mainland China-focused Chinese information order	culturally and politically diverse Chinese information order across various Chinese societies

The two patterns reflect the cultural-political preferences of users and are governed by their attitudes and perceptions towards various information sources. Indeed, as with the theories of ‘network gatekeeping’ (Barzilai-Nahon, 2008) and ‘relevance filtration and accreditation’ (Benkler, 2006, p. 169), the two cultural thickening patterns are the results of gatekeeping processes created by self-appointed users on the networks. User-contributors must produce relevant content, and user-readers must find the content credible. Thus, obtaining specific geographic and linguistic information from these processes reveals the information orders that such cultural thickening processes produce. Baidu Baike was found to produce a commercially and politically correct mainland China-focused Chinese information order, whereas Chinese Wikipedia produces a culturally and politically diverse Chinese information order for various Chinese societies by filtering out copyright-dubious, self-promoting or advertisement-based content but keeping reliable information sources across Chinese-speaking regions. The different patterns thus demonstrate two Chinese information orders at work.

7.1.4 Cultural politics and the filtering/content regime. The two cultural thickening patterns indicate a plausible explanation as to why two such different Chinese information orders have come into being. This thesis has argued that the filtering/content regime produced both separation and integration effects, rather than isolation or indigenization effects. The block-then-diffusion hypothesis proposed in Chapter 4 suggested that, in the context of mainland Chinese Internet diffusion rates and the Internet filtering of Chinese Wikipedia, the censorship/filtering regime gave Baidu Baike a strong comparative advantage in accessing and maintaining mainland Chinese Internet users. Effectively, the filtering/content regime reshaped the centre of gravity within each Chinese web sphere, thereby showing how Chinese Wikipedia and Baidu Baike operated within different borders and information orders.

Different experiences of information literacy, including using search engines and online encyclopaedias, also mattered here. The two cultural thickening patterns exemplified by Baidu Baike and Chinese Wikipedia show that the censorship/filtering regime produced two distinct Chinese web spheres: one is the censored and domesticated mainland Chinese sphere, and another is the sphere that is marginalized from mainland Chinese users' experiences. In other words, the censorship/filtering regime prevented the genuine exchange and integration of Chinese-language information. Also, the more 'protected' Baidu Baike did not receive substantial contributions from outside mainland China, be it in the form of power users, contributors, or simply information content, whereas the more 'integrative' Chinese Wikipedia was often rendered inaccessible or less visible for the majority of mainland Chinese users. It is clear then that from the viewpoint of the Chinese-language Internet, there are distinct Chinese web spheres with different patterns of ordering Chinese-language information, revealing the cultural politics among Chinese societies. These cultural-political manoeuvres can be interpreted as a containment strategy by Beijing to prevent political opponents from influencing mainland Chinese users.

To conclude, this thesis has provided findings about the relationship between Chinese societies and Chinese information systems. As regards the effects of China's Internet censorship/filtering regime, Table 7-2 summarizes how the proposed analysis differs from the existing isolation and indigenization (cultural protection) views. The isolation view holds that the censorship/filtering regime has a strong isolating effect (i.e. isolation of China from the world), whereas the indigenization view argues that the effect is rather minimal. A third view has been proposed here: China's censorship/filtering regime does in fact have a critical impact in isolating the majority of mainland users from seeing the rest of the Chinese-speaking world. Nevertheless, although the basic geocultural hypothesis of cultural affinity may bring Chinese web spheres closer together because of a common language, this thesis does not downplay the political and

ideological differences between mainland China and the rest of the Chinese-speaking world, notably Hong Kong and Taiwan. The persistent significance of these differences was demonstrated in global TV studies about Greater China (Chan, 1996, 2009). These differences continue to have significant impact on the Web, as shown by the webometric and content findings here. In effect, China's Internet censorship/filtering regime picks the winners and losers for the mainland Chinese users (by giving comparative advantages to Baidu and Baidu Baike over Google and Chinese Wikipedia). While not isolating 'all' mainland Chinese users from the world, it does, however, keep other Chinese views that are outside mainland China's control from going mainstream in mainland China, protecting mainland China from the rest of Chinese-speaking world. The geocultural 'Chinese cluster' therefore has clear political and ideological barriers: For example, despite integrating diverse Chinese information, Chinese Wikipedia is shown to be in a peripheral position for the majority of Chinese users in terms of visibility and web structure. Thus, new insights can be gained from seeing the patterns whereby information is marginalized and mobilized because of its geographical, linguistic, cultural and political features – rather than merely because of the provision or the lack of information access. The impact of the information blockage is not so much whether mainland Chinese are isolated from the world or not, but rather how certain Chinese cultural thickening patterns and Chinese information orders cannot compete with the ones preferred by Beijing.

Table 7-2

Views on the impact of the Great Firewall (GFW) of China

	Isolation of China from the world	Indigenization (cultural protection) of China	'Protection' of mainland China from the rest of the Chinese-speaking world
Impact intensity	strong	little	crucial
Main effects	GFW isolates China from the world.	GFW does not prevent the formation of a single Chinese cluster.	GFW lowers the intensity of exchange between the majority of mainland Chinese users and the rest of the Chinese-speaking world, thereby shaping a Chinese information order desired by Beijing.
Network graphs	Two parallel Internet universes: China and the rest	A single Chinese cluster that is not isolated from the world.	Two overlapping Chinese Internet spaces, but the majority of mainland Chinese visited only one frequently.
Views on Beijing authorities	Authoritarian regime that keeps tight control over Chinese Internet users.	Main national cultural protector and political stabilizer that protect the national interest from foreign influence	Authoritarian regime that competes to remain dominant in various Chinese information spheres, and uses filtering/censorship regime to create comparative advantage in accessing mainland Chinese users commercially, politically and culturally
Views on Hong Kong and/or Taiwan	Peripheral (but integral to China's core)	Peripheral (but bridging between China and the world)	In-between centrality (and even alternative core)
Overall pictures	China's departure from the international norms	(Natural) media choices made by (mainland) Chinese Internet users.	Struggles and competitions in the Chinese-language Internet for certain "desired" Chinese Information order

Hence, these three views differ in their views of Beijing, Hong Kong, and Taiwan, with different interpretations of geopolitical and information networks. Both the isolation and cultural protection (indigenization) views see Hong Kong and Taiwan's role as peripheral; however, the findings here and elsewhere, including evidence presented by Taneja & Wu (2013), show that it is Hong Kong and Taiwan that provides a 'bridge' between mainland China and the rest of the world, functioning as important Chinese information hubs: they are not only

distinct from China and the West, but also the core nodes in Chinese-language popular music, film, and TV production (Gold, 1993). Thus, they both serve as ‘in-between’ information hubs between mainland China and the world. In relation to this in-between view, the findings have shown that Google and Wikipedia’s Chinese information provisions not only mix and integrate information from various Chinese web spheres, but also accommodate (and thus in a sense ‘protect’) the diversity of Chinese cultural and linguistic practices (e.g. both Google and Wikipedia accommodate non-Mandarin dialects such as Cantonese). From the evidence presented, it becomes clear that the actual impact of China’s Internet censorship/filtering regime is not so much about international politics, but rather about instilling a national Chinese information order that Beijing desires, with a dominant information order promoting simplified Chinese at its national core which marginalizes the communicative influence of Chinese information sources outside mainland China, including Chinese Wikipedia. The so-called Great Firewall of China thus reinforces a cultural-political barrier among the Chinese-speaking societies, which effectively creates a comparative advantage for Beijing.

7.2 Implications for research and policy

This thesis has several implications for research and policy, especially concerning the politics of web spaces. First, better understanding of the geolinguistic characteristics of web spaces can avoid overly simplistic generalizations. Second, along with other cultural-political factors, geolinguistic factors set limits to open collaboration and thus have implications for design. Third, there are a number of cultural-political implications relating to the uses of the internet and its content for mobilizing and marginalizing certain groups. By examining the web’s information-processing capacity in accordance with certain geolinguistic specifications, researchers can understand the dynamics of national and transnational webs. Fourth, this thesis has put the development of information and knowledge into the larger context of geolinguistic dynamics

since the Enlightenment, including the tensions between universal knowledge and its diffusion. Altogether, the politics of web spaces may be improved by considering the cultural-political relationship between users and content sources.

7.2.1 Geolinguistic dynamics. Researchers, policymakers and information system designers can avoid overly simple generalizations by examining the specific geolinguistic characteristics of web spaces.

Without these specific characteristics, it is also to conflate China's Internet and the Chinese-speaking Internet. Admittedly, mainland China and the Mandarin Chinese language constitute an indispensable part of the modern Chinese political landscape, and recent developments of the Chinese-language Internet developments in particular. Nonetheless, researchers should not overlook the roles of Hong Kong, Taiwan, and overseas Chinese, or disregard the existence of other Chinese languages and written scripts. In particular, research about the Internet and the design of the Internet in dealing with 'Chineseness' require concepts and methodological tools that make analytical sense of qualifying and quantifying what specifically is 'Chinese'. These concepts and tools must be transparent and subject to cultural-political scrutiny. To illustrate, this thesis has specified several geolinguistic variants of Chinese users and content, and gone on to quantify and provide qualitative analysis of the interactions among these variants.

We have seen that Chinese geolinguistic codes and the corresponding webometric data can be used to compare the preferences of the two websites in drawing from Chinese information sources, contributors and readership. In contrast to Baidu Baike, Chinese Wikipedia operates better in integrating content and users across Chinese-speaking regions, which is likely due to the relative clarity of its policy and design in handling 'Chineseness'. More precise geographical and linguistic specificity benefits researchers as well as practitioners in understanding and building 'national webs' or 'localized webs', also beyond the Chinese context. Picking up the questions raised by the 'national

web studies' (R. Rogers, 2013) and work on internationalization/localization (Dunne, 2006), these efforts contribute both to the empirical understanding and methodological sophistication of what constitutes a 'Cyrillic Web space' (Beumers, Hutchings, & Rulyova, 2008), an 'Iranian web' (R. Rogers, 2013), the 'Arabic online world' (Wahba et al., 2013). Researchers and practitioners can thus better delineate the object of their research or design a web space demarcated by languages, regions, authorship, topics, actual users, national domain names or localized information systems, and the like. Because a more precise specification of geographical and linguistic differences can be supported quantitatively by web data and qualitatively by historical context and other evidence, such efforts can provide complementary methods that connect two types of disciplines: information science and webometrics on one hand and social/political sciences and area studies on the other.

7.2.2 The limits to open collaboration. User-generated encyclopaedias have been a major object for such research on open collaboration. Arguably, the Internet's potential for overcoming boundaries depends on the definition of open collaboration (Riehle, 2012): it has to be 'egalitarian', 'meritocratic', and 'self-organizing'. While both Chinese-language encyclopaedia websites claim to promote open collaboration, the findings indicate their geographical and linguistic limits and thus the cultural politics of open collaboration for Chinese Internet users. This thesis has demonstrated how different Chinese web spaces are open, with different levels and scope of openness, to the participation of sources and users from different Chinese societies. There are, of course, different cultural and political limits, for example, those due to political censorship and copyright. These limits also correspond to the geographic and linguistic differences in the Chinese context: Hong Kong and Taiwan experience more free and democratic societies that use traditional Chinese script, as against a less free and democratic society in mainland China where simplified Chinese script is used. The definition of open collaboration therefore cannot be taken for granted;

it is important to specify rather than assume how the collaborative values of ‘egalitarian’, ‘meritocratic’ and ‘self-organizing’ are actually defined by users, shared, and implemented through policies and computer code.

Geographical and linguistic factors are likely to be built into the platforms of open collaboration, rendering such ‘openness’ geographically and linguistically configured. Not all languages or regions are treated as ‘equal’: some are more ‘meritocratic’ than others; the constituent members of the ‘self-organizing’ body of contributors reveal how the Chineseness is defined in the case of Chinese-language online encyclopaedias. Thus, to specify these limits is useful in governing these open collaboration projects. There are also limits to how such geographic and linguistic differences can be used as a proxy for cultural and political differences, and what applies in this research on the Chinese Internet may not be applicable elsewhere. Still, it is valuable to examine geographic and linguistic limits first so as to pave the way for further cultural-political analysis. Further work will be needed to establish whether geographical and linguistic specification discourages or encourages open collaboration in other contexts. This thesis has shown, in any event, that Baidu Baike limits open collaboration among mainland Chinese users while Chinese Wikipedia has overcome the geolinguistic differences of such collaboration.

Cross-cultural design practices are needed to engage ‘social affordances based on a rich understanding of meaningful contextualized activity’ (Sun, 2012). Geolinguistic analysis and data as provided here can enrich such understanding. Moreover, as media researchers question how a localized and personalized Internet such as with filter bubbles and media bias (Zuckerman, 2013) constitute barriers to cosmopolitanism, geolinguistic analysis can help researchers, policy-makers and designers to improve socio-technical affordances with better cross-cultural, cross-regional and/or cross-lingual design.

7.2.3 Mobilizing and marginalizing users and content. Focusing on geographic and linguistic features can explain, and possibly predict, how certain groups of users and/or content are mobilized or marginalized.

The thesis has illustrated the usefulness of specifying geographic and linguistic aspects of Chineseness for understanding the cultural politics of the Chinese-language Internet. As we have seen, web spheres can be more fragmented (Baidu Baike) or integrated (Chinese Wikipedia). Researchers and practitioners can seek to analyse and then improve the features of web spheres that address geographical, linguistic, and platform factors. For instance, this thesis has examined how geographical and linguistic patterns lead to integration, exclusion, or separation. Chinese Wikipedia was found to be more integrative than Baidu Baike, and Google search more so than Baidu or Yahoo search. Without first pinpointing geographical and linguistic patterns, the thesis would not have been able to make this argument about Chinese information integration. Looking to specifics avoids overly simply generalizations about the existence of a single ‘Chinese information cluster’ (Taneja & Wu, 2013). Future research on the Chinese Internet should continue to examine ‘cultural thickening’ among Chinese-speaking regions and, where possible, among different Chinese-speaking Internet users around the world.

Broadly speaking, more empirical research and methodological tools are needed to examine how particular sources or groups of users are mobilized while others are marginalized. Further research detailing linguistic and geographical patterns and examining other platforms are needed to see how the mobilization and marginalization dynamics play out in different cultural-political contexts. By examining these patterns in a transparent fashion, we can then reveal how information mobilization and marginalization takes place.

7.2.4 Understanding the geolinguistic dynamics of knowledge. As the Web is ‘increasingly grounded with geographical and linguistic specificity by platform and space’ (R. Rogers, 2013, p. 58), we need to reflect on the ramifications of such

groundedness, which range from internationalization/localization of information systems to the cultural protectionism policies of national (or even provincial) media systems. In this regard, past efforts at understanding the Chinese Internet by academics, policy-makers, and industry practitioners has revolved around cultural protectionism, authoritarian control, civic engagement, and similar issues. Yet as we have seen, we can become more specific about geographical and linguistic patterns in dealing with these issues if we do not implicitly assume a particular idea of 'Chineseness' from the start.

Geographical and linguistic analysis provides important theoretical and methodological tools to examine how information flow and knowledge exchange are regulated and shaped. This kind of analysis highlights the relationship between technological and institutional features of the medium on one side and of the geographical and linguistic preferences of medium-users on the other. This is why geolinguistics, a subfield of linguistics and geography, aims to map out the spoken and written forms of language activities, usually in the historical context of the nationalization of writing standards and the globalization of media. This is also why global TV studies developed 'geolinguistic analysis' to chart new patterns of information flows and audience preferences (Chan, 1996, 2009; Sinclair et al., 1996). Geographical and linguistic factors can be studied, designed, and now even computer-coded to rank, regulate, and rearrange information flows across different regions and different groups of users. If '[h]umanity has always been readjusting to developments in the flow of information' (Gleick, 2011), then such readjusting efforts have important cultural and political ramifications.

Indeed, if we consider the historical context of encyclopaedias, the Enlightenment, nation-states, and various political transformations can be related to various new flows of information and knowledge, including new vernacular languages in print media and various knowledge platforms for disseminating scientific and revolutionary ideas, followed by mass-literacy and

the social mobilization of the citizens of modern nation-states. From this longer-term historical perspective, understanding the geolinguistic dynamics of information can provide researchers, policymakers and technology developers with knowledge about how information is processed and diffused in the world.

7.3 Understanding the geolinguistic dynamics of knowledge and information

It is useful to draw together the three concepts of the thesis—processability, web spheres, and cultural thickening—to explain the geolinguistic dynamics of knowledge and the role of the Internet in shaping these dynamics.

7.3.1 Information systems and geolinguistic processability. Modern information systems, as the latest incarnations of media-language systems, feature the capacity to process content and information activities at a larger scale, synchronously, and for different groups of users. While users may be ‘self-organizing’ for open collaboration (Riehle, 2012), I argue that geographic and linguistic factors, more so than the types of information systems, have shaped the new possibilities for information and communication processing activities. The shaping of these activities has also been theorized in terms of ‘network gatekeeping’ (Barzilai-Nahon, 2008) or ‘relevance filtration and accreditation’ (Benkler, 2006, p. 169). I have proposed the term ‘processability’ to conceptualize the enabling conditions, impacts, and design of the media-language system (which includes writing, information, and communication systems) contend that geographic and linguistic factors shaping this system.

Geolinguistic processability is essential in considering the effects of social inclusion and exclusion, or integration and separation, on the Web. Geolinguistic processability and its cultural-political implications demands research which includes analysis of the socio-technical processes that shape geographical and linguistic information content and users, including local domain names, country codes, language tags, locale settings, geo-IP information, input methods, and the like. These factors are important because geographic and linguistic information must be extracted, coded, processed in ways built into the system design. Other

related factors include users' preferences, choices, and capacities to find and manipulate information. Geolinguistic processability therefore has cultural and political implications because it regulates information flow, reshapes social boundaries, and depends on civic literacies for participation.

7.3.2 Web spheres and cultural thickening. Different information systems (including web platforms) demarcate and order different sets of information sources based on geolinguistic processability. Conceptualized by researchers of digital methods as 'a device-demarcated source set' (R. Rogers, 2013, p. 118), web spheres are increasingly communicative and information spaces demarcated by geolinguistic factors. As a research strategy, researchers can specify different geolinguistically-demarcated web spheres to understand the societies they seek to represent.

Taking Chinese-language information systems as example, social researchers can examine different geolinguistic versions of major search engines (as in Chapter 6), compare the geographic and linguistic distribution of online references (as in Chapter 5), and contrast the system designs and policies dealing with geographic and linguistic differences (as in Chapter 4). By comparing information outcomes, researchers can advance national web studies while avoiding issues of methodological nationalism. Thus, the concept of a web sphere can help researchers to understand online public spheres in such a way that 'information, knowledge, and sociality are organized by recommender systems', or simply by reference to 'web epistemology' (R. Rogers, 2013, p. 26). Web spheres are more than static objects for information systems research since they are culturally and politically shaped like knowledge itself. This thesis has presented a case where knowledge and information are demarcated differently for various Chinese-language web spheres that also correspond to different groups of users.

To further examine the cultural-political dynamics between two or more web spheres, I borrowed from media studies the concept of 'cultural thickening'

(Hepp & Couldry, 2009; Löfgren, 1997). Media cultures are ‘a form of thickening of translocal classification systems and formations of the articulation of meaning’ (Hepp, 2013). Though this concept has so far mainly been used to examine national broadcasting and global TV, it can also be applied to information systems to the Web. In this sense, it is similar to classical sociological interpretations of ICTs in terms of everyday routines (Schroeder & Ling, 2013). Researchers can approach web spheres that produce cultural thickening effects for certain groups of users and content. If two web spheres share overlapping content or users, then cultural thickening entails that these two web spheres are brought closer together or become more integrated. Otherwise, if little content is shared among users, then the lack of cultural thickening between the two spheres can be said to promote separation. By integrative or distancing effects among web spheres, researchers can analyse the local, national, and transnational dynamics of media cultures and how web spheres are culturally and politically shaped. ‘Geolinguistic analysis’, developed by global TV studies to examine patterns of information flows and audience preferences (Chan, 1996, 2009; Sinclair et al., 1996) has been extended here to include the data collection and analysis techniques of information science (including webometrics) on the one hand, and the geographically/linguistically rich understanding of social and political science (including area studies) on the other.

7.3.3 Knowledge politics and geolinguistic dynamics. The social impact and shaping of modern web spheres raises a classic question about knowledge: How is information and knowledge produced and spread? Because this process depends on transmission and storage media of writing systems, processability is a useful way to understand the history of nation-states and the Enlightenment as examples of knowledge politics with different geolinguistic dynamics.

On the one hand, for modern nation-states, the scalability of political and communicative systems requires the processability of interactions. The print era has a geolinguistic systematic bias towards a geographically bounded

monolingual communicative space. Building national states and national economies requires the formation of ‘nerves of government’ as feedback loops between the government and the governed (Deutsch, 1953, 1966a, 1966b; Deutsch & Foltz, 2010). The geolinguistic dynamics of knowledge thus include the need to spread literacy; industrialization requires, among other things, national languages and national education systems. The public sphere is ‘a linguistically-constituted public space’ formed by ‘a communicative structure’ (Habermas, 1989, p. 360). Spreading knowledge in this context thus entails building a national communicative space. On the Web and with an Internet infrastructure, geolinguistic processability often reinstates such national differences or groupings, including national domain names, country codes, language tags, locale settings, and input methods which often correspond to certain targeted geolinguistic group of fellow-users. In this sense, the classic notion of a national geographically bounded monolingual communicative space might be reintroduced on the Web.

On the other hand, if the Enlightenment is defined as ‘[t]he process of spreading certain kinds of information, knowledge, understanding and attitudes’ (Chisick, 2005, p. 262), it can also be understood in terms of processability. Indeed, enlightenment means a spreading movement of light/knowledge that eliminates the darkness of ignorance. Some geographic and linguistic dynamics of the Enlightenment have been highlighted by recent research (Israel, 2001; Roche, 2006; Withers, 2008). The public sphere of the pre-1750 Enlightenment bred a reading culture that engendered the rise of the vernacular languages (replacing Latin), which led to the wide diffusion of ideas and knowledge (Israel, 2001). As a prime example of knowledge diffusion, Enlightenment encyclopaedias revealed the geographical and linguistic centre of the Enlightenment movement: they emerged in the French language from Parisian publishing houses (Roche, 2006), with some geolinguistic groups of users more ready than others to take part in the Enlightenment. A number of political and

knowledge revolutions based on some of the universal ideals of the Enlightenment, such as modern democracy (Israel, 2010) and human rights (Israel, 2011), were also spread across national boundaries and around the world. Such Enlightenment attitudes associated with knowledge and information diffusion provided the basis for civic engagement within a new cultural-political space. The Internet and its potential for enhancing civic engagement and civic culture (Benkler, 2006; Rheingold, 2008; Scammell, 2000; Shirky, 2010) can be understood as a part of this continuing historical process.

Yet the two parts of this process of knowledge intertwined with politics, one national and the other universal, appear to be at odds with one another. While citizens must have the capacity to process knowledge as a functioning literate member of the (national) society, they may all share a set of (universal) values that any nation-state should respect and practice. Therein lie the tensions between national cultures and universal human knowledge. Knowledge, universal or not, must be spread through a writing system or other medium that has geolinguistic dynamics and boundaries. The light of knowledge, due to different transmission media with geolinguistic differences, is bound to change as it travels, causing different effects in different corners of the world. From this perspective, the development of the Internet is a natural – but variable – extension of the global Enlightenment movement.

7.3.4 The geolinguistic dynamics of Chinese modernization - and beyond.

Ultimately, the geolinguistic dynamics and reach of media systems constitute the boundaries of human knowledge and social development. China, once the centre of the world as the ‘middle kingdom’ that spread its civilization to Korea, Vietnam, and Japan through the medium of classic Chinese and Confucian teachings, became subject to modernization. This process has distinct geolinguistic dynamics. During modern China’s colonial and Republican period, the foreigner-controlled treaty ports concentrated along the coastal regions of China constituted the major sites for modern China’s public sphere (S. R.

MacKinnon, 1997; Rowe, 1990). Mass education and national language were debated and developed around the May Fourth movement in the Republican era, with limited reach and success (Fairbank, 1983). Later, in the Communist era, different regimes on both sides of the Cold War developed different forms education, culture and identity: Taiwan, Hong Kong, and Macau sided with South Korea, Japan, and the U.S., whereas mainland China sided with North Korea, Vietnam, and Soviet Russia (Nisic, 2001).

The cultural-political development of the region can thus be seen as a particular geolinguistic refraction of the Enlightenment movement, wherein both East Asian and ethnic Chinese regimes seek modernization of both their media-language and cultural-political systems. Note that the modernization of Chinese media-language systems faced historical challenges in appropriating largely alphabet-based information and communication systems to process Chinese-language information. Nonetheless, the rise of Chinese and East Asian use of the Internet has challenged, even if it has rejected, the effect of a Eurocentric alphabet. It can be argued that the regional success of the Internet is due to the regional achievements of general literacy programs, as Ogawa, Jones, & Williamson (1993) point out when they compare this success with Latin America. As with the recent development concerning citizenship in Hong Kong, Taiwan, and mainland China (Xing, Ng, & Cheng, 2012), the cultural-political development of information continues to confront the geolinguistic differences that have been examined here.

In this sense a pair of geographic metaphors used by historians of modern China such as Fairbank (1983) and Wang (1996, 2004) can be extended to the geolinguistic dynamics of the Chinese modernization process: 'Maritime China' refers to modern maritime activities that connect China to the world, whereas 'Continental China' refers to the Chinese agrarian-bureaucratic tradition. Despite being peripheral to geographical China, Maritime China has grown into a richer and bigger segment of Chinese society as the leading economic and

cultural centre of China, whereas continental China views maritime interests as 'narrow and self-serving', of which Hong Kong and Taiwan serving as the 'extreme manifestations' (G. Wang, 1996b, p. 17). A linguistic dimension can be added to this pair of geographic metaphors: Maritime China tends to be more tolerant of language variations and exchanges (Han Chinese southern dialects are concentrated along the East coast of southern China). Continental China tends to be less tolerant because of its bureaucratic tradition that insists on one unified Chinese national language. Baidu Baike and Chinese Wikipedia can thus be understood as the latest incarnations of Continental China and Maritime China, at least in the domains of knowledge processing for Chinese-language users. Geolinguistically, Baidu Baike sticks to simplified Chinese in mainland China, whereas Chinese Wikipedia integrates mainland China alongside the major regions that constitute Maritime China, which even includes Singapore and Malaysia. It is worth mentioning that, although not studied in this thesis, Wikipedia has hosted other southern Chinese dialect versions such as Cantonese, Wu, Min Nan, Hakkanese, and others. Put in this larger context, the diffusion of the Internet across East Asian and mainland Chinese regions exhibits geolinguistic dynamics that shape the reach of knowledge. Therefore, what is known as the Great Firewall of China should be viewed as a cultural-political project that Continental China has imposed on Maritime China as part of the latest twist in the geolinguistic dynamic of Chinese modernization. Effectively, Chinese national boundaries are reinforced, shifted or overcome on Baidu Baike and Chinese Wikipedia because of these geolinguistic dynamics.

Geolinguistic dynamics, understood by means of cultural thickening, hold the key for understanding national and other boundaries of information and communication systems. When we look beyond the Chinese case and consider the Enlightenment and nation-states in terms of knowledge politics with different geolinguistic dynamics, the centrality of major encyclopaedias is evident: They reflect changes and challenges for different societies in the world in

preparing their members to engage with information and knowledge in a way that is at the heart of civic learning experiences and cultural-political membership. For instance, how are Arabic-speaking Internet users engaging across Arabic-speaking regions? The significance of online user-generated encyclopaedias, when compared to their print predecessors, lies in their scalability and ability to reach different geolinguistic groups of content and users. For instance, what does the development of a separate Egyptian Arabic Wikipedia mean when it becomes a competing alternative to the pre-existing Arabic Wikipedia? Online user-generated encyclopaedias demand our attention in the context of the geolinguistic dynamics of knowledge. Such geolinguistic dynamics are also evident in other online platforms beyond encyclopaedias, such as search engines and microblogs. The dynamics of civic information engagement within and across different geolinguistic groups of users and/or content remain an open question. The three proposed concepts of web spheres, processability, and cultural thickening have the potentials to provide new understanding of how the Internet and modern media-language systems shape civic engagement across boundaries beyond the Chinese language context or beyond the scope of encyclopaedic knowledge.

References

Abbreviations used for author names below include the following: Baidu Baike (BB), Baidu Baike contributors (BB contributors:), Chinese Wikipedia (zhWP), and English Wikipedia (enWP).

- Ackland, R. (2013). *Web social science: concepts, data and tools for social scientists in the digital age*. London: Sage. Retrieved from <http://librarytitles.ebrary.com/id/10772991>
- Albert, M., Jacobson, D., & Lapid, Y. (2001). *Identities, Borders, Orders: Rethinking International Relations Theory*. Minneapolis: University of Minnesota Press.
- Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to "webometrics." *Journal of Documentation*, 53(4), 404-426. doi:10.1108/EUM0000000007205
- Alvestrand, H. T. (1998, January). IETF Policy on Character Sets and Languages. IETF (Best Current Practice). Retrieved from <http://tools.ietf.org/html/rfc2277>
- American Planning Association. (2006). *Planning and Urban Design Standards*. Hoboken NJ: John Wiley & Sons.
- Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.
- Anderson, B. (1992). *Long-distance nationalism: World capitalism and the rise of identity politics*. Amsterdam: Centre for Asian Studies Amsterdam.
- Anderson, B. (2001). Western Nationalism and Eastern Nationalism: Is there a difference that matters? *New Left Review*, 9.
- Anderson, M. (2011, June). Parlez-vous Facebook? *IEEE Spectrum*, 48(6). Retrieved from <http://spectrum.ieee.org/at-work/innovation/parlezvous-facebook>
- Ayers, P., Matthews, C., & Yates, B. (2008). *How Wikipedia works: and how you can be a part of it*. San Francisco: No Starch Press.
- Baark, E. (1997). *Lightning wires: The telegraph and China's technological modernization, 1860-1890*. Westport CT: Greenwood Press.
- Baeza-Yates, R., Castillo, C., Telefónica, C., & Efthimiadis, E. N. (2007). Characterization of national Web domains. *ACM Transactions on Internet Technology*, 7(2), Article No. 9. doi:10.1145/1239971.1239973
- Bamman, D., O'Connor, B., & Smith, N. (2012). Censorship and deletion practices in Chinese social media. *First Monday*, 17(3-5). doi:10.5210/fm.v17i3.3943
- Banerjee, I. (2007). *The Internet and Governance in Asia: A Critical Reader*. Singapore: Asian Media Information and Communication Centre.

- Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). Omnipedia: bridging the Wikipedia language gap. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (pp. 1075–1084). New York: ACM. doi:10.1145/2207676.2208553
- Bar-Ilan, J. (2006). Web links and search engine ranking: The case of Google and the query “jew.” *Journal of the American Society for Information Science and Technology*, 57(12), 1581–1589. doi:10.1002/asi.20404
- Bar-Ilan, J. (2007a). Google Bombing from a Time Perspective. *Journal of Computer-Mediated Communication*, 12(3). doi:10.1111/j.1083-6101.2007.00356.x
- Bar-Ilan, J. (2007b). Manipulating search engine algorithms: the case of Google. *Journal of Information, Communication and Ethics in Society*, 5(2/3), 155–166. doi:10.1108/14779960710837623
- Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century—A review. *Journal of Informetrics*, 2(1), 1–52. doi:10.1016/j.joi.2007.11.001
- Barmé, G. R. (2010, January 29). The Harmonious Evolution of Information in China [Web log post]. Retrieved from <http://www.thechinabeat.org/?p=1422>
- Barnett, G. A., Chung, C. J., & Park, H. W. (2011). Uncovering Transnational Hyperlink Patterns and Web-Mediated Contents: A New Approach Based on Cracking.com Domain. *Social Science Computer Review*, 29(3), 369–384. doi:10.1177/0894439310382519
- Barzilai-Nahon, K. (2008). Toward a theory of network gatekeeping: A framework for exploring information control. *Journal of the American Society for Information Science and Technology*, 59(9), 1493–1512. doi:10.1002/asi.20857
- Battelle, J. (2005). *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture* (1st ed.). New York: Portfolio.
- BB. (2006). Baidu Baike help [bǎidù bǎikē bāngzhù 百度百科 帮助]. Retrieved August 9, 2010, from http://web.archive.org/web/20060424182803/http://www.baidu.com/search/baike_help.html
- BB. (2009, September 14). Kedou Group Rules [kēdǒu tuán zǒngzé 蝌蚪团总则]. Retrieved October 7, 2012, from <http://www.baidu.com/search/baike/odp/main.html>
- BB. (2012a). Baidu Baike help [bǎidù bǎikē bāngzhù 百度百科 帮助]. Retrieved August 9, 2012, from http://www.baidu.com/search/baike_help.html
- BB. (2012b). Baike Kedou Group Rules (April 26, 2012) [bǎikē kēdǒu tuán zhāngchéng 百科蝌蚪团章程 (2012 年 4 月 26 日修订)]. Retrieved October 7, 2012, from <http://tieba.baidu.com/p/1556813452>
- BB. (2012c). Ranking of user contribution [gòngxiàn bǎng 贡献榜]. Retrieved December 5, 2012, from <http://baike.baidu.com/star/contribution/grow.html?func=congrow>

- BBC. (2011, March 31). Google's China exit "exaggerated." *BBC*. Retrieved from <http://www.bbc.co.uk/news/business-12917322>
- BB contributors. (2012, September 30). Baike Kedou Group [bǎikē kēdǒu tuán 百科蝌蚪团]. In *Baidu Baike*. Retrieved from <http://baike.baidu.com/view/881001.htm>
- Beilock, R., & Dimitrova, D. V. (2003). An exploratory model of inter-country Internet diffusion. *Telecommunications Policy*, 27(3-4), 237-252. doi:10.1016/S0308-5961(02)00100-3
- Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven: Yale University Press.
- Bennett, W. L., & Wells, C. (2009). Civic Engagement: Bridging Differences to Build a Field of Civic Learning. *International Journal of Learning and Media*, 1(3), 1-10. doi:10.1162/ijlm_a_00029
- Bermejo, F. (2009). Audience manufacture in historical perspective: from broadcasting to Google. *New Media & Society*, 11(1-2), 133-154. doi:10.1177/1461444808099579
- Beumers, B., Hutchings, S., & Rulyova, N. (2008). *The Post-Soviet Russian Media: Conflicting Signals*. Routledge.
- Bilenko, M., & White, R. W. (2008). Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *Proceeding of the 17th international conference on World Wide Web* (pp. 51-60).
- Biuk-Aghai, R. P. (2006). Visualizing Co-Authorship Networks in Online Wikipedia. In *International Symposium on Communications and Information Technologies, 2006. ISCIT '06* (pp. 737-742). doi:10.1109/ISCIT.2006.339838
- Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227. doi:10.1002/asi.20077
- Blaikie, P. (1978). The Theory of the Spatial Diffusion of Innovations: A Spacious Cul-De-Sac. *Progress in Human Geography*, 2(2), 268-295. doi:10.1177/030913257800200204
- Blondheim, M. (1994). *News over the wires: the telegraph and the flow of public information in America, 1844-1897*. Cambridge: Harvard University Press.
- Bolsover, G., Dutton, W. H., Law, G., & Dutta, S. (2013). Social Foundations of the Internet in China and the New Internet World: A Cross-National Comparative Perspective. In *China and the New Internet World, an International Communication Association (ICA) Preconference*. Oxford Internet Institute, University of Oxford. Retrieved from <http://papers.ssrn.com/abstract=2276482>
- Borgman, C. L. (2007). *Scholarship in the digital age: information, infrastructure, and the Internet*. Cambridge: MIT Press.

- Bosworth, A. (2004). Globalization in the Information Age: Western, Chinese and Arabic Writing Systems. *Globalization*, 4(2). Retrieved from <http://globalization.icaap.org/content/v4.2/bosworth.html>
- Brady, A.-M. (2009). *Marketing Dictatorship: Propaganda and Thought Work in Contemporary China*. Lanham: Rowman & Littlefield.
- Bresnahan, T. F., & Yin, P.-L. (2006a). *Economic and Technical Drivers of Technology Choice: Browsers* (HBS Working Paper No. 06-008). Retrieved from <http://hbswk.hbs.edu/item/5439.html>
- Bresnahan, T. F., & Yin, P.-L. (2006b). Standard setting in markets: the browser war. In S. Greenstein & V. Stango (Eds.), *Standards and Public Policy*. Cambridge: Cambridge University Press.
- Brettel, M., & Spilker-Attig, A. (2010). Online advertising effectiveness: a cross-cultural comparison. *Journal of Research in Interactive Marketing*, 4(3), 176–196. doi:10.1108/17505931011070569
- Bruce, S., & Yearley, S. (2006). Diffusion of Innovation. *The Sage dictionary of sociology* (p. 73). London: SAGE.
- Brunn, S. D. (1998). A treaty of Silicon for the treaty of Westphalia? New territorial dimensions of modern statehood. *Geopolitics*, 3(1), 106–131. doi:10.1080/14650049808407610
- Bruns, A. (2008). Life beyond the public sphere: Towards a networked model for political deliberation. *Information Polity*, 13(1), 71–85.
- Bruns, A., Burgess, J., Highfield, T., Kirchhoff, L., & Nicolai, T. (2011). Mapping the Australian Networked Public Sphere. *Social Science Computer Review*, 29(3), 277–287. doi:10.1177/0894439310382507
- Burgess, J. E., Foth, M., & Klaebe, H. G. (2006). Everyday creativity as civic engagement: A cultural citizenship view of new media. In *Communications Policy & Research Forum*. Sydney. Retrieved from <http://eprints.qut.edu.au/5056>
- Burgmer, C. (2009, August 6). Python, Unicode and the digital divide [Web log post]. Retrieved from <http://cburgmer.nfshost.com/content/python-unicode-and-digital-divide>
- Butler, B., Joyce, E., & Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *Proceedings of the 26th annual SIGCHI conference on Human factors in computing systems* (pp. 1101–1110). New York: ACM. doi:10.1145/1357054.1357227
- Cairncross, F. (2001). *The Death of Distance: How the Communications Revolution is Changing Our Lives*. Boston: Harvard Business Press.
- Callahan, W. A. (2005). Nationalism, Civilization and Transnational Relations: the discourse of Greater China. *Journal of Contemporary China*, 14(43), 269–289. doi:10.1080/10670560500065629

- Cammaerts, B., & Van Audenhove, L. (2005). Online political debate, unbounded citizenship, and the problematic nature of a transnational public sphere. *Political Communication*, 22(2), 179–196. doi:10.1080/10584600590933188
- Carey, J. W. (1988). Technology and Ideology: The Case of the Telegraph. In *Communications as Culture: Essays on Media and Society* (pp. 201–230). London: Routledge.
- Cartier, C. (2011). *Globalizing South China*. Oxford: Blackwell.
- Castells, M. (2008). The New Public Sphere: Global Civil Society, Communication Networks, and Global Governance. *The ANNALS of the American Academy of Political and Social Science*, 616(1), 78–93. doi:10.1177/0002716207311877
- Chang A. T. (2005, October 17). Letters: Google contradicts itself. *Taipei Times*, p. 8.
- Chan, J. M. (1996). Television in Greater China: Structure, Exports, and Market Formation. In J. Sinclair, E. Jacka, & S. Cunningham (Eds.), *New Patterns in Global Television: Peripheral Vision*. Oxford: Oxford University Press.
- Chan, J. M. (2009). Toward TV Regionalization in Greater China and Beyond. TV China: Institutions, Programming, and Audiences in Greater China and the Chinese Diaspora. In Y. Zhu & C. Berry (Eds.), *TV China*. Bloomington: Indiana University Press.
- Charlton, G. (2012, February 13). Why Wikipedia is top on Google: the SEO truth no-one wants to hear [Web log post]. Retrieved from <https://econsultancy.com/blog/9009-why-wikipedia-is-top-on-google-the-seo-truth-no-one-wants-to-hear>
- Chase, M. S. (2011, October 5). Chinese suspicion over US intentions. *Asia Times Online*. Retrieved from <http://www.atimes.com/atimes/China/MJ05Ado2.html>
- Chau, M., Fang, X., & Yang, C. C. (2007). Web searching in Chinese: A study of a search engine in Hong Kong. *Journal of the American Society for Information Science and Technology*, 58(7), 1044–1054. doi:10.1002/asi.20592
- Chebotarev, T., & Ingersoll, J. S. (Eds.). (2004). *Russian and East European books and manuscripts in the United States: proceedings of a conference in honor of the fiftieth anniversary of the Bakhmeteff Archive of Russian and East European History and Culture*. New York: Routledge.
- Chen, B., Yeh, Y.-M., Huang, Y.-M., & Chen, Y.-T. (2006). Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information. In *ICASSP 2006 Proceedings* (Vol. I, pp. 969–972). IEEE. doi:10.1109/ICASSP.2006.1660184
- Chen, D.-Y., & Lee, C.-P. (2008). To reinforce or to mobilize?: tracing the impact of internet use on civic engagement in Taiwan. In *Proceedings of the 2nd international conference on Theory and practice of electronic governance* (pp. 394–401). New York: ACM. doi:10.1145/1509096.1509178

- Chen, J. (2008). *Essays on auction mechanisms and resource allocation in keyword advertising* (Unpublished doctoral dissertation). University of Texas at Austin, Austin.
- Chen, P. (1993). Modern Written Chinese in Development. *Language in Society*, 22(4), 505–537. doi:10.2307/4168472
- Chen, Z. (2012). What are the distinct features of entries of Baidu Baike and those of Chinese Wikipedia? What makes them different? [百度百科词条与中文维基百科条目各有什么特点? 原因是什么?] [Web log post]. Retrieved from <http://www.zhihu.com/question/19710565>
- China Digital Times. (2011). The Great Chinese LAN (local area network). Retrieved December 8, 2012, from [http://chinadigitaltimes.net/space/The_Great_Chinese_LAN_\(local_area_network\)](http://chinadigitaltimes.net/space/The_Great_Chinese_LAN_(local_area_network))
- Chisick, H. (2005). *Historical Dictionary of the Enlightenment*. Lanham: Scarecrow Press.
- Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T.-H., & Chen, H. (2004). Internet searching and browsing in a multilingual world: An experiment on the Chinese Business Intelligence Portal (CBizPort). *Journal of the American Society for Information Science and Technology*, 55(9), 818–831. doi:10.1002/asi.20025
- Chu, X. (2013). *Appendix B. Self Censorship Efforts on Chinese Wikipedia*. Retrieved from https://docs.google.com/file/d/oB8ztBERe_FUwLWxUXolaeWF3aEo
- CIC. (2009). *China Search Engine Market Report 2009*. Beijing: China IntelliConsulting Corporation. Retrieved from <http://tech.sina.com.cn/z/2009ssdc/index.shtml>
- Cliff, A., Pred, A., & Hagerstrand, T. (1992). Classics in human geography revisited: Hagerstrand, T. 1967: Innovation diffusion as a spatial process. Chicago: University of Chicago Press. Translation and postscript by Allan Pred. *Progress in Human Geography*, 16(4), 541–544. doi:10.1177/030913259201600403
- CNNIC. (2006, September 16). Chinese Search Engine Market Survey Report 2006. Retrieved November 19, 2011, from <http://xtlv.cn/html/Dir/2006/11/06/4216.htm>
- CNNIC. (2007, September 26). 2007 Survey Report on Search Engine Market in China. Retrieved November 19, 2011, from <http://www.cnnic.cn/html/Dir/2007/10/10/4838.htm>
- CNNIC. (2009, March 5). China Search Engine Report 2008 Advertisers and Users Behavior Study. Retrieved November 19, 2011, from <http://www.cnnic.cn/html/Dir/2009/03/05/5483.htm>
- CNNIC. (2013). *Statistical Report on Internet Development in China (January 2013)* (No. 31st). Beijing: China Internet Network Information Center.

- Cole, J. M. (2009). *Democracy in Peril: Taiwan's Struggle for Survival From Chen Shui-Bian to Ma Ying-Jeou*. New York: iUniverse.
- Cote, P. (n.d.). *Effective Cartography: Mapping with Quantitative Data*. Cambridge, Mass.: Harvard Graduate School of Design. Retrieved from <http://www.gsd.harvard.edu/gis/manual/normalize/>
- Couldry, N. (2012). *Media, Society, World: Social Theory and Digital Media Practice*. Cambridge: Polity.
- Couldry, N., & Hepp, A. (2013). Comparing Media Cultures. In F. Esser & T. Hanitzsch (Eds.), *Handbook of Comparative Communication Research*. Abingdon: Routledge.
- Coulmas, F. (2000). The Nationalization of Writing. *Studies in the Linguistic Science*, 30(1), 47–50.
- Couvering, E. V. (2004). New Media? The Political Economy of Internet Search Engines. In *International Association of Media & Communications Researchers*. Porto Alegre, Brazil.
- Couvering, E. V. (2008). The History of the Internet Search Engine: Navigational Media and the Traffic Commodity. In A. Spink & M. Zimmer (Eds.), *Web Search* (Vol. 14, pp. 177–206). Berlin: Springer Berlin Heidelberg.
- Čuhalev, J. (2006, October 13). Ranking of Wikipedia articles on search engines for searches about its own articles [Web log post]. Retrieved from <http://www.jurecuhalev.com/blog/2006/10/13/seeing-lots-of-wikipedia-in-your-google-searches/>
- Dahlberg, L. (2001). The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. *Information, Communication & Society*, 4(4), 615–633. doi:10.1080/13691180110097030
- Dahlberg, L. (2005). The Corporate Colonization of Online Attention and the Marginalization of Critical Communication? *Journal of Communication Inquiry*, 29(2), 160–180. doi:10.1177/0196859904272745
- Dahlgren, P. (2005). The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation. *Political Communication*, 22(2), 147–162. doi:10.1080/10584600590933160
- Damm, J. (2007). The Internet and the fragmentation of Chinese society. *Critical Asian Studies*, 39, 273–294. doi:10.1080/14672710701339485
- Damm, J. (2008). Technological Empowerment: The Internet, State, and Society in China. By Yongnian Zheng. Stanford, Calif.: Stanford University Press, 2008. xviii, 272 pp. \$50.00 (cloth). *The Journal of Asian Studies*, 67(03), 1080–1081. doi:10.1017/S0021911808001435
- Darnton, R. (1979). *The Business of Enlightenment: A Publishing History of the Encyclopedie 1775*. Cambridge: Harvard University Press.
- Davis, E. L. (Ed.). (2009). *Encyclopedia of Contemporary Chinese Culture*. London: Routledge.

- Davis, M. (2010, January 28). Unicode nearing 50% of the web [Web log post]. Retrieved from <http://googleblog.blogspot.com/2010/01/unicode-nearing-50-of-web.html>
- Davis, M. (2012, February 3). Unicode over 60 percent of the web [Web log post]. Retrieved from <http://googleblog.blogspot.com/2012/02/unicode-over-60-percent-of-web.html>
- DCCI. (2012). *Blue Book of China MicroBlog*. Beijing: Data Center of China Internet (DCCI). Retrieved from <http://www.chinainternetwatch.com/1760/chinese-weibo-users-proportion-on-different-platforms/>
- Dean, J. (2003). Why the Net is not a Public Sphere. *Constellations*, 10(1), 95–112. doi:10.1111/1467-8675.00315
- DeFrancis, J. (1989). *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu: University of Hawaii Press.
- Delanty, G. (2006). *Europe and Asia beyond East and West*. London: Routledge.
- Delanty, G. (2009). *The Cosmopolitan Imagination: The Renewal of Critical Social Theory*. Cambridge: Cambridge University Press.
- DeLisle, J. (2010). Soft Power in a Hard Place: China, Taiwan, Cross-Strait Relations and US Policy. *Orbis*, 54(4), 493–524. doi:10.1016/j.orbis.2010.07.002
- DePalma, D. A. (2002). Internationalization and Localization. In *Business without borders: a strategic guide to global marketing*. New York: John Wiley and Sons.
- Department of Cultural Affairs. (2005). The 3rd Chinese Character Festival. Retrieved February 4, 2012, from <http://www.chinesecharacter.culture.gov.tw/3rdhanzi/english.htm>
- Department of Cultural Affairs. (2011, April 21). An Introduction to the Chinese Character Festival [TEXT]. Retrieved February 4, 2012, from <http://english.doca.taipei.gov.tw/ct.asp?xItem=1118369&ctNode=31734&mp=119002>
- Derthick, K., Tsao, P., Kriplean, T., Borning, A., Zachry, M., & McDonald, D. W. (2011). Collaborative Sensemaking during Admin Permission Granting in Wikipedia. In A. A. Ozok & P. Zaphiris (Eds.), *Online Communities and Social Computing* (pp. 100–109). Berlin: Springer Berlin Heidelberg.
- Derudder, B., & Witlox, F. (2005). An appraisal of the use of airline data in assessing the world city network: a research note on data. *Urban Studies*, 42(13), 2371–2388. doi:10.1177/0969776412453149
- Deutsch, K. W. (1951). Mechanism, Organism, and Society: Some Models in Natural and Social Science. *Philosophy of Science*, 18(3), 230–252.
- Deutsch, K. W. (1953). The Growth of Nations: Some Recurrent Patterns of Political and Social Integration. *World Politics*, 5(2), 168–195. doi:10.2307/2008980

- Deutsch, K. W. (1966a). *Nationalism and Social Communication: An Inquiry into the Foundations of Nationality* (2nd ed.). Cambridge: MIT Press.
- Deutsch, K. W. (1966b). *The nerves of government: models of political communication and control*. New York: Free Press.
- Deutsch, K. W., & Foltz, W. J. (2010). *Nation Building in Comparative Contexts*. New Brunswick: AldineTransaction.
- Dickie, M. (2006, May 10). China launches version of Wikipedia. *Financial Times*. London. Retrieved from <http://www.ft.com/cms/s/2/bd89e998-e056-11da-9e82-0000779e2340.html>
- Ding, S., & Saunders, R. A. (2006). Talking up China: An analysis of China's rising cultural power and global promotion of the Chinese language. *East Asia*, 23(2), 3–33. doi:10.1007/s12140-006-0021-2
- Doctorow, C. (2009, June 2). Search algorithms are editorial decisions. Retrieved February 15, 2012, from <http://boingboing.net/2009/06/02/search-algorithms-ar.html>
- Doreian, P., Batagelj, V., & Ferligoj, A. (2004). Generalized blockmodeling of two-mode network data. *Social Networks*, 26(1), 29–53. doi:10.1016/j.socnet.2004.01.002
- Drozдова, K., & Gaubatz, K. T. (2009). Structured, Focused Uncertainty: Information Analysis for Multi-Method Comparative Case Studies. In *APSA 2009 Meeting Paper*. Toronto. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1451646
- Dugan, L. (2011, July 27). 140 Characters On Chinese Twitter Is More Like 500 Characters On Twitter.com. Retrieved April 15, 2013, from http://www.mediabistro.com/alltwitter/140-characters-on-chinese-twitter-is-more-like-500-characters-on-twitter-com_b11951
- Dunay, P., & Krueger, R. (2011). *Facebook Marketing For Dummies*. Hoboken: John Wiley & Sons.
- Dunleavy, P., Margetts, H., Bastow, S., Pearce, O., & Tinkler, J. (2007). *Government on the internet: progress in delivering information and services online*. UK: National Audit Office. Retrieved from http://www.nao.org.uk/publications/nao_reports/06-07/0607529.pdf
- Dunne, K. J. (2006). *Perspectives on Localization*. Amsterdam: John Benjamins.
- Dutta, S., Dutton, W. H., & Law, G. (2011). The New Internet World: A Global Perspective on Freedom of Expression, Privacy, Trust and Security Online. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1810005
- Dutton, W. H., & Eynon, R. (2009). Networked Individuals and Institutions: A Cross-Sector Comparative Perspective on Patterns and Strategies in Government and Research. *The Information Society*, 25, 198–207. doi:10.1080/01972240902848914

- Easingwood, C. J., Mahajan, V., & Muller, E. (1983). A Nonuniform Influence Innovation Diffusion Model of New Product Acceptance. *Marketing Science*, 2(3), 273–295. doi:10.1287/mksc.2.3.273
- Einhorn, B. S., Bruce. (2010, November 11). How Baidu Won China. *BusinessWeek: Online Magazine*. Retrieved from http://www.businessweek.com/magazine/content/10_47/b4204060242597_page_6.htm
- Eisenstein, E. L. (1979). *The printing press as an agent of change: communications and cultural transformations in early modern Europe, volumes I and II*. Cambridge: Cambridge University Press.
- Enquiro. (2007, June 15). Chinese Eye Tracking Study: Baidu Vs Google. Retrieved July 9, 2009, from <http://searchengineland.com/chinese-eye-tracking-study-baidu-vs-google-11477>
- enWP. (2013). Chinese Wikipedia. In *English Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Chinese_Wikipedia
- Ess, C. (1998). Cosmopolitan Ideal or Cybercentrism? A Critical Examination of the Underlying Assumptions of “The Electronic Global Village.” *APA Newsletter on Computers*, 9(2). Retrieved from http://www.cddc.vt.edu/digitalfordism/fordism_materials/ess.htm
- Ess, C. (2002). Computer-mediated colonization, the renaissance, and educational imperatives for an intercultural global village. *Ethics and Information Technology*, 4(1), 11–22. doi:10.1023/A:1015227723904
- Etling, B., Kelly, J., & Faris, R. (2009). Mapping Chinese Blogosphere. In *7th Annual Chinese Internet Research Conference (CIRC 2009)*. Annenberg School for Communication, University of Pennsylvania, Philadelphia, US. Retrieved from <http://www.global.asc.upenn.edu/circ/webcast.html>
- Etling, B., Kelly, J., Faris, R., & Palfrey, J. (2010). Mapping the Arabic blogosphere: politics and dissent online. *New Media & Society*, 12(8), 1225–1243. doi:10.1177/1461444810385096
- Fairbank, J. K. (1983). *The Cambridge history of China: Republican China 1912-1949, Part 1*. Cambridge: Cambridge University Press.
- Fallows, D. (2008). Search Engine Use. Retrieved November 19, 2011, from <http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx>
- Farrar, L. (2009, October 15). It's tricky for wikis and online encyclopedias in China. Retrieved October 25, 2009, from <http://edition.cnn.com/2009/TECH/10/14/wiki.china/>
- Fei, X. (1991). *Zhōnghuá mínzú yánjiū xīn tàn suǒ* (中华民族研究新探索). Beijing: China Social Sciences Press (中国社会科学出版社).
- Fidler, D. P. (1998). The Internet and the Sovereign State: The Role and Impact of Cyberspace on National and Global Governance. *Indiana Journal of Global Legal Studies*, 5(2), 415–421.
- Fildes, J. (2007, August 15). Wikipedia “shows CIA page edits.” *BBC*. Retrieved from <http://news.bbc.co.uk/2/hi/technology/6947532.stm>

- Fitton, L. (2012). *Twitter For Dummies*. Hoboken: John Wiley & Sons.
- Fitzgerald, J. (1995). The Nationless State: The Search for a Nation in Modern Chinese Nationalism. *The Australian Journal of Chinese Affairs*, (33), 75. doi:10.2307/2950089
- Fløysand, A., & Jakobsen, S.-E. (2011). The complexity of innovation: A relational turn. *Progress in Human Geography*, 35(3), 328–344. doi:10.1177/0309132510376257
- Foot, K. A., & Schneider, S. M. (2002). Online Action in Campaign 2000: An Exploratory Analysis of the U.S. Political Web Sphere. *Journal of Broadcasting & Electronic Media*, 46(2), 222–244. doi:10.1207/s15506878jobem4602_4
- Forte, A., & Bruckman, A. (2008). Scaling Consensus: Increasing Decentralization in Wikipedia Governance. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual* (pp. 157–157). doi:10.1109/HICSS.2008.383
- Fossum, J. E., & Schlesinger, P. (2007). *The European Union and the Public Sphere: A Communicative Space in the Making?*. Abingdon: Routledge.
- Foster, W., & Goodman, S. E. (2000). *The Diffusion of the Internet in China*. Center for International Security and Cooperation (CISAC), Stanford University.
- Fraser, N. (2007). Transnationalizing the public sphere: On the legitimacy and efficacy of public opinion in a post-Westphalian world. *Theory, Culture & Society*, 24(4), 7–30.
- French, H. W. (2006, November 29). China News: Chinese-Language Wikipedia Presents Different View of History. *The New York Times*. Retrieved from <http://chinadigitaltimes.net/2006/11/chinese-language-wikipedia-presents-different-view-of-history-howard-w-french/>
- Fu, K., Chan, C., & Chau, M. (2013). Assessing Censorship on Microblogs in China: Discriminatory Keyword Analysis and Impact Evaluation of the “Real Name Registration” Policy. *IEEE Internet Computing*, 17(3), 42–50.
- Fu, K., & Chau, M. (2013). Reality Check for the Chinese Microblog Space: a random sampling approach. *PLOS ONE*, 8(3), e58356.
- Fung, A. Y. H., & Lee, C.-C. (1994). Hong Kong's changing media ownership: Uncertainty and dilemma. *International Communication Gazette*, 53(1-2), 127–133. doi:10.1177/001654929405300109
- Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities* (Vol. 8). New York: Wiley.
- Geiger, R. S. (2009). Does Habermas Understand the Internet? The Algorithmic Construction of the Blog/Public Sphere. *Gnovis*, 10(1), 1–29.
- Geiger, R. S. (2010). Bot Politics: The Domination, Subversion, and Negotiation of Code in Wikipedia. In *Critical Point of View Conference, Amsterdam* (Vol. 26, p. 27).

- Geiger, R. S., & Ribes, D. (2010). The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 117–126). New York: ACM. doi:10.1145/1718918.1718941
- Gellner, E. (1983). *Nations and nationalism*. Ithaca: Cornell University Press.
- George, A. L., & Bennett, A. (2004). The Method of Structured, Focused Comparison. In *Case Studies and Theory Development in the Social Sciences* (Vol. 70). Cambridge: MIT Press.
- Gleick, J. (2011). *The Information: A History, a Theory, a Flood*. New York: Pantheon.
- Goggin, G., & McLelland, M. (2009). The internationalization of the internet and its implications for media studies. In D. K. Thussu (Ed.), *Internationalizing Media Studies* (pp. 294–307). Abingdon: Routledge.
- Goldsmith, J. L., & Wu, T. (2008). *Who Controls the Internet? Illusions of a Borderless World: Illusions of a Borderless World*. New York: Oxford University Press.
- Gold, T. B. (1993). Go With Your Feelings: Hong Kong and Taiwan Popular Culture in Greater China. *The China Quarterly*, 136, 907–925. doi:10.1017/S0305741000032380
- Google. (2011a). Google in Your Language. Retrieved June 24, 2011, from www.google.com/transconsole
- Google. (2011b). Language Tools. Retrieved June 24, 2011, from http://www.google.com/language_tools?hl=en
- Google. (2012a). Google Permissions. Retrieved February 15, 2012, from <http://www.google.com/permissions/>
- Google. (2012b). Google Public DNS. Retrieved March 26, 2012, from <http://code.google.com/speed/public-dns/>
- Google Support. (2011). Search preferences : Search history and settings - Web Search Help. Retrieved June 24, 2011, from <http://www.google.com/support/websearch/bin/static.py?hl=en&page=guide.cs&guide=1224171&answer=35892&rd=1>
- Graells, E., & Baeza-Yates, R. (2008). Evolution of the Chilean Web: A Larger Study. In *Latin American Web Conference, 2008. LA-WEB '08*. (pp. 108–114). doi:10.1109/LA-WEB.2008.19
- Graham, M., Hale, S. A., & Stephens, M. (2011). *Geographies of the World's Knowledge*. London: Convoco!
- Gray, M. (2007, May). Google Love Affair with Wikipedia - Graywolf's SEO Blog. Retrieved December 2, 2011, from <http://www.wolf-howl.com/google/google-love-affair-with-wikipedia/>
- Gries, P. H. (2004). *China's new nationalism*. Berkeley: University of California Press.

- Grimmelmann, J. (2009). The Google Dilemma. *New York Law School Law Review*, Jan. 2009. Retrieved from http://works.bepress.com/james_grimmelmann/19/
- Grosswiler, P. (2004). Dispelling the Alphabet Effect. *Canadian Journal of Communication*, 29(2). Retrieved from <http://www.cjc-online.ca/index.php/journal/article/view/1432>
- Guo, S., & Guo, B. (2010). *Greater China in an Era of Globalization*. Lanham: Rowman & Littlefield.
- Habash, N. Y. (2010). *Introduction to Arabic Natural Language Processing*. San Rafael: Morgan & Claypool.
- Habermas, J. (1989). *The Structural Transformation of the Public Sphere: An Inquiry Into a Category of Bourgeois Society*. Cambridge: MIT Press.
- Hafner, K. (2007, August 19). Seeing Corporate Fingerprints in Wikipedia Edits. *The New York Times*. Retrieved from <http://www.nytimes.com/2007/08/19/technology/19wikipedia.html>
- Hagestad II, W. (2012). *21st Century Chinese Cyberwarfare*. Ely: IT Governance.
- Hale, S. A. (2012). Net Increase? Cross-lingual Linking in the Blogosphere. *Journal of Computer-Mediated Communication*, 17(2), 135–151. doi:10.1111/j.1083-6101.2011.01568.x
- Halfaker, A., & Riedl, J. (2012). Bots and Cyborgs: Wikipedia's Immune System. *Computer*, 79–82. doi:10.1109/MC.2012.82
- Hallin, D. C., & Mancini, P. (Eds.). (2004). *Comparing Media Systems: Three Models of Media and Politics*. Cambridge: Cambridge University Press.
- Hallin, D. C., & Mancini, P. (Eds.). (2011). *Comparing Media Systems Beyond the Western World*. Cambridge: Cambridge University Press.
- Hargittai, E. (2007). The Social, Political, Economic, and Cultural Dimensions of Search Engines: An Introduction. *Journal of Computer-Mediated Communication*, 12(3), 769–777. doi:10.1111/j.1083-6101.2007.00349.x
- Hayek, F. A. (1967). *Studies in Philosophy, Politics and Economics*. Chicago: University of Chicago Press.
- Hayek, F. A. (1984). *Rules and order: a new statement of the liberal principles of justice and political economy*. Chicago: University of Chicago Press.
- Haythornthwaite, C. (2005). Social networks and Internet connectivity effects. *Information, Communication & Society*, 8(2), 125–147. doi:10.1080/13691180500146185
- Hearne, R. (2006, August 12). SERP Click Through Rate of Google Search Results – AOL-data.tgz – Want to Know How Many Clicks The #1 Google Position Gets? Retrieved December 2, 2011, from <http://www.redcardinal.ie/search-engine-optimisation/12-08-2006/clickthrough-analysis-of-aol-datatz/>
- He, B. (2003). Why is Establishing Democracy so Difficult in China: The Challenge of China's National Identity Question. *Contemporary Chinese Thought*, 35(1), 71–92.

- Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on Communities and technologies* (pp. 11–20). New York: ACM. doi:10.1145/1556460.1556463
- Hecht, B., & Gergle, D. (2010). The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 291–300). New York: ACM. doi:10.1145/1753326.1753370
- He, H. Y. (2001). da Zhonghua 大中華 (Greater China). *Dictionary of the political thought of the People's Republic of China* (中华人民共和国政治文化用语大典) (pp. 47–48). Armonk: M.E. Sharpe.
- Hepp, A. (2009). Transculturality as a Perspective: Researching Media Cultures Comparatively. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 10(1). Retrieved from <http://www.qualitative-research.net/index.php/fqs/article/view/1221>
- Hepp, A. (2013). *Cultures of Mediatization*. Cambridge: Polity.
- Hepp, A., & Couldry, N. (2009). What should comparative media research be comparing? Towards a transcultural approach to “media cultures.” In D. K. Thussu (Ed.), *Internationalizing Media Studies*. London: Routledge. Retrieved from <http://eprints.lse.ac.uk/52470/>
- Hepp, A., Hjarvard, S., & Lundby, K. (2010). Mediatization – Empirical perspectives: An introduction to a special issue. *Communications*, 35(3). doi:10.1515/comm.2010.012
- Higgins, A. (2012a, January 12). China denounces “Hong Konger” trend. *The Washington Post*. Retrieved from http://www.washingtonpost.com/world/asia_pacific/china-denounces-hong-konger-trend/2012/01/10/gIQAmyvNqP_story_1.html
- Higgins, A. (2012b, January 21). Tycoon prods Taiwan closer to China. *The Washington Post*. Retrieved from http://www.washingtonpost.com/world/tycoon-prods-taiwan-closer-to-china/2012/01/20/gIQAhsWmFQ_story.html
- Hobbs, J. (2009). Cultural diffusion. In *World Regional Geography* (6th ed.). Belmont: Cengage Learning.
- Ho, D. D. (2006). To Protect and Preserve: Resisting the Destroy the Four Olds Campaign, 1966–1967. In J. Esherick, P. Pickowicz, & A. G. Walder (Eds.), *The Chinese cultural revolution as history*. Stanford: Stanford University Press.
- Högselius, P. (2006). National Systems of Innovation and Creative Destruction: A small-country perspective. In T. Kalvet & R. Kattel (Eds.), *Creative Destruction Management: Meeting the Challenges of the Techno-economic Paradigm Shift* (pp. 31–50). Tallinn: PRAXIS.
- Holland, J. H. (1995). *Hidden order: how adaptation builds complexity*. Reading MA: Addison-Wesley.

- Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters in Twitter: A Large Scale Study. In *Proceedings of the Fifth International Conference on Weblogs and Social Media* (pp. 518–521). Barcelona: AAAI.
- Hook, B., & Twitchett, D. C. (1991). *The Cambridge encyclopedia of China*. Cambridge: Cambridge University Press.
- Hopkins, H. (2009, January 23). Britannica 2.0: Wikipedia Gets 97% of Encyclopedia Visits [Web log post]. Retrieved from http://weblogs.hitwise.com/us-heather-hopkins/2009/01/britannica_20_wikipedia_gets_9.html
- Howland, D. R. (1996). *Borders of Chinese Civilization: Geography and History at Empire's End*. Durham NC: Duke University Press.
- Huang, X. (2002). An Introduction to Huaxia Communication Studies (华夏传播研究刍议). *Journalism & Communication* (新闻与传播研究), 4. Retrieved from <http://dspace.xmu.edu.cn/dspace/handle/2288/7838>
- Hughes, C. (2004). Controlling the Internet Architecture Within Greater China. In F. Mengin (Ed.), *Cyber China: Reshaping National Identities in the Age of Information* (pp. 71–90). Basingstoke: Palgrave Macmillan.
- Human Rights Watch. (2006). *"Race to the Bottom": Corporate Complicity in Chinese Internet Censorship*. New York: Human Rights Watch.
- Hussain, S., & Mohan, R. (2008). Localization in Asia Pacific. In F. Librero & P. B. Arinto (Eds.), *Digital Review of Asia Pacific 2007/2008*. Montréal: Orbicom and the International Development Research Centre (IDRC). Retrieved from <http://www.idrc.ca/openebooks/377-5/>
- Huynh, A. T. (2006, May 11). Baidupedia, the Chinese Wikipedia. Retrieved September 24, 2012, from <http://www.dailytech.com/Baidupedia+the+Chinese+Wikipedia/article2272.htm>
- IANA. (2011, January 11). IETF Language Subtag Registry. Retrieved April 11, 2011, from <http://www.iana.org/assignments/language-subtag-registry>
- IDATE. (2011). *World Internet Usage & Markets*. IDATE Consulting and Research. Retrieved from http://www.idate.org/en/Research-store/Collection/Market-Data-Reports_23/World-Internet-Usage-Markets_584.html
- Innis, H. A. (2007). *Empire and Communications*. Lanham MD: Rowman & Littlefield.
- iResearch. (2013). *Sina's Weibo New Advertising Model Marks Monetization of the Long-tail Marketing Strategy* (Press release). iResearch Consulting Group. Retrieved from <http://www.iresearchchina.com/views/4809.html>
- Iskold, A. (2007, August 6). Rethinking "Crossing The Chasm" [Web log post]. Retrieved from http://www.readwriteweb.com/archives/rethinking_crossing_the_chasm.php

- Israel, J. I. (2001). *Radical Enlightenment: Philosophy and the Making of Modernity 1650-1750* (1st ed.). Oxford: Oxford University Press.
- Israel, J. I. (2010). *A revolution of the mind: Radical Enlightenment and the intellectual origins of modern democracy*. Princeton: Princeton University Press.
- Israel, J. I. (2011). *Democratic enlightenment: philosophy, revolution, and human rights 1750-1790*. Oxford: Oxford University Press.
- Jakubowicz, K. (2010). Introduction. Media Systems Research: An Overview. In B. Dobek-Ostrowska, M. Glowacki, K. Jakubowicz, & M. Sükösd (Eds.), *Comparative Media Systems: European and Global Perspectives* (pp. 1–21). Budapest: Central European University Press.
- Jansen, B. J., Brown, A., & Resnick, M. (2007). Factors relating to the decision to click on a sponsored link. *Decision Support Systems*, 44(1), 46–59. doi:10.1016/j.dss.2007.02.009
- Jansen, B. J., & Mullen, T. (2008). Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business*, 6(2), 114–131. doi:10.1504/IJEB.2008.018068
- Jansen, J. (2011). *Understanding Sponsored Search: Core Elements of Keyword Advertising*. Cambridge: Cambridge University Press.
- Jeffres, L. W., Atkin, D. J., Bracken, C. C., & Neuendorf, K. A. (2004). Cosmopolitanism in the Internet Age. *Journal of Computer-Mediated Communication*, 10(1). doi:10.1111/j.1083-6101.2004.tb00227.x
- Jiang, M. (2007). *Citizen Interactions with Chinese Government Networks: Information Technology, Institutions, and Agency* (Ph.D. Dissertation). Purdue University, US.
- Jiang, M., & Akhtar, A. (2011). Peer into the Black Box of Chinese Search Engines: A Comparative Study of Baidu, Google, and Goso. In *9th Chinese Internet Research Conference (CIRC 2011)*. Washington, D.C.: Institute for the Study of Diplomacy. Georgetown University.
- Ji, F. Y. (2004). *Linguistic engineering: language and politics in Mao's china*. Honolulu: University of Hawaii Press.
- Jirik, J. (2004). Television in Greater China: a single geo-cultural region, or an asymmetrically interdependent series of geolinguistic regions? In *Joint International Conference on Chinese Communication*. Shanghai.
- John, N. (2010). *The History of the Internet in Israel: A study in the sociology of technology* (Ph.D. Dissertation). Hebrew University of Jerusalem. Retrieved from <http://www.sociothink.com/>
- Jones, R. (2007, June 26). 96.6% of Wikipedia Pages Rank in Google's Top 10. Retrieved December 2, 2011, from <http://www.thegooglecache.com/white-hat-seo/966-of-wikipedia-pages-rank-in-googles-top-10/>
- Jucquois-Delpierre, M. (2007). Fictional reality or real fiction: how can one decide?: The strengths and weaknesses of information science concepts

- and methods in the media world. *Journal of Information, Communication & Ethics in Society*, 5(2/3), 235–252. doi:10.1080/14616700306488
- Jung, G. (2008). *The Increasing Relevance of Online Marketing*. München: GRIN Verlag.
- Kalish, S., & Lilien, G. L. (1986). A Market Entry Timing Model for New Technologies. *Management Science*, 32(2), 194–205.
- Keane, M. (2007). *Created in China: The Great New Leap Forward*. London: Routledge.
- Keniston, K. (2001). Language, power, and software. In C. Ess & F. Sudweeks (Eds.), *Culture, technology, communication: Towards an intercultural global village* (pp. 283–306). New York: SUNY Press.
- Khanna, A. (2011, October 26). Google drives traffic to Wikipedia, but half of readers look for Wikipedia content — Wikimedia blog [Web log post]. Retrieved from <http://blog.wikimedia.org/2011/10/26/search-and-wikipedia/>
- Kiiski, S., & Pohjola, M. (2002). Cross-country diffusion of the Internet. *Information Economics and Policy*, 14(2), 297–310. doi:10.1016/S0167-6245(01)00071-3
- Kim, B.-K. (2005). Internationalizing the Internet. In *Internationalizing the Internet: the co-evolution of influence and technology* (pp. 1–9). Cheltenham: Edward Elgar.
- King, G., Pan, J., & Roberts, M. (2012). How censorship in China allows government criticism but silences collective expression. In *APSA 2012 Annual Meeting Paper*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2104894
- Kipnis, A. (2006). Suzhi: A Keyword Approach. *The China Quarterly*, 186, 295–313. doi:10.1017/S0305741006000166
- Koike, Y. (2012, 31). Yuriko Koike: China's Soft-Power Offensive in Taiwan. *Straits Times: Global Perspective*. Retrieved from http://www.straitstimes.com/Project_Syndicate/Story/STIStory_761249.html
- Konieczny, P. (2009). Governance, Organization, and Democracy on the Internet: The Iron Law and the Evolution of Wikipedia. *Sociological Forum*, 24(1), 162–192. doi:10.1111/j.1573-7861.2008.01090.x
- Koutsogiannis, D., & Mitsikopoulou, B. (2006). Greeklish and Greekness: Trends and Discourses of “Glocalness.” *Journal of Computer-Mediated Communication*, 9(1), 0–0. doi:10.1111/j.1083-6101.2003.tb00358.x
- Kraidy, M. M. (2011). The Rise of Transnational Media Systems: Implications of Pan-Arab Media for Comparative Research. In D. C. Hallin & P. Mancini (Eds.), *Comparing Media Systems Beyond the Western World* (pp. 177–222). Cambridge: Cambridge University Press.
- Lai, C. P.-Y. (2007). *Media in Hong Kong: press freedom and political change, 1967-2005*. London: Routledge.

- Lai, R., & Rahman, S. (2012). Analytic of China Cyberattack. *The International Journal of Multimedia & Its Applications (IJMA)*, 4(3). doi:10.5121/ijma.2012.4304
- Lall, R. (2014, January 15). Rajiv Lall: AAP and the politics of urbanisation. *Business Standard India*. Retrieved from http://www.business-standard.com/article/opinion/rajiv-lall-aap-and-the-politics-of-urbanisation-114011501284_1.html
- Lam, O. (林藹雲). (2012). Hong Kong: Battle against 50 Cents at Wikipedia [Web log post]. Retrieved from <http://advocacy.globalvoicesonline.org/2012/10/27/hong-kong-battle-against-50-cents-at-wikipedia/>
- Leamer, E. E., & Storper, M. (2001). The Economic Geography of the Internet Age. *Journal of International Business Studies*, 32(4), 641–665. doi:10.1057/palgrave.jibs.84909988
- Lee, C. C. (2001). Rethinking political economy: Implications for media and democracy in Greater China. *Javnost-The Public*, 8(4), 81–102.
- Lee, J. T.-H., Nedilsky, L. V., & Cheung, S.-K. (2012). *China's Rise to Power: Conceptions of State Governance*. Basingstoke: Palgrave Macmillan.
- Lee, L. Z. (2005, November). Chinese Characters: A Quick Social, Political and Linguistic Survey. *CCAPS Translation and Localization*, 18. Retrieved from http://www.ccaps.net/newsletter/10-05/art_1en.htm
- Lee, N. (2012). *Facebook Nation: Total Information Awareness*. Berlin: Springer.
- Lee, P., & Rice, R. (1998). *Telecommunications and Development in China*. New York: Hampton Press.
- Lee, S. O. (2007). *Taiwanese identity construction: A discourse analysis of Chinese textbooks from the 1970s to 2004* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Indiana.
- Lee, X.-Y., & Luo, Z.-C. (2009). Study on Evolving Tendency and Policy Environment of Chinese Language Wikipedia [Zhōngwén wéijī bǎikē yǎnhuà qūshì yǔ zhèngcè huánjìng jiégòu yánjiū 中文维基百科演化趋势与政策环境结构研究]. *Journal of Intelligence(情报杂志)*, 2.
- Leibold, J. (2011). Blogging Alone: China, the Internet, and the Democratic Illusion? *The Journal of Asian Studies, First View*, 1–19. doi:10.1017/S0021911811001550
- Leong, K. Y., Liu, H., & Wu, O. P. (1998). Web Internationalization and Java Keyboard Input Methods. Presented at the INET Conference 1998. Retrieved from http://www.isoc.org/inet98/proceedings/5f/5f_2.htm
- Leung, L. (2009). User-generated content on the internet: an examination of gratifications, civic engagement and psychological empowerment. *New Media & Society*, 11(8), 1327–1347. doi:10.1177/1461444809341264
- Levinson, P. (1997). *The Soft Edge: A Natural History and Future of the Information Revolution*. London: Routledge.

- Liao, H.-T. (2008). A webometric comparison of Chinese Wikipedia and Baidu Baike and its implications for understanding the Chinese-speaking Internet. In *9th annual Internet Research Conference: Rethinking Community, Rethinking Place*. Copenhagen.
- Liao, H.-T. (2009a). Are Chinese characters not modern enough? An essay on their role online. *GLIMPSE: The Art + Science of Seeing*, 2(1), 16–24.
- Liao, H.-T. (2009b). Conflict and Consensus in the Chinese version of Wikipedia. *IEEE Technology and Society Magazine*, 28(2), 49–56.
doi:10.1109/MTS.2009.932799
- Liao, H.-T. (2009c). Special speech zones and diversity in the Chinese-written Internet. In *7th Annual Chinese Internet Research Conference (CIRC 2009)*. Annenberg School for Communication, University of Pennsylvania, Philadelphia, US.
- Liao, H.-T. (2011). *Needing to Have a Voice: Linguistic Grouping in the Digital Networked Environment* (ISD Working Papers in New Diplomacy). Washington, D.C.: Institute for the Study of Diplomacy. Georgetown University. Retrieved from
<http://isd.georgetown.edu/files/Needing%20to%20Have%20a%20Voice.pdf>
- Liao, H.-T. (2013a). How does localization influence online visibility of user-generated encyclopedias? A case study on Chinese-language Search Engine Result Pages (SERPs). In *Proceedings of the 9th International Symposium on Open Collaboration*. New York: ACM.
- Liao, H.-T. (2013b, April 16). How much can one express in 140 characters? Comparison between English and other languages like Chinese [Web log post]. Retrieved from
<http://people.oii.ox.ac.uk/hanteng/2013/04/16/how-much-can-one-express-in-140-characters-comparison-between-english-and-other-languages-like-chinese/>
- Liao, H.-T. (2014). “User Generated Content” (用户生成内容). In L. Cheng (Ed.), *The Internet in China* (First.). Great Barrington: Berkshire Publishing.
- Liao, H.-T., & Petzold, T. (2010). Analysing geo-linguistic dynamics of the World Wide Web: The use of cartograms and network analysis to understand linguistic development in Wikipedia. *Cultural Science*, 3(2).
- Lih, A. (2006, November 13). Chinese Wikipedia’s Surge in Growth [Web log post]. Retrieved from
<http://www.andrewlih.com/blog/2006/11/13/chinese-wikipedias-surge-in-growth/>
- Lih, A. (2009). *The Wikipedia Revolution: How a Bunch of Nobodies Created the World’s Greatest Encyclopedia*. New York: Hyperion.
- Língcái Web Promotion. (2010, June 12). Web Promotion: Using Baidu Baike (网络推广之百度百科推广实战篇). Retrieved February 23, 2013, from
<http://www.admin5.com/article/20100612/241906.shtml>

- Li, R. (1997). China in transition: Nationalism, regionalism and transnationalism. *Contemporary Politics*, 3(4), 365–380. doi:10.1080/13569779708449939
- Li, R. (1999). *Business War in Silicon Valley (guīgǔ shāngzhàn 硅谷商战)*. Beijing: Tsinghua University Press(清华大学出版社).
- Loch, K. D., Straub, D. W., & Kamel, S. (2003). Diffusing the Internet in the Arab world: the role of social norms and technological cultururation. *IEEE Transactions on Engineering Management*, 50(1), 45–63.
- Löfgren, O. (1997). Scenes From a Troubled Marriage Swedish Ethnology and Material Culture Studies. *Journal of Material Culture*, 2(1), 95–113. doi:10.1177/135918359700200105
- Löfgren, O. (2001). The nation as home or motel? Metaphors of media and belonging. *Sociologisk Årbok*, 1(1-34).
- Logan, R. K. (1986). *The alphabet effect: the impact of the phonetic alphabet on the development of western civilization*. New York: St. Martin's.
- Logan, R. K. (2000). *The Sixth Language: Learning a Living in the Internet Age*. Toronto: Stoddart.
- Logan, R. K. (2007). *The extended mind: the emergence of language, the human mind, and culture*. Toronto: University of Toronto Press.
- Long, W. (2006, April 7). Baidu Steals from Chinese Wikipedia [百度染指維基百科] [Web log post]. Retrieved from <http://blog.yam.com/williamlong/article/5910151>
- Lowe, J. (2011). *A Thousand Fibers Connect Us: Wikipedia's Global Reach*. Presented at the WikiSym 2011. Retrieved from <http://blog.wikimedia.org/2011/10/06/a-thousand-fibers-connect-us-wikiviz-winner-visualize-wikipedias-global-reach/>
- Lowther, W., Shih, H., & Chao, V. Y. (2011, May 4). Freedom House lowers Taiwan's press ranking. *Taipei Times*. Taipei. Retrieved from <http://www.taipeitimes.com/News/front/archives/2011/05/04/2003502357>
- Lubrano, A. (1997). *The telegraph how technology innovation caused social change*. New York: Garland.
- Luhmann, N. (1995). *Social Systems* (Stanford.). Stanford: Stanford University Press.
- Lundvall, B.-A., & Borrás, S. (2005). Science, technology and innovation policy. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 599–631). Oxford: Oxford University Press.
- Luo, T., & Li, C. (2010). How to develop public opinion monitoring products? - The modes of operation of the Public Opinion Monitoring Unit, People's Net. [如何开发輿情监测产品?——人民网輿情监测室的运作模式]. *Chinese Journalist (中国记者)*, (6). doi:10.3969/j.issn.1003-1146.2010.06.028
- Luo Z.-C., & Fu Z. (2008). The Impact Analyze of Some External Forces on the Wikipedia Order Process (外部因素对维基百科序化过程的影响分析). *Document, Information & Knowledge (圖書情報知識)*, 123, 28–33.

- Lutz, M. (2009, December 26). Arabic-language Internet is coming of age. *Los Angeles Times*. Retrieved from <http://articles.latimes.com/2009/dec/26/business/la-fi-arabic-internet26-2009dec26>
- Mac an Airchinnigh, M. (2012). Digital Spaces in Popular Culture. Retrieved from <http://www.tara.tcd.ie/jspui/handle/2262/61729>
- MacKinnon, R. (2008). Flatter world and thicker walls? Blogs, censorship and civic discourse in China. *Public Choice*, 134(1), 31–46. doi:10.1007/s11127-007-9199-0
- MacKinnon, S. R. (1997). Toward a History of the Chinese Press in the Republican Period. *Modern China*, 23(1), 3–32. doi:10.1177/009770049702300101
- Ma F. (2008). An introduction: the Order of Information Science [导言: 情报学中的序]. *Document, Information & Knowledge (圖書情報知識)*, 2008(3), 28–33.
- Ma, F. (2012). *The Basic Principles of Network Information Ordering: Web 2.0 Mechanism [网络信息序化原理: Web 2.0 机制]*. Beijing: Science Press (科学出版社).
- Ma, F., & Xia, Y. (2008). Mechanism of ordering in Wikipedia based on the CAS theory (基于 CAS 理论的维基百科序化机制研究). *Library Tribune (图书馆论坛)*, 28(6), 85–92.
- Mahajan, V., Muller, E., & Bass, F. M. (1990). New Product Diffusion Models in Marketing: A Review and Directions for Research. *Journal of Marketing*, 54(1), 1–26. doi:10.2307/1252170
- Malaga, R. A. (2008). Worst practices in search engine optimization. *Communications of the ACM*, 51(12), 147–150. doi:10.1145/1409360.1409388
- Margetts, H. Z., & Escher, T. (2006). Governing from the Centre? Comparing the Nodality of Digital Governments. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1755762
- Marres, N. (2012). The redistribution of methods: on intervention in digital social research, broadly conceived. *The Sociological Review*, 60, 139–165. doi:10.1111/j.1467-954X.2012.02121.x
- Martin, J. D., & El-Toukhy, S. (2011). Blogging for Sovereignty: An analysis of Palestinian Blogs. In T. Dumova & R. Fiordo (Eds.), *Blogging in the Global Society: Cultural, Political and Geographical Aspects* (pp. 148–160). Hershey PA: Idea Group.
- MasPoster. (2009, December 17). Baidu Japan Releases Japanese Input Method Beta. Retrieved March 14, 2010, from <http://eming.com/en/baidu-japan-releases-japanese-input-method-beta/>
- Massa, P. (2011). Social networks of Wikipedia. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (pp. 221–230). New York: ACM. doi:10.1145/1995966.1995996

- Massa, P., & Scrinzi, F. (2012). Manypedia: Comparing Language Points of View of Wikipedia Communities. In *Proceeding WikiSym '12*. doi:10.1145/2462932.2462960
- MaxMind. (2012). MaxMind - GeoLite Country. Retrieved March 26, 2012, from <http://www.maxmind.com/app/geolitecountry>
- Ma, Y. (2009, June 23). The Cultural Significance of "Writing in Simplified Chinese Characters While Maintaining a Reading Knowledge of Orthodox Forms" in Mainland China. Retrieved February 4, 2012, from <http://english.president.gov.tw/Default.aspx?tabid=1136&itemid=20800&rmid=3049>
- Maynard, M., & Tian, Y. (2004). Between global and glocal: content analysis of the Chinese Web Sites of the 100 top global brands. *Public Relations Review*, 30(3), 285–291. doi:10.1016/j.pubrev.2004.04.003
- Mcgrath, C., & Zell, D. (2001). The Future of Innovation Diffusion Research and its Implications for Management: A Conversation with Everett Rogers. *Journal of Management Inquiry*, 10(4), 386–391. doi:10.1177/1056492601104012
- McKenna, M. G., & Naftulin, H. (2000). Challenges in the multicultural HCI development environment. In *CHI '00 extended abstracts* (pp. 362–362). New York: ACM. doi:10.1145/633292.633509
- McLuhan, M. (1964). *Understanding Media: The Extensions of Man* (1st ed.). New York: McGraw-Hill.
- McLuhan, M., & Logan, R. K. (1977). Alphabet, Mother of Invention. *ETC: A Review of General Semantics*, 34(4), 373–83.
- McNichol, T. (2004, January 22). Engineering Google Results to Make a Point. *New York Times*.
- Mengin, F. (2004). The Changing Role of the State in Greater China in the Age of Information. In F. Mengin (Ed.), *Cyber China: Reshaping National Identities in the Age of Information* (1st ed.). New York: Palgrave.
- Meyrowitz, J. (1997). Shifting Worlds of Strangers: Medium Theory and Changes in "Them" Versus "Us"*. *Sociological Inquiry*, 67(1), 59–71. doi:10.1111/j.1475-682X.1997.tb00429.x
- Meyrowitz, J. (2010). Media evolution and cultural change. In J. R. Hall, L. Grindstaff, & M.-C. Lo (Eds.), *Handbook of Cultural Sociology*. London: Routledge.
- Midgley, D. F., & Dowling, G. R. (1978). Innovativeness: The Concept and Its Measurement. *The Journal of Consumer Research*, 4(4), 229–242.
- Mingpao. (2010a, April 22). Addicted to writing Wikipedia, Engineer become fulltime Chinese Wikipedian [Gōngchéngshī xiě shàngyǐn biàn quánzhí wéijī rén 工程師寫上癮 變全職維基人]. *Mingpao.com* (明報). Hong Kong. Retrieved from <http://dailynews.sina.com/bg/chn/chnpolitics/mingpao/20100422/15521377592.html>

- Mingpao. (2010b, April 23). Chinese Wikipedia's Administrator Recall Election failed. Disputes over political censorship over "June Fourth" and "Xi Yang" articles [Wéijī yònghù bàmiǎn zhēngyì guǎnlǐ yuán shībài: hōng zhèngzhì shēnchá, guòlǜ "liùsì" and "xí yáng" 维基用户罢免争议管理员失败 轰政治审查 过滤「六四」「席扬」]. *Mingpao.com* (明報). Hong Kong. Retrieved from <http://life.mingpao.com/cfm/dailynews3b.cfm?File=20100423/nalgh/gha1h.txt>
- Mingpao. (2010c, April 25). Narrowly escaped from recall election, Amin dShizhao: I am not "50 Cent Party", editorial policies unchanged [Xiǎn zāo bàmiǎn wéijī guǎnlǐ yuán: Wǒ fēi wǔmáo biānjí fāngzhēn bù biàn 險遭罷免維基管理員：我非五毛編輯方針不變]. *Mingpao.com* (明報). Hong Kong. Retrieved from <http://life.mingpao.com/cfm/dailynews3b.cfm?File=20100425/nalgh/ghc1.txt>
- Mitter, R. (2008). *Modern China: A Very Short Introduction*. Oxford: Oxford University Press.
- Mok, D., & Wellman, B. (2007). Did distance matter before the Internet?: Interpersonal contact and support in the 1970s. *Social Networks*, 29(3), 430–461. doi:10.1016/j.socnet.2007.01.009
- Mok, D., Wellman, B., & Carrasco, J. (2010). Does Distance Matter in the Age of the Internet? *Urban Studies*, 47(13), 2747–2783. doi:10.1177/0042098010377363
- Moore, G. A. (1999). *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers* (Rev. ed.). New York: HarperBusiness.
- Morley, D., & Robins, K. (1995). *Spaces of Identity: Global Media, Electronic Landscapes and Cultural Boundaries*. London: Routledge.
- Morris, M., & Ogan, C. (2002). The Internet as Mass Medium. In D. McQuail (Ed.), *McQuail's reader in mass communication theory* (pp. 134–145). London: SAGE.
- Mueller, M. L. (2010). *Networks and States: The Global Politics of Internet Governance*. MIT Press.
- Muncaster, P. (2012, May 30). Chinese micro-blogs a hit with police: but Sina users urged to snitch on each other. *The Register*. Retrieved from http://www.theregister.co.uk/2012/05/30/police_china_weibo_censorship/
- Murphy, R. (2004). Turning Peasants into Modern Chinese Citizens: "Population Quality" Discourse, Demographic Transition and Primary Education. *The China Quarterly*, 177, 1–20. doi:10.1017/S0305741004000025
- Myers, G. (2010). *Discourse of Blogs and Wikis*. New York: Continuum.
- NetEase. (2012, April 7). More reliable than Baidu Baike: Chinese Wikipedia [Bǐ bǎidù bǎikē kàopǔ de wéijī bǎikē 比百度百科靠谱的维基百科] [Web log

- post]. Retrieved from
<http://data.163.com/12/0407/03/7UF6PBBD00014MTN.html>
- Netpop Research. (2007). *Chinese consumers have dramatically surpassed Americans in adopting Web 2.0 behavior, relying heavily on social media for guidance in purchase decisions* (Press release). San Francisco. Retrieved from
<http://www.marketingcharts.com/wp/interactive/chinese-surpass-americans-in-web-20-use-2257/>
- Neubig, G., & Duh, K. (2013). How much is said in a tweet? A multilingual, information-theoretic perspective. In *AAAI Spring Symposium on Analyzing Microtext*. Stanford, California. Retrieved from
<http://www.phontron.com/paper/neubig13sam.pdf>
- Nguyen, C. (2011, March). Search Engine Market share by country. Retrieved December 1, 2011, from
<http://www.chandlernguyen.com/2011/03/search-engine-market-share-by-country-mar-2011.html>
- Nickles, D. P. (2003). *Under the Wire: how the telegraph changed diplomacy*. Cambridge: Harvard University Press.
- Nielsen Online. (2008). *Wikipedia U.S. Web Traffic Grows 8,000 Percent In Five Years, Driven By Search*. New York: Nielsen Online. Retrieved from
<http://news.softpedia.com/news/Wikipedia-Traffic-Mostly-from-Google-85703.shtml>
- Nisic, N. (2001, April 28). La Chine, "Le Pays sous le Ciel" [China, "The country under heaven"]. *Le dessous des cartes*. Retrieved from
<http://www.arte.tv/fr/la-chine-le-pays-sous-le-ciel/392,CmC=548402,view=maps.html>
- Nunberg, G. (2002, November 30). Will the Internet Always Speak English? *The American Prospect*. Retrieved from
http://www.prospect.org/cs/articles?article=will_the_internet_always_speak_english
- Nunberg, G. (2003, May 18). As Google Goes, So Goes the Nation. Retrieved October 25, 2009, from
<http://www.nytimes.com/2003/05/18/weekinreview/18NUNB.html>
- Nystedt, D. (2007, August 6). Baidu May Be Worst Wikipedia Copyright Violator. *IDG News Service*. Retrieved from
<http://www.pcworld.com/article/id,135550-page,1/article.html>
- Ogawa, N., Jones, G. W., & Williamson, J. G. (1993). *Human resources in development along the Asia-Pacific Rim*. Oxford: Oxford University Press.
- Oh, H., Curley, S. P., & Subramani, M. R. (2008). The Death of Distance?: The Influence of Computer Mediated Communication on Perceptions of Distance. In *ICIS* (p. 149). Retrieved from
<https://misrc.umn.edu/workshops/2008/fall/Oh.pdf>

- OII. (2007). *Oxford Union Debate: Internet and Democratisation, Part 6*. Oxford. Retrieved from http://webcast.oii.ox.ac.uk/?view=Webcast&ID=20070518_194
- Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Arto, L. (2012). The people's encyclopedia under the gaze of the sages: a systematic review of scholarly research on Wikipedia. doi:10.2139/ssrn.2021326
- Old Geng. (2010, March 26). Baidu Baike and Chinese Wikipedia in Old Geng's eyes [老耿眼中的百度百科与维基百科] [Web log post]. Retrieved from <http://www.laogeng.org/archives/939.html>
- Ong, A. (1997). *Ungrounded Empires: The Cultural Politics of Modern Chinese Transnationalism*. New York: Routledge.
- Ong, R. (2007). *China's Security Interests in the 21st Century*. London: Routledge.
- Ong, W. J. (1982). *Orality and Literacy: The Technologizing of the Word*. London: Methuen.
- Ong, W. J. (1992). Writing is a technology that restructures thought. In P. Downing, S. D. Lima, & M. P. Noonan (Eds.), *The Linguistics of Literacy*. Amsterdam: John Benjamins.
- Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management*, 45(2), 272–279. doi:10.1016/j.ipm.2008.10.001
- Papacharissi, Z. (2002). The virtual sphere The internet as a public sphere. *New Media & Society*, 4(1), 9–27. doi:10.1177/14614440222226244
- Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin.
- Park, H. W., & Thelwall, M. (2003). Hyperlink Analyses of the World Wide Web: A Review. *Journal of Computer-Mediated Communication*, 8(4). doi:10.1111/j.1083-6101.2003.tb00223.x
- Petzold, T., & Liao, H.-T. (2011). Geo-linguistic analysis of the World Wide Web: The use of cartograms and network analysis to understand linguistic development in Wikipedia. In D. Araya, Y. Breindl, & T. J. Houghton (Eds.), *Nexus: New Intersections in Internet Research* (pp. 55–75). New York: Peter Lang.
- Petzold, T., Liao, H.-T., Hartley, J., & Potts, J. (2012). A world map of knowledge in the making: Wikipedia's inter-language linkage as a dependency explorer of global knowledge accumulation. *Leonardo: Art, Science and Technology*, 45(3), 284–284. doi:10.1162/LEON_a_00376
- Phillips, A., & Davis, M. (2009, September). RFC 5646 - Tags for Identifying Languages. Retrieved May 21, 2011, from <http://tools.ietf.org/html/rfc5646>
- Porche, I., Paul, C., York, M., Serena, C. C., & Sollinger, J. M. (2013). *Redefining Information Warfare Boundaries for an Army in a Wireless World*. Santa Monica: Rand Corporation.

- Postman, N. (2006). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. London: Penguin.
- Prescott, M. B. (1997). Understanding the Internet as an innovation. *Industrial Management & Data Systems*, 97(3), 119 – 124.
doi:10.1108/02635579710173185
- Press, L., Burkhart, G., Foster, W., Goodman, S., Wolcott, P., & Woodard, J. (1998). An Internet diffusion framework. *Communications of the ACM*, 41(10), 21–26. doi:10.1145/286238.286242
- Press, L., Foster, W., Wolcott, P., & McHenry, W. (2003). The Internet in India and China. *Information Technologies and International Development*, 1(1), 41–60.
- PricewaterhouseCoopers. (2011). *IAB Internet Advertising Revenue Report*. New York: The Interactive Advertising Bureau. Retrieved from <http://www.iab.net/AdRevenueReport>
- Prigogine, I., & Stengers, I. (1984). *Order out of chaos: man's new dialogue with nature*. New York: Bantam.
- Qiang, X. (2010, July 23). User-generated content online now 50.7% of total. *China Daily*. Beijing. Retrieved from http://www.chinadaily.com.cn/business/2010-07/23/content_11042851.htm
- Qiang, X., & Link, P. (2013, January 5). In China's Cyberspace, Dissent Speaks Code. *Wall Street Journal*. Retrieved from <http://online.wsj.com/news/articles/SB10001424127887323874204578219832868014140>
- Qiu, J. L. (2005). The Internet in China: technologies of freedom in a statist society. In M. Castells (Ed.), *The Network Society: A Cross-Cultural Perspective* (pp. 99–124). Cheltenham, UK: Edward Elgar.
- Reagle, J. (2008). *In good faith: Wikipedia collaboration and the pursuit of the universal encyclopedia* (PhD thesis). New York University, New York.
- Reporters Without Borders. (2007). Cyberdissidents imprisoned to date for their activities on the Internet. Retrieved December 20, 2007, from http://www.rsf.org/rubrique.php3?id_rubrique=119
- Rheingold, H. (2008). Using Participatory Media and Public Voice to Encourage Civic Engagement. In W. L. Bennett (Ed.), *Civic Life Online: Learning How Digital Media Can Engage Youth* (pp. 97–118). Cambridge: MIT Press.
- Riehle, D. (2012, September 28). Definition of Open Collaboration [Web log post]. Retrieved from <http://www.wikisym.org/2012/09/28/definition-of-open-collaboration/>
- Rivière, P. (2010). Alfred Metraux: empiricist and romanticist. In R. Parkin & A. de Sales (Eds.), *Out of the Study and Into the Field: Ethnographic Theory and Practice in French Anthropology* (p. 140). Oxford: Berghahn.

- Roche, D. (2006). Encyclopedias and the diffusion of knowledge. In M. Goldie & R. Wokler (Eds.), *The Cambridge History of Eighteenth-Century Political Thought* (pp. 172–194). Cambridge: Cambridge University Press.
- Rogers, E. M. (2003). *Diffusion of Innovations, 5th Edition*. New York: Free Press.
- Rogers, R. (2012). Mapping and the Politics of Web Space. *Theory, Culture & Society*, 29(4-5), 193–219. doi:10.1177/0263276412450926
- Rogers, R. (2013). *Digital methods*. Cambridge: MIT Press.
- Rowe, W. T. (1990). Review Article : The Public Sphere in Modern China. *Modern China*, 16(3), 309–329. doi:10.1177/009770049001600303
- Rubin, J. (2014, January 2). No perfect candidates [Web log post]. Retrieved from <http://www.washingtonpost.com/blogs/right-turn/wp/2014/01/02/no-perfect-candidates/>
- Russell, J. (2011). Why Yahoo! –not Google– rules Taiwan’s webspace. Retrieved December 1, 2011, from <http://asiancorrespondent.com/55695/focus-on-taiwan-where-yahoo-not-google-rules-the-countrys-webspace/>
- Saint-Simon, H. (1810). *Henri Saint-Simon (1760-1825): Selected Writings on Science, Industry, and social organisation*. (K. Taylor, Ed.). London: Croom Helm.
- Sanchez, N., & Mesa, D. (2011). Language spheres. Retrieved March 29, 2014, from <https://wiki.digitalmethods.net/Dmi/DmiSummer2011LanguageSpheres>
- Sassen, S. (1998). On the Internet and Sovereignty. *Indiana Journal of Global Legal Studies*, 5(2), 545–559.
- Sauper, C., & Barzilay, R. (2009). Automatically generating Wikipedia articles: a structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 208–216). Stroudsburg PA: Association for Computational Linguistics.
- Scammell, M. (2000). The Internet and Civic Engagement: The Age of the Citizen-Consumer. *Political Communication*, 17(4), 351–355. doi:10.1080/10584600050178951
- Schlesinger, P. (2000). The nation and communicative space. In H. Tumber (Ed.), *Media Power, Professionals and Policies* (pp. 99–115). London: Routledge.
- Schlesinger, P. (2001). Communications Theories of Nationalism. In A. S. Leoussi (Ed.), *Encyclopaedia of Nationalism* (pp. 26–30). Edison NJ: Transaction.
- Schneider, N.-C., & Gräf, B. (2011). *Social Dynamics 2.0: Researching Change in Times of Media Convergence: Case Studies from the Middle East and Asia*. Berlin: Frank & Timme.
- Schneider, S. M., & Foot, K. A. (2004). The Web as an Object of Study. *New Media & Society*, 6(1), 114–122. doi:10.1177/1461444804039912

- Schneider, S. M., & Foot, K. A. (2005). Web Sphere Analysis: An Approach to Studying Online Action. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet* (pp. 157–170). Oxford: Berg.
- Schroeder, R., & Ling, R. (2013). Durkheim and Weber on the social implications of new information and communication technologies. *New Media & Society*. doi:10.1177/1461444813495157
- Segev, E. (2008). Search Engines and Power: A Politics of Online (Mis-) Information. *Webology*, 5(2). Retrieved from <http://www.webology.org/2008/v5n2/a54.html>
- SEMPO. (2011). *SEMPO State of Search Marketing Report 2011*. SEMPO Institute. Retrieved from <http://econsultancy.com/uk/reports/sempos-state-of-search>
- Shao, G., Lu, J., & Wu, J. (2012). New media and civic engagement in China: the case of the Xiamen PX event. *China Media Research*, 8(2), 76.
- Sherrod, L. R., Torney-Purta, J., & Flanagan, C. A. (2010). *Handbook of Research on Civic Engagement in Youth*. Hoboken: John Wiley & Sons.
- Sheyholislami, J. (2011). Linguistic Minorities on the Internet. In K. S. Amant & S. Kelsey (Eds.), *Computer-Mediated Communication Across Cultures: International Interactions in Online Environments*. Hershey PA: Idea Group.
- Shigeru, N. (中山茂). (2001). The Digital Revolution and East Asian Science. In A. K. L. Chan, G. K. Clancey, & H.-C. Loy (Eds.), *Historical Perspectives on East Asian Science, Technology, and Medicine*. Singapore: World Scientific.
- Shim, J., & Yang, J. (2009). Why is Wikipedia not more widely accepted in Korea and China? Factors affecting knowledge-sharing adoption. *Decision Line*, 40(2), 12–15.
- Shirky, C. (2010). *Cognitive Surplus: Creativity and Generosity in a Connected Age*. New York: Penguin.
- Siebert, F. S., Peterson, T., & Schramm, W. (1963). *Four theories of the press: the authoritarian, libertarian, social responsibility, and Soviet communist concepts of what the press should be and do*. Urbana: University of Illinois Press.
- Silverwood-Cope, S. (2012, February 8). Wikipedia: Page one of Google UK for 99% of searches [Web log post]. Retrieved from <http://www.intelligentpositioning.com/blog/2012/02/wikipedia-page-one-of-google-uk-for-99-of-searches/>
- Sinclair, J. (1996). Culture and Trade: Some Theoretical and Practical Considerations. In E. G. McAnany & K. T. Wilkinson (Eds.), *Mass media and free trade* (p. 444). Austin: University of Texas Press.
- Sinclair, J., Jacka, E., & Cunningham, S. (1996). *New Patterns in Global Television: Peripheral Vision*. Oxford: Oxford University Press.

- Singh, C., Lehal, G. S., Sengupta, J., Sharma, D. V., & Goyal, V. (2011). *Information Systems for Indian Languages: International Conference, ICISIL 2011, Patiala, India, March 9-11, 2011. Proceedings*. Berlin: Springer.
- Slingshot SEO. (2011). *Google & Bing Click-Through Rates* (White paper). Retrieved from <http://www.slingshotseo.com/resources/white-papers/google-ctr-study/>
- So, A. Y. (2004). One country, three systems? State, nation, and civil society in the making of citizenship in the Chinese triangle of Mainland-Taiwan-Hong Kong. In A. S. M. Ku & N. Pun (Eds.), *Remaking citizenship in Hong Kong: community, nation, and the global city* (pp. 235–253). London: Routledge.
- Sorasak, P., & Kosona, C. (2010). Cambodia. In S. Akhtar & P. Arinto (Eds.), *Digital Review of Asia Pacific 2009-2010*. IDRC.
- Specter, M. (1996, April 14). World, Wide, Web - 3 English Words. *The New York Times*. Retrieved from <http://query.nytimes.com/gst/fullpage.html?res=9E0DE2DA1139F937A25757CoA960958260&pagewanted=all>
- Spindler, S. (2010). *Online Marketing: How to Increase International Sales with Search Engine Optimisation*. München: GRIN Verlag.
- Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer*, 35(3), 107–109. doi:10.1109/2.989940
- StatCounter. (2011). Top 5 Search Engines in China/Hong Kong/Singapore/Taiwan from Nov 2010 to Nov 2011. Retrieved December 1, 2011, from http://gs.statcounter.com/#search_engine-CN-monthly-201011-201111
- Stavridis, J. (2012). *James Stavridis: A Navy Admiral's thoughts on global security*. TEDGlobal 2012. Retrieved from www.ted.com/talks/james_stavridis_how_nato_s_supreme_commander_thinks_about_global_security.html
- Steinmueller, W. E. (2010). Economics of Technology Policy. In *Handbook of the Economics of Innovation* (Vol. 2, pp. 1181–1218). Amsterdam: Elsevier.
- Strang, D., & Tuma, N. B. (1993). Spatial and Temporal Heterogeneity in Diffusion. *American Journal of Sociology*, 99(3), 614–639.
- Sui, C. (2011, May 22). China-Taiwan joint dictionary challenge. *BBC*. Retrieved from <http://www.bbc.co.uk/news/world-asia-pacific-13467598>
- Sum, N.-L. (2004). Cyber-Capitalism and the Remaking of Greater China. In F. Mengin (Ed.), *Cyber China: Reshaping National Identities in the Age of Information* (1st ed., pp. 205–236). Basingstoke: Palgrave Macmillan.
- Sun, H. (2012). *Cross-cultural technology design: creating culture-sensitive technology for local users*. Oxford: Oxford University Press.
- Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton: Princeton University Press.

- Suo, H. (2007). The development of Wiki sites in China: An exploratory study. [中国维基网站发展原因之初探]. In *Harmonious Society, Civic Society, and Mass Media* [和谐社会、公民社会与大众媒介]. Beijing: China University Communication Press [中国传媒大学出版社]. Retrieved from <http://academic.mediachina.net/article.php?id=5476>
- Tai, Z. (2006). *The Internet in China: Cyberspace and Civil Society* (1st ed.). London: Routledge.
- Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, 34(1), 73–81. doi:10.1016/j.socnet.2011.05.006
- Taneja, H., & Wu, A. X. (2013). How Does the Great Firewall of China Affect Online User Behavior? Isolated “Internets” as Culturally Defined Markets on the WWW. Presented at the 11th Annual Chinese Internet Research Conference (CIRC 2013), Oxford, UK. Retrieved from <http://arxiv.org/abs/1305.3311>
- Tan, F. B., Corbett, P. S., & Wong, Y.-Y. (1999). *Information Technology Diffusion in the Asia Pacific: Perspectives on Policy, Electronic Commerce and Education*. Hershey PA: Idea Group.
- Tang, X. (2014). *Annual Report on Development of New Media in China (2014)* 《中国新媒体发展报告》 (No. 5). Beijing, China: Institute of Journalism and Media, Chinese Academy of Social Sciences (CASS). Retrieved from <http://dcf.qingke.cn/content.do?id=20214&copID=800312&channelID=1768&classifiedid=0>
- Tan, Y. F. (2012, January 29). Online public opinion monitoring: an exploratory report [Tànzào wǎngluò yúqíng jiāncè 探照网络舆情监测]. *Caijing* (财经), 2012(312). Retrieved from <http://magazine.caijing.com.cn/2012-01-29/111641737.html>
- Tapscott, D., & Williams, A. D. (2008). *Wikinomics: How Mass Collaboration Changes Everything*. New York: Portfolio.
- Tarde, G. de. (1962). *The laws of imitation*. Gloucester MA: P. Smith.
- Tarde, G. (2010). *Gabriel Tarde On Communication and Social Influence: Selected Papers*. Chicago: University of Chicago Press.
- Tatum, C. (2005). Deconstructing Google bombs. *First Monday*, 10(10). doi:10.5210/fm.v10i10.1287
- Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science*, 34(4), 605–621. doi:10.1177/0165551507087238
- Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1), 1–116. doi:10.2200/Soo176ED1Vo1Y200903ICR004
- Thelwall, M., & Smith, A. (2002). Interlinking between Asia-Pacific University Web sites. *Scientometrics*, 55(3), 363–376. doi:10.1023/A:1020410728852

- Thelwall, M., & Vaughan, L. (2004). Webometrics: An introduction to the special issue. *Journal of the American Society for Information Science and Technology*, 55(14), 1213–1215. doi:10.1002/asi.20076
- Thelwall, M., & Wilkinson, D. (2003). Graph structure in three national academic Webs: Power laws with anomalies. *Journal of the American Society for Information Science and Technology*, 54(8), 706–712. doi:10.1002/asi.10267
- Thomas, H. (2009). *Twitter Marketing: An Hour a Day*. Hoboken: John Wiley & Sons.
- Thussu, D. K. (2009). *Internationalizing Media Studies*. Abingdon: Routledge.
- Tolosa, G., Bordignon, F., Baeza-Yates, R., & Castillo, C. (2007). Characterization of the Argentinian Web. *Cybermetrics*, 11(1), Paper 3.
- Totok, A. (2009). *Modern Internet Services: Exploiting Service Usage Information for Optimizing Service Management*. Saarbrücken: VDM Verlag.
- Unger, J., & Barne, G. (1996). *Chinese Nationalism*. Armonk: M.E. Sharpe.
- Unicode Inc. (2011a). CLDR Project - Unicode Common Locale Data Repository. Retrieved June 13, 2011, from <http://cldr.unicode.org/>
- Unicode Inc. (2011b). Unicode Standard. Retrieved March 4, 2011, from <http://unicode.org/standard/standard.html>
- Unicode Inc. (2011c, April 27). The Unicode Consortium Members. Retrieved May 3, 2011, from <http://www.unicode.org/consortium/memblogo.html>
- "Unicode". *The Universal Telegraphic Phrase-Book. A code of cypher words for commercial, domestic and familiar phrases in ordinary use in inland and foreign telegrams.* (1886) (2nd ed.). London: Cassell & Company.
- United States Congress. (1981). *Computer-based national information systems : technology and public policy issues*. Washington DC: DIANE.
- US Army. (2009). *Tactics in Counterinsurgency FM 3-24.2*. Headquarters, Department of the Army Head.
- User:430072. (2006, 2012). Talk: Baidu Baike [Talk:百度百科] [Discussion page]. Retrieved November 21, 2012, from <http://zh.wikipedia.org/wiki/Talk:%E7%99%BE%E5%BA%A6%E7%99%BE%E7%A7%91>
- User:Emijrp. (2012, December 4). List of Wikipedians by number of edits [Document page]. Retrieved December 5, 2012, from http://zh.wikipedia.org/w/index.php?title=User:Emijrp/List_of_Wikipedians_by_number_of_edits&oldid=22141797
- User:Gāoyúnhuimǐn[高云慧敏]. (2011, April 3). Fifth Year Anniversary of Baidu Baike: My five years with Baidu Baike [【百科五周年】走过百科这五年] [Web log post]. Retrieved from <http://hi.baidu.com/bwrclldvkbopzd/item/742f6031f914a88af5e4adc2>
- User:Gāoyúnhuimǐn[高云慧敏]. (2012, March 10). Survey: What needs to be improved with Baidu Baike? [请问大家: 如果你认为百度百科不好或不够

- 好, 不好在哪?]. *Baidu Baike-Baidu Tieba* [百度百科吧 百度贴吧]. Online forum comment. Retrieved from <http://tieba.baidu.com/p/1021698407>
- User:ljq513. (2012, July 22). How deep the resentment is this: "Any use of this image by Baidu Baike is not permitted" [这怨念有多深啊: "本图片不允许百度百科使用"]. Retrieved November 20, 2012, from <http://tieba.baidu.com/p/1742518344>
- User:Lorenzarius, & other contributors. (2010). Wikipedia:Aviod-Regional-Centrism[维基百科:避免地域中心] [Official policy page]. Retrieved December 20, 2007, from <http://zh.wikipedia.org/wiki/Wikipedia:%E9%81%BF%E5%85%8D%E5%9C%B0%E5%9F%9F%E4%B8%AD%E5%BF%83>
- User:Shizhao. (2013, February 11). Editing difference on Zhonghua minzu (中华民族). Retrieved March 20, 2013, from <http://zh.wikipedia.org/w/index.php?title=%E4%B8%AD%E5%8D%8E%E6%B0%91%E6%97%8F&diff=24961369&oldid=15170027>
- User:Stevenfruitsmaak. (2006, 12). File:Belgium provinces regions striped.png. In *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc. Retrieved from http://en.wikipedia.org/w/index.php?title=Communities,_regions_and_language_areas_of_Belgium&oldid=490470967
- User:Wǒnǎiyěyúnhè(我乃野云鹤). (2012, November 13). Editing difference on Zhonghua minzu (中华民族). Retrieved March 20, 2013, from <http://baike.baidu.com/diff/?vid1=36658111&vid2=36183173>
- Van Kemenade, W. (2010). *China, Hong Kong, Taiwan, Inc.: The Dynamics of a New Empire*. New York: Random House.
- Varian, H. R. (2007). The Economics of Internet Search. Presented at the Angelo Costa lecture, Rome. Retrieved from <http://people.ischool.berkeley.edu/~hal/Papers/2007/costa-lecture.pdf>
- Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, 40(4), 693–707. doi:10.1016/S0306-4573(03)00063-3
- Vaughan, L., & Zhang, Y. (2007). Equal Representation by Search Engines? A Comparison of Websites across Countries and Domains. *Journal of Computer-Mediated Communication*, 12(3). doi:10.1111/j.1083-6101.2007.00355.x
- Volkmer, K. (2011). How Far Can Media Systems Travel? Applying Hallin and Mancini's Comparative Framework outside the Western World. In D. C. Hallin & P. Mancini (Eds.), *Comparing Media Systems Beyond the Western World* (pp. 224–244). Cambridge: Cambridge University Press.
- Von Hippel, E. (2005). *Democratizing Innovation*. Cambridge: MIT Press.
- Wahba, K. M., Taha, Z. A., & England, L. (2013). *Handbook for Arabic Language Teaching Professionals in the 21st Century*. London: Routledge.

- Wang, A. (2013, December 4). Using social media to engage Wikipedia readers and editors in China [Web log post]. Retrieved from <https://blog.wikimedia.org/2013/12/04/using-social-media-to-engage-editors-in-china/>
- Wang, G. (1991). *The Chineseness of China: selected essays*. Oxford: Oxford University Press.
- Wang, G. (1993). Greater China and the Chinese Overseas. *The China Quarterly*, 136, 926–948. doi:10.1017/S0305741000032392
- Wang, G. (1996a). Openness and Nationalism: Outside the Revolution. In J. Unger & G. Barne (Eds.), *Chinese Nationalism*. Armonk: M.E. Sharpe.
- Wang, G. (1996b). *The Revival of Chinese Nationalism*. Leiden: International Institute for Asian Studies.
- Wang, G. (2004). Maritime China in transition. In G. Wang & C.-K. Ng (Eds.), *Maritime China in transition 1750-1850*. Wiesbaden: Harrassowitz Verlag.
- Wang, J. (1996). *High Culture Fever: Politics, Aesthetics, and Ideology in Deng's China*. Berkeley: University of California Press.
- Warncke-Wang, M., Uduwage, A., Dong, Z., & Riedl, J. (2012). In Search of the Ur-Wikipedia: Universality, Similarity, and Translation in the Wikipedia Inter-language Link Network. In *Proceeding WikiSym '12*. doi:10.1145/2462932.2462959
- Webber, M., Lutz, J. M., & Brown, L. A. (2006). Classics in human geography revisited: Brown, L.A. 1981: Innovation diffusion: a new perspective. London: Methuen. *Progress in Human Geography*, 30(4), 487–489. doi:10.1191/0309132506ph620xx
- Wellman, B., Quan-Haase, A., Boase, J., Chen, W., Hampton, K., Díaz, I., & Miyata, K. (2003). The Social Affordances of the Internet for Networked Individualism. *Journal of Computer-Mediated Communication*, 8(3). doi:10.1111/j.1083-6101.2003.tb00216.x
- Wells, H. G. (1937, August). World Brain: The Idea of a Permanent World Encyclopaedia. In *new Encyclopédie Française*. Retrieved from https://sherlock.ischool.berkeley.edu/wells/world_brain.html
- Wessler, H., Peters, B., Brüggemann, M., Kleinen-von KönigsLöw, K., & Sifft, S. (2008). *Transnationalization of public spheres*. Basingstoke: Palgrave Macmillan. Retrieved from http://www.sfb597.uni-bremen.de/download/en/publikationen/sfbReihen_flyer_The_Transnationalization_of_Public_Spheres.pdf
- Wiener, N. (1948). *Cybernetics Or Control and Communication in the Animal and the Machine*. Cambridge: MIT Press.
- Wikimedia Foundation. (2012a). Terms of use - Meta [使用條款 - Meta]. Retrieved November 20, 2012, from http://meta.wikimedia.org/wiki/Terms_of_use/zh-hant
- Wikimedia Foundation. (2012b). Wikipedias in multiple writing systems. Retrieved December 4, 2012, from

- http://meta.wikimedia.org/wiki/Wikipedias_in_multiple_writing_systems
- Wikimedia Meta. (2012, March 15). Global Development [Web log post]. Retrieved from https://meta.wikimedia.org/wiki/Global_Development
- Wilkinson, D., & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison: Trending Twitter Topics in English: An International Comparison. *Journal of the American Society for Information Science and Technology*, 63(8), 1631–1646. doi:10.1002/asi.22713
- Wimmer, A., & Schiller, N. G. (2002). Methodological nationalism and beyond: nation–state building, migration and the social sciences. *Global Networks*, 2(4), 301–334. doi:10.1111/1471-0374.00043
- Windrum, P. (2004). Leveraging technological externalities in complex technologies: Microsoft's exploitation of standards in the browser wars. *Research Policy*, 33(3), 385–394. doi:10.1016/j.respol.2003.09.002
- Winseck, D. R. (2007). *Communication and empire: media, markets, and globalization, 1860-1930*. Durham: Duke University Press.
- Withers, C. W. J. (2008). *Placing the Enlightenment: Thinking Geographically about the Age of Reason*. Chicago: University of Chicago Press.
- Wolcott, P., Press, L., Mchenry, W., Goodman, S., & Foster, W. (2001). A framework for assessing the global diffusion of the Internet. *Journal of the Association for Information Systems*, 2, 1–50.
- Wong, K.-F., Li, W., & Xu, R. (2009). *Introduction to Chinese Natural Language Processing*. San Rafael: Morgan & Claypool.
- Woodside, A. (2006). Mandarin Management Theorists? In *Lost modernities*. Cambridge: Harvard University Press.
- Woo, E. (2007, November 13). Baidu's Censored Answer to Wikipedia. *BusinessWeek*. Retrieved from http://www.businessweek.com/globalbiz/content/nov2007/gb20071113_725400.htm
- Wu, J. (2007). World without borders: wildlife trade on the Chinese-language Internet. *Traffic Bulletin*, 21(2), 75–84.
- Xiao, Q. (2009, April 30). Baidu's Internal Monitoring and Censorship Document Leaked (1) (Updated). Retrieved November 20, 2012, from <http://chinadigitaltimes.net/2009/04/baidus-internal-monitoring-and-censorship-document-leaked/>
- Xing, J., Ng, P.-S., & Cheng, C. (2012). *General Education and the Development of Global Citizenship in Hong Kong, Taiwan and Mainland China: Not Merely Icing on the Cake*. London: Routledge.
- Xu, Z. (2009). *Who am I? Who are the others? [Wo zhe yu ta zhe: Zhongguo li shi shang de nei wai fen ji] (我者與他者：中國歷史上的內外分際)*. Hong Kong: Chinese University Press.

- Yang, G. (2003). The Internet and the Rise of a Transnational Chinese Cultural Sphere. *Media Culture Society*, 25(4), 469–490.
doi:10.1177/01634437030254003
- Yang, G. (2011). Internet and Civil Society. In W. S. Tay & A. Y. So (Eds.), *Handbook of Contemporary China*. Singapore: World Scientific.
- Yang, G. (2012). A Chinese Internet? History, practice, and globalization. *Chinese Journal of Communication*, 5(1), 49–54.
doi:10.1080/17544750.2011.647744
- Yang, K. C. C. (2007). A comparative study of Internet regulatory policies in the Greater China Region: Emerging regulatory models and issues in China, Hong-Kong SAR, and Taiwan. *Telematics and Informatics*, 24(1), 30–40.
doi:10.1016/j.tele.2005.12.001
- Yang, M.-J. (2006a, May 30). Baidu launched Wikipaida, China style - News series on Wikipedias III [百度推出中國式維基百科-維基百科系列報導(三)]. Retrieved September 24, 2012, from <http://www.ptt.cc/bbs/Wikipedia/M.1149214602.A.496.html>
- Yang, M.-J. (2006b, May 30). China loses online high grounds by blocking Wikipedia - News series on Wikipedias II [中國自喪重要網路陣地 -維基百科系列報導(二)]. Retrieved September 24, 2012, from <http://www.ptt.cc/bbs/Wikipedia/M.1149214602.A.496.html>
- Yang, Y. (2011, February 25). China's "Wikipedia" Submits Complaint about Baidu. *Economic Observer News*, 508, 28.
- Yang, Y., & Xie, L. (2008). Subword Latent Semantic Analysis for Texttiling-Based Automatic Story Segmentation of Chinese Broadcast News. In *6th International Symposium on Chinese Spoken Language Processing, 2008*. (pp. 1–4). doi:10.1109/CHINSL.2008.ECP.101
- Yin, P.-L. (2006, April 10). Lessons from the Browser Wars — Q&A with Pai-Ling Yin. Retrieved from <http://hbswk.hbs.edu/item/5288.html>
- Young, R. D. (2011, August 10). Top Google Ranking Captures 18.2% of Clicks. Retrieved December 2, 2011, from <http://searchenginewatch.com/article/2100616/Top-Google-Ranking-Captures-18.2-of-Clicks-Study>
- Yunker, J. (2002). *Beyond Borders: Web Globalization Strategies*. Indianapolis: New Riders.
- Yu Z.-T., Fan X.-Z., Guo J.-Y., & Geng Z.-M. (2006). Answer Extracting for Chinese Question-Answering System Based on Latent Semantic Analysis. *Chinese Journal of Computers (计算机学报)*, 29(10), 1889.
- Zhai, Q. (2009). 1959: Preventing Peaceful Evolution. *China Heritage Quarterly*, (18). Retrieved from http://chinaheritagenewsletter.anu.edu.au/features.php?searchterm=018_1959preventingpeace.inc&issue=018
- Zhang, J.-C., & Qin, Y. (2012). Impact of Internet Use on Civic Engagement in Chinese Rural Areas: A Preliminary Research. In A. Manoharan & M.

- Holzer (Eds.), *Active Citizen Participation in E-Government*. Hershey: IGI Global. Retrieved from <http://www.igi-global.com/chapter/impact-internet-use-civic-engagement/63376>
- Zhang, X., & Zhu, F. (2011). Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. *American Economic Review*, 101(4), 1601–1615. doi:10.1257/aer.101.4.1601
- Zhao, S. (2004). *A Nation-State by Construction: Dynamics of Modern Chinese Nationalism*. Stanford: Stanford University Press.
- Zhao, S. (2008). Chinese character modernisation in the Digital Era: A historical perspective. In R. B. Kaplan (Ed.), *Language Planning and Policy in Asia: Japan, Nepal, Taiwan and Chinese characters*. Bristol: Multilingual Matters.
- Zhao, S., & Baldauf, R. B. (2007a). Language Planning, Naming and Character Use in China. *Current Issues in Language Planning*, 8(3), 283. doi:10.2167/cilp117.0
- Zhao, S., & Baldauf, R. B. J. (2007b). *Planning Chinese Characters: Reaction, Evolution or Revolution?*. Berlin: Springer.
- Zhao, Y. (2011). Understanding China's Media System in a World Historical Context. In D. C. Hallin & P. Mancini (Eds.), *Comparing Media Systems Beyond the Western World* (pp. 143–175). Cambridge: Cambridge University Press.
- Zheng, Y. (2007). The Internet, Civic Engagement, and Public Distrust. In *Technological Empowerment: The Internet, State, and Society in China*. Stanford: Stanford University Press.
- Zheng, Y. (2011). *China on the Sea: How the Maritime World Shaped Modern China*. Leiden: Brill.
- Zhonghua minzu (zhōnghuá mínzú 中华民族). (2013, March 20). Retrieved March 20, 2013, from <http://baike.baidu.com/history/id=12371>
- Zhou, Y. (2005). *Historicizing Online Politics: Telegraphy, the Internet, and Political Participation in China*. Stanford: Stanford University Press.
- Zhu, J. H. (1997). Growth, competition and survival: forecasting of Chinese language television based on the S-curve model. In W. Xie, X. Cai, H. Huang, & Z. Shi (Eds.), *Perspectives on Chinese-language television* (pp. 41–50). Guangzhou: Flower City Press.
- Zhu, J. J. H., & Wang, E. (2005). Diffusion, use, and effect of the internet in China. *Communications of the ACM*, 48(4), 49. doi:10.1145/1053291.1053317
- Zhu, J. J. H., & Zhou, H. (2002). Information Accessibility, User Sophistication, and Source Credibility: The Impact of the Internet on Value Orientations in Mainland China. *Journal of Computer-Mediated Communication*, 7(2). Retrieved from <http://jcmc.indiana.edu/vol7/issue2/china.html>

- Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., & Wallach, D. S. (2012). Tracking and Quantifying Censorship on a Chinese Microblogging Site. *arXiv:1211.6166*. Retrieved from <http://arxiv.org/abs/1211.6166>
- zhWP. (2008a). Chinese Wikipedia [Zhōngwén wéijī bǎikē 中文維基百科]. In *Chinese Wikipedia*. Retrieved from <http://zh.wikipedia.org/wiki/%E4%B8%AD%E6%96%87%E7%BB%B4%E5%9F%BA%E7%99%BE%E7%A7%91>
- zhWP. (2008b). Wikipedia:Wikimania 2005 Frankfurt [維基百科:2005 年維基媒體國際大會] [Document page]. Retrieved March 14, 2008, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:2005%E5%B9%B4%E7%B6%AD%E5%9F%BA%E5%AA%92%E9%AB%94%E5%9C%8B%E9%A%9B%E5%A4%A7%E6%9C%83&variant=zh-tw>
- zhWP. (2009). Wikipedia:About Chinese Wikipedia [維基百科:关于中文維基百科] [Document page]. Retrieved February 11, 2009, from <http://zh.wikipedia.org/wiki/Wikipedia:%E5%85%B3%E4%BA%8E%E4%B8%AD%E6%96%87%E7%BB%B4%E5%9F%BA%E7%99%BE%E7%A7%91>
- zhWP. (2012a). Wikipedia:Administrators [維基百科:管理员] [Official policy page]. Retrieved October 7, 2012, from <http://zh.wikipedia.org/wiki/Wikipedia:%E7%AE%A1%E7%90%86%E5%91%98>
- zhWP. (2012b). Wikipedia:Articles for deletion-copyright violation[維基百科:頁面存廢討論/疑似侵權] [Discussion page]. Retrieved November 20, 2012, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:%E9%A0%81%E9%9D%A2%E5%AD%98%E5%BB%A2%E8%A8%E8%AB%96/%E7%96%91%E4%BC%BC%E4%BE%B5%E6%AC%8A&oldid=23762964>
- zhWP. (2012c). Wikipedia:Baidu Baike violates Wikipedia's copyright [維基百科:百度百科對維基百科的侵權] [Discussion page]. Retrieved November 20, 2012, from <http://zh.wikipedia.org/zh-hant/Wikipedia:%E7%99%BE%E5%BA%A6%E7%99%BE%E7%A7%91%E5%B0%8D%E7%B6%AD%E5%9F%BA%E7%99%BE%E7%A7%91%E7%9A%84%E4%BE%B5%E6%AC%8A>
- zhWP. (2012d). Wikipedia:Edit war(維基百科:編輯戰). Retrieved December 6, 2012, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:%E7%B7%A8%E8%BC%AF%E6%88%Bo&oldid=23646608>
- zhWP. (2012e). Wikipedia:Taskforce for saving new contribution from copyright violation [維基百科:侵权拯救工作小组] [Discussion page]. Retrieved November 20, 2012, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:%E4%BE%B5%E6%9D%83%E6%8B%AF%E6%95%91%E5%B7%A5%E4%BD%9C%E5%B0%8F%E7%BB%84&oldid=23675358>

- zhWP. (2012f). Wikipedia:What Wikipedia is not [維基百科:維基百科不是什麼] [Official policy page]. Retrieved February 21, 2013, from <http://zh.wikipedia.org/wiki/Wikipedia:維基百科不是什麼>
- zhWP. (2012g). Wikipedia:Wheel_war(維基百科:車輪戰) [Official policy page]. Retrieved December 6, 2012, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:%E8%BB%8A%E8%BC%AA%E6%88%Bo&oldid=23171133>
- zhWP. (2012h). Wikipedia:Wikipedia Political edit war [維基百科:維基政治編輯戰] [Document page]. Retrieved December 6, 2012, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:%E7%BB%B4%E5%9F%BA%E6%94%BF%E6%B2%BB%E7%BC%96%E8%BE%91%E6%88%98>
- zhWP. (2012i, November 27). Wikipedia:Vandalism(維基百科:破壞) [Official policy page]. Retrieved December 4, 2012, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:%E7%A0%B4%E5%9D%8F&oldid=23708329>
- zhWP. (2012j, November 28). Wikipedia:Article_titles [Wikipedia:命名常规] [Official policy page]. Retrieved December 4, 2012, from <http://zh.wikipedia.org/w/index.php?title=Wikipedia:%E5%91%BD%E5%90%8D%E5%B8%B8%E8%A7%84&oldid=23852853>
- zhWP. (2012k, November 28). Wikipedia talk:Processing Traditional-Simplified/Archive6 (維基百科討論區:繁簡处理/档案 6). In *Chinese Wikipedia*. Retrieved from http://zh.wikipedia.org/w/index.php?title=Wikipedia_talk:%E7%B9%81%E7%AE%80%E5%A4%84%E7%90%86/%E6%A1%A3%E6%A1%886&oldid=11582876
- zhWP. (2013a). Talk:Zhonghua_minzu [Talk:中华民族]. Retrieved March 20, 2013, from <http://zh.wikipedia.org/w/index.php?title=Talk:%E4%B8%AD%E5%8D%8E%E6%Bo%91%E6%97%8F&oldid=25122836>
- zhWP. (2013b). Zhonghua minzu (中华民族) editing history. Retrieved March 20, 2013, from <http://zh.wikipedia.org/w/index.php?title=%E4%B8%AD%E5%8D%8E%E6%Bo%91%E6%97%8F&offset=&limit=500&action=history>
- Zuckerman, E. (2013). *Rewire: Digital Cosmopolitans in the Age of Connection* (1st ed.). New York: W. W. Norton & Company.
- Zuo, X. (2006). When the Internet is no longer the tool of the Western “Peaceful Evolution” (Dāng hùliánwǎng bù zài shì xīfāng “héping yǎnbiàn” de gōngjù 当互联网不再是西方“和平演变”的工具). *Network Security Technology & Application* (网络安全技术与应用), 4, 8–10.