Mining the Automotive Industry: A Network Analysis of Corporate Positioning and Technological Trends

Niklas Stoehr[1]*, Fabian Braesemann[2]†, Shi Zhou[1]

[1] University College London, Department of Computer Science, UK
[2] University of Oxford, Saïd Business School, Oxford OX1 1HP, UK
f.braesemann@sbs.ox.ac.uk

Abstract. The digital transformation is driving revolutionary innovations and new market entrants threaten established sectors of the economy such as the automotive industry. Following the need for monitoring shifting industries, we present a network-centred analysis of car manufacturer web pages. Solely exploiting publicly-available information, we construct large networks from web pages and hyperlinks. The network properties disclose the internal corporate positioning of the three largest automotive manufacturers, Toyota, Volkswagen and Hyundai with respect to innovative trends and their international outlook. We tag web pages concerned with topics like e-mobility & environment or autonomous driving, and investigate their relevance in the network. Toyota and Hyundai are concerned with e-mobility throughout large parts of their web page network; Volkswagen devotes more specialized sections to it, but reveals a strong focus on autonomous driving. Sentiment analysis on individual web pages uncovers a relationship between page linking and use of positive language, particularly with respect to innovative trends. Web pages of the same country domain form clusters of different size in the network that reveal strong correlations with sales market orientation. Our approach is highly transparent, reproducible and data driven, and could be used to gain complementary insights into innovative strategies of firms and competitive landscapes.

Keywords: Automotive Industry, Network Analysis, Complex Networks, Digitization, Web Page Mining, Competition

Code and Data: `www.github.com/Braesemann/MiningAutomotiveIndustry`

1 Introduction

1.1 Motivation

Environmental change and the ongoing digitization cause large-scale transformations in the economy, as boundaries of production, distribution and consumption are reshaped [1–4]. Industries are impacted differently depending on factors

*ORCID-ID: 0000-0003-2867-0236
†ORCID-ID: 0000-0002-7671-1920

like contribution to greenhouse gas emissions, automation capabilities, customer proximity, and labour complexity [2]. One of the biggest, yet most strongly affected industries is the automotive industry [3, 5–7].

According to estimates of the International Organization of Motor Vehicle Manufacturers (OICA), more than 5 % of the world's total manufacturing employment is directly involved in the production of vehicles and parts [8]. However, the economic importance of the automotive industry reaches far beyond that. Many of the most groundbreaking innovations of the 20th-century, mass production, just-in-time and multi-divisional business organization originated in automotive companies and left a recognizable geographical footprint [9].

The need to develop environmentally sustainable mobility solutions and digitization will drive further innovation in fields relevant for the automotive industry such as e-mobility, connected infrastructure, and autonomous driving. They provide unparalleled opportunities for value creation, but also major risks. Online programming platforms provide highly specialised knowledge and allow users without much previous know-how to develop innovative digital technologies [10–12]. With the emergence of such novel digital technologies, new market participants might introduce innovative services, which could threaten the established ecosystem of the automotive industry [6]. Such radical shifts are creating the need for analytic feedback and status reports. Not only executive boards, policy-makers and investors, but more importantly, millions of employees are interested in a successful and sustainable transformation and therefore rely on objective assessments of car manufacturers' positioning with regards to novel developments and technologies.

1.2 Contributions

In this paper, we provide a new perspective on the possibilities for analyzing shifting industries. Using only publicly available data, we present a network-centered approach that allows for conclusions on innovative orientation and international outlook of the world's leading car manufacturers, Toyota, Volkswagen, and Hyundai.[3]

We exploit the web page structure of the firms to obtain complex networks, with web pages considered as nodes connected by hyperlinks. To gain insights into the companies' orientation with respect to innovative trends, we analyse the shape and properties of the web page networks and we tag individual web pages that mention relevant keywords to observe their position in the network.[4] As a result, we identify commonalities and differences between the car manufacturers in

[3] These are the three biggest automotive manufacturers by production numbers in 2016: Toyota (10.2 mio. vehicles), the Volkswagen Group (10.1 mio cars), and Hyundai with 7.9 mio. units [13].

[4] For this part of the analysis, which hinges on the identification of specific keywords, we analyze the US domains of the company websites (`www.toyota.com`, `www.vw.com`, `www.hyundaiusa.com`), as the United States is the second largest global car market behind China, and the largest English-language car market.

the positioning of pages with innovative keywords and their network centrality: pages dealing with innovative topics tend to be higher ranked.

Furthermore, we perform sentiment analysis on the content of each web page in the network to understand the use of positive language as a means of customer communication. Evidently, well connected pages are more likely to show a positive sentiment than peripheral ones. Equally, pages focusing on innovative topics are characterised by a positive sentiment. Additionally, we analyze the prevalence of different country domains and target markets throughout the network.[5] We find that the leading car manufacturers focus their web presence on a limited number of large national markets, with their home market having a disproportional extensive web page network.

Our approach uses only publicly available information: content directly published by the companies and structural meta-information available on the web. It is, thus, highly transparent and can be reproduced to monitor company web pages and industries to gain insights into their innovative and market strategies.

2 Related Work

This work is related to two strands of literature: studies on the transformation of the automotive industry and network analyses of web pages.

Change in the automotive industry has been intensively investigated in past studies [9, 14–17], focusing on changing geographies and challenges of the technological transformation, and on technical, environmental, and management implications. Moreover, network analysis has been applied in the field, mainly to understand Supply-Chain systems in the automotive industry [18, 19].

While web page networks of car manufacturers have, to our knowledge, not been investigated previously, similar approaches were applied in tourism and e-commerce. For instance, Baggio et al. investigate the links between tourism destination websites to find the statistical characteristics of the underlying graph [20], and Wang et al. establish principles of an improved link structure for a hypothetical e-supermarket website [21].

Additionally, many pioneering studies in network science have focused on the world wide web: in 2000, Broder et al. discovered the bow tie structure of the world wide web [22], and Meusel et al. aimed to understand it's growth mechanism in constructing a giant network of large parts of the world wide web [23]. Barabási and Albert used web data to demonstrate their theory on the emergence of scaling in random networks [24], and the PageRank algorithm was introduced for ranking web pages in a directed graph of the world wide web [25].

The content and structure of web pages has previously been investigated from the perspective of search engine optimization [26], data retrieval [27], website

[5] For the international comparison, we use the manufacturers' international web pages (`www.toyota-global.com`, `www.volkswagen.com`, `www.hyundai.com/worldwide`), as a starting point for the data collection.

design [28], and web navigation [29].

Our approach builds on the previously mentioned studies, but assumes a different perspective. Instead of analysing content and structure of web pages to optimise properties of individual websites, we explore the potential of applying network analysis of web pages to gain insights into business models and innovative orientation of companies and industries.

3 Methodology

Our approach consists of four steps. First, we apply web-crawling on the company web pages in order to obtain the network data. Secondly, we visualize the networks and derive properties and centrality measures from the data. Thirdly, we search for relevant keywords on innovative topics and apply sentiment analysis to the web pages. Lastly, we compare the manufacturers with respect to their international orientation in 25 national markets.

After retrieving the network data, we turn towards a comprehensive analysis of link structure and node meta-information, using a number of tools. For visualization and exploration, we use the graph visualization tools *Gephi* [30] and *Graphviz* [31]. The Python package *NetworkX* [32] is used for obtaining the network and node properties. The search engine optimization software *Screaming Frog* and *SEO Powersuite* are employed to get additional node information.

To obtain the network data, we implement a crawler able to retrieve an extensive number of pages from the web. For this purpose, we use the *Python* package *Beautiful Soup* for HTML and XML parsing. Essentially, the crawler starts at a single initial web page, where it retrieves the web page including all it's hyperlinks. Following the breadth-first paradigm, it then visits all the web pages that the start page links to and stores them as nodes connected via the hyperlinks (links). The crawler then repeats this process on the second level of web pages before it goes on to the next level. We terminate the crawler after the hyperlinks on the 6th level have been added to the network (Fig. 1A). The data was collected in December 2018.

Based on the key themes identified in the literature [3, 5–7], we establish three major innovative trends affecting the automotive industry: (1) 'e-mobility and environment', (2) 'autonomous driving and artificial intelligence', and (3) 'connectivity and shared mobility'. These trends are identified in the textual components[6] of each web page by counting relevant keywords.[7]

[6] The textual components are all HTML tags (predominantly "title" and "body") of a web page. Specifically excluded are the tags "script", "style", "head", "[document]". This way, we only include textual components visible to the user.

[7] The list of keywords has been created prior to looking at any company website, based only on the qualitative definitions in the literature. The respective keywords are:

E-mobility & environment: e-mobility, battery, environment, biological, eco, ecological, electric, hybrid, environment, environmental-friendly;

For the sentiment analysis, we use the natural language processing toolbox *TextBlob* [33] and analyze the textual components of each individual web page with regards to their polarity [34] on a continuous, symmetric sentiment scale ranging from -1 (negative) to +1 (positive).

To capture the international orientation of the car manufacturers, we examine the peculiarity and prevalence of web pages of different country domains, starting from the international landing page of the three firms. The domain affiliation is derived from the country tag, e.g '.ca' for Canada, '.uk' for United Kingdom etc. The number of web pages per country domain are then compared with the size of the national market in terms of car sales and registrations [35].

4 Results

4.1 Network Structure

The obtained web page networks of the US domains of Hyundai, Toyota, and Volkswagen show distinct structures (Fig. 1B). In the Toyota and Volkswagen networks, several navigational pages connect the major elements of the websites, while the web pages of Hyundai fan out into only two major components.

The networks also differ with regards to the positioning of innovative topics.[8] In contrast to Volkswagen, Hyundai and Toyota refer to 'e-mobility & environment' in large parts of their networks; also in the sections 'model-selection' and 'configurator'. This result most likely reflects their focus on hybrid vehicles. All manufacturers refer to 'connectivity & sharing' only in peripheral categories ('offers' and 'operations'). The topic 'autonomous driving & artificial intelligence' is discussed in designated areas on Hyundai's 'newthinking' and Volkswagen's 'media' website; in contrast, it is essentially absent from Toyota's web page.

Despite these differences, the networks show a similar scale-free degree distribution and comparable network properties (Fig. 1C and Fig. 1D). According to the high nodal average degree of 93 and a network density of 0.07, the web page network of Toyota with its 2,654 nodes and 353,327 links is the largest and most densely connected one.

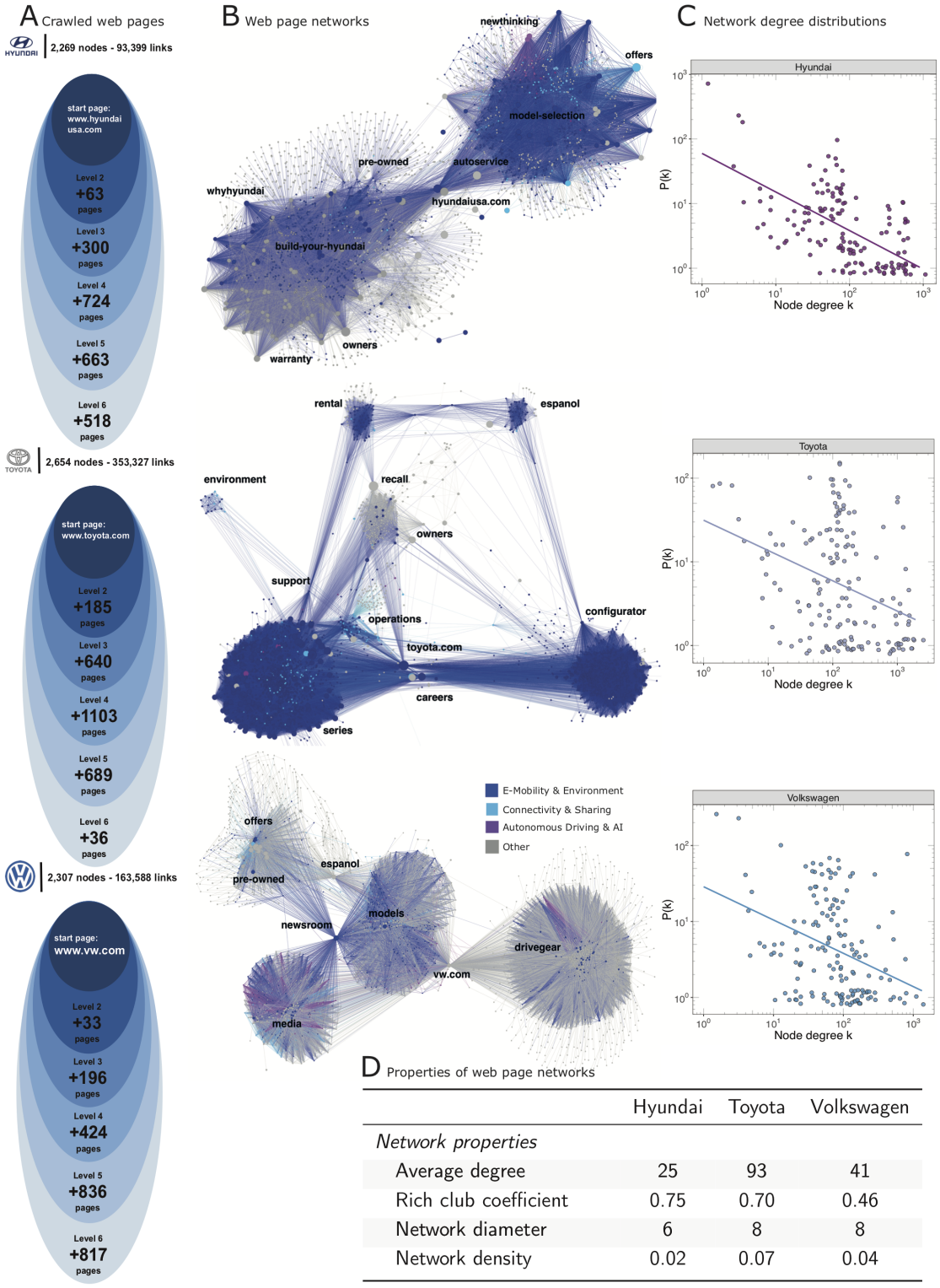
4.2 Sentiment and Centrality

To understand the role of positive language on the car manufacturer web pages, we display the Volkswagen network coloured according to sentiment (Fig. 2A).

Connectivity & shared mobility: connectivity, shared, mobility, sharing, interconnectedness, cloud, cloud computing, wifi, 5G;

Autonomous driving & artificial intelligence: autonomous, self-driving, ai, machine learning, artificial intelligence, intelligent, neural network, algorithm.

[8] In the network visualizations, the nodes are coloured according to the keyword-category that appears most often in a web page. If none of the keywords occurs, the node is coloured in grey.



D Properties of web page networks

	Hyundai	Toyota	Volkswagen
Network properties			
Average degree	25	93	41
Rich club coefficient	0.75	0.70	0.46
Network diameter	6	8	8
Network density	0.02	0.07	0.04

Fig. 1 (**A**) Crawled web pages of three car manufacturers: crawlers collected all sub-pages (nodes) and hyperlinks (links: in- and out-links are combined as undirected links) to a depth of 6 levels in a breadth-first manner. (**B**) Resulting web page networks: pages mentioning keywords on innovative trends are highlighted. The networks differ in their structure and positioning of innovative contents. (**C**) Degree distribution of the networks: they reveal to be scale-free. (**D**) Network Properties: average node degree, average rich club coefficient, network diameter, and network density vary between the three manufacturer web pages.

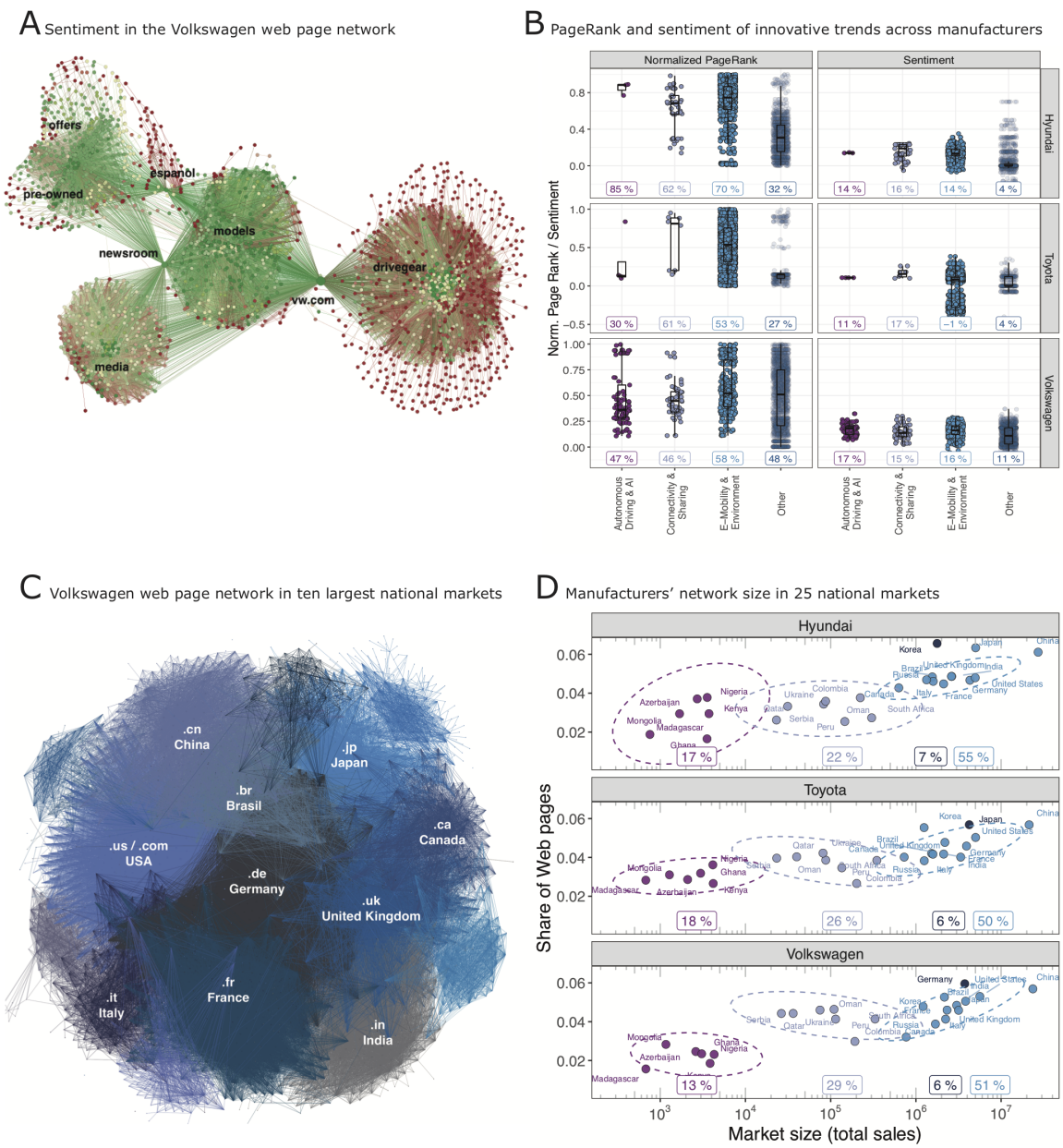


Fig. 2 (**A**) Distribution of sentiment in the Volkswagen web page network: nodes with intermediate degree centrality show, on average, a more positive sentiment (inset). (**B**) Distribution of PageRank centrality and sentiment per topic and manufacturer: pages on innovative topics (in particular e-mobility and environment) tend to be higher ranked and to have more positive sentiment than other sub-pages. (**C**) International Volkswagen web page network (crawled from international landing page) in ten largest car markets: Volkswagen's German home market network forms a central cluster, densely connected with other key key markets in Europe and North America. (**D**) Global sales (log-scale) and share of web pages in market groups: the three manufacturers similarly focus their web presence on a cluster of large car markets, with their home market web page network being disproportionately extensive.

Qualitatively, we note a positive correlation between node degree and sentiment: more central nodes appear to have a more positive sentiment. As the inset shows, node degree centrality and sentiment are actually characterised by an inverse U-shaped relation. A potential reasoning for this finding could be that most topical content is provided on pages with an intermediate degree centrality (e. g. pages that are linked to in the 'media' or 'newsroom' sections of the website). In contrast, less central nodes might rather describe specific topics such as warranty issues or model specifications, and the most central nodes are likely navigational pages, characterised by a more neutral sentiment.

In all three manufacturer web page networks, sites dealing with innovative content tend to be more central than other web pages (Fig. 2B, left panel). The figure displays the distribution of the normalized PageRank per page (PageRank divided by the highest ranked page's value) and the average rank per category (labels below the box plots). Pages with the theme 'e-mobility & environment' are, on average, more central than 'other' web pages with a normalized PageRank of 0.7 vs. 0.32 (Hyundai), 0.53 vs. 0.27 (Toyota), and 0.57 vs. 0.48 (Volkswagen).

Similarly, sites on innovative themes use, on average, a more positive language than other web pages (Fig. 2C, right panel). An exception is Toyota's content on e-mobility with a slightly negative average sentiment of -0.01. While we could not find direct evidence of a relationship between centrality and sentiment per topic category, the results nonetheless suggest that the car manufacturers display content on innovative trends at prominent positions in their networks.

4.3 Country Domain Analysis

For analysing the international orientation of the car manufacturers, we crawl web page network data from the firms international landing page. Figure 2C shows the Volkswagen network of the sub-domains referring to the ten largest national markets in terms of car sales and registrations. The country domains agglomerate in densely connected clusters of different size, indicating the importance of the respective target markets. Volkswagen's German web page cluster is disproportionately large and connected with most other markets. Hence, it is at the core of the international network, reflecting Volkswagen's focus on it's home market.

This finding does not only hold for Volkswagen, but also for the other two car manufacturers (Fig 2D). The figure shows the size of the sub-domain network (number of web pages) per manufacturer in 25 national markets for which we could obtain the number of passenger car sales [35]. Applying a k-means clustering algorithm on the number of sales (log-scale) yields three distinct groups of national markets: small markets of African and Asian countries (less than 1,000 cars sold in 2017), a second cluster of Global South countries with an intermediate market size (1,000 – 1,000,000 cars), and group of twelve large markets (more than 1 million cars), consisting of OECD and Newly Industrialized countries. These clusters are highlighted in the figure.

The three manufacturers show a similar pattern in terms of their web presence in these groups: 13 – 18 % of the firms' web pages target the group of small markets, while 22 – 29 % relate to the medium-sized markets. With 50 – 55 % of all their web pages, the firms' online presence clearly focuses on the group of large national markets with more than a million annually sold cars. The home market's web page network of each company is disproportionately large with 6 – 7 % of all web pages.

This finding provides insights into the competitive strategies of the three firms, as they clearly focus their marketing activities and product range (which determines the size of their national online presences) to a limited number of particularly important markets.

5 Discussion

5.1 Summary

In the digital era, the automotive industry remains one of the cornerstones of global manufacturing, not only in terms of employment and trade, but more importantly for its crucial role in introducing new technologies and organizational methods [9]. This work presents a network-centered approach to gain insights into the innovative focus of three large car manufacturers by analyzing the firms' web pages. In crawling publicly available content and structural meta-information from the websites of Toyota, Hyundai, and Volkswagen, we construct complex networks and analyze their properties. This reveals the companies' focus on themes such as e-mobility, connectivity, or autonomous driving.

Within their web page networks, the three firms centrally present content on the topic 'e-mobility & environment'; however, the topics 'connectivity & sharing' and 'autonomous driving & artificial intelligence' are dealt with only in peripheral sections. The companies tend to make use of more positive language on pages on innovative trends. Our analysis of national sub-domains, moreover, shows that the manufacturers concentrate their online marketing efforts on a limited number of target markets, with a particular focus on their home market.

5.2 Limitations and Future Research

In this study, we gained insights into a company's innovative orientation by considering the network structure and distribution of keywords on corporate websites. Our approach maintains the hierarchical structure of firms' online presence, which is relevant to grasp the prioritization and allocation of topics.

However, this study is exploratory. Thus, it comes with some limitations and future research is needed to establish generalizability and applicability of our findings beyond the case presented here. The innovative themes have been identified via keyword tagging: it might be questioned, in how far keywords alone provide enough information about a website's content. Moreover, mentioning

keywords and writing about innovative trends on the website does not necessarily go along with actual innovative activity. However, the differences between the three firms already show varying prioritization, and the similarities with respect to the national target markets gives insights into the competitive landscape of the automotive industry.

Web-crawling in an iterative manner, as we have presented it here, yields different results, depending on the starting page. Thus, in order to gain more robust results, the data collection should be conducted from several starting pages. Moreover, websites change quite frequently; hence, our approach is highly time-sensitive. To leverage it's potential, the websites should be monitored over longer periods of time, so that changing structures and topics can be identified.

As our approach exploits only publicly available information and is easily reproducible, it promotes transparency and it might be used as a complementary tool to monitor shifting industries and competitive landscapes.

References

1. M. Langley, World Economic Forum, Digital transformation: Understanding the impact of digitalization on society. *World Economic Forum*, (2017).
2. European Economic and Social Committee, Impact of digitalisation and the on-demand economy on labour markets and the consequences for employment and industrial relations. *European Economic and Social Committee*, (2017).
3. B. Weinelt, World Economic Forum, Digital transformation: Digital transformation of the automotive industry. *World Economic Forum*, (2016).
4. F. Braesemann,V. Lehdonvirta, O. Kssi, ICTs and the Urban-Rural Divide: Can Online Labour Platforms Bridge the Gap? in: *SocArxiv preprint*, 10.31235/osf.io/wbxd7, (2019)
5. OECD, Recent developments in the automobile industry. *Economics Department Policy Notes*, No. 7, (2011).
6. D. Mohr, P. Gao, HW. Kaasm, D. Wee, Disruptive trends that will transform the auto industry. *McKinsey & Company*, 1(January), (2016): 1-9.
7. F. Kuhnert, Five trends transforming the automotive industry, *Pricewaterhouse-Coopers*, (2018).
8. OICA, Total manufacturing employment 2018, `http://www.oica.net/category/production-statistics/2017-statistics/` (accessed 2018-12-16).
9. M. Ferrazzi, A. Goldstein, The new geography of automotive manufacturing, *International Economics*, (2011), Chathamhouse.
10. F. Braesemann, N. Stoehr, M. Graham, Global networks in collaborative programming. *Regional Studies, Regional Science*, 6(1), (2019) 371-373.
11. F. Stephany, F. Braesemann, Coding together - coding alone: The role of trust in collaborative programming, in: *SocArxiv preprint*, 10.31235/osf.io/8rf2h, (2019)
12. F. Stephany, F. Braesemann, An exploration of Wikipedia data as a measure of regional knowledge distribution, in: *International Conference on Social Informatics*, 10540 (2017), 31-40, Springer, Cham
13. OICA. World motor vehicle production 2016, `http://www.oica.net/wp-content/uploads/World-Ranking-of-Manufacturers.pdf`(accessed 2018-12-16).

14. M. Traub, H. Voegel, E. Sax, T. Streichert, and J. Haerri, Digitalization in automotive and industrial systems, in: *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, (2018): 1203-1204, IEEE.
15. L. Fridman et al., Mit autonomous vehicle technology study: Large-scale deep learning based analysis of driver behavior and interaction with automation, *arXiv preprint*, arxXiv:1711.06976, (2017).
16. H.-O. Günther, M. Kannegiesser, and N. Autenrieb, The role of electric vehicles for supply chain sustainability in the automotive industry, *Journal of Cleaner Production*, 90, (2015): 220-233.
17. K. D. Thoben, S. Wiesner, and T. Wuest, "Industrie 4.0" and smart manufacturing a review of research issues and application examples, *International Journal of Automation Technology*, 11(1), (2017): 4-19.
18. T. Kito, A. Brintrup, S. New, F. Reed-Tsochas, The structure of the toyota supply network: An empirical analysis. *Said Business School WP* 3, (2014).
19. A. Swaminathan, G. Hoetker, and W. Mitchell. Network structure and business survival: The case of u.s. automobile component suppliers. *University of Illinois at Urbana-Champaign*, (2002).
20. R. Baggio, M.-A. Corigliano, and V. Tallinucci, The websites of a tourism destination: A network analysis, *Information and Communications Technologies in Tourism*, (2007): 279-288.
21. Y. Wang, D. Wang, and W. H. Ip, Optimal design of link structure for e-supermarket website, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 36(2), (2006): 338-355.
22. A. Broder et al., Graph structure in the web, *Computer Networks*, 33(1-6), (2000): 309-320.
23. R. Meusel, S. Vigna, O. Lehmberg, C. Bizer, Graph structure in the web — revisited: A trick of the heavy tail, *Proceedings of the 23rd International Conference on World Wide Web*, (2014) 427–432, ACM.
24. A.-L. Barabási, R. Albert, Emergence of scaling in random networks, *Science*, 286(5439), (1999): 509-512.
25. L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bringing order to the web, *Stanford InfoLab*, (1999).
26. A. Ortiz-Cordova, B. J. Jansen, Classifying web search queries to identify high revenue generating customers, *Journal of the American Society for Information Science and Technology*, 63(7), (2012): 1426-1441.
27. T. Gowda and C. A. Mattmann, Clustering web pages based on structure and style similarity (application paper), *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, (2016): 175-180.
28. H. A. Wan and C.-W. Chung, Web page design and network analysis, *Internet Research*, 8(2), (1998): 115-122.
29. S. Sahebi, F. Oroumchian, and R. Khosravi, An enhanced similarity measure for utilizing site structure in web personalization systems, *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 3, (2008): 82-85.
30. M Bastian, S. Heymann, and Mathieu Jacomy, Gephi: An open source software for exploring and manipulating networks, *Third International AAAI Conference on Weblogs and Social Media*, (2009).
31. J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, G. Woodhull, Graph drawing software: Graphviz and dynagraph - static and dynamic graph drawing tools, *Graph Drawing Software*, Springer, (2003): 127-148.

32. S. A. Schult, A. A. Hagberg and P. J. Swart, Exploring network structure, dynamics, and function using networkx, *Proceedings of the 7th Python in Science Conference (SciPy2008)*, (2008).
33. S. Loria, Textblob: Simplified text processing, 2018, https://textblob.readthedocs.io/en/dev/index.html (accessed 2018-12-16).
34. L. Bing, Sentiment Analysis and Opinion Mining, *Synthesis Lectures on Human Language Technologies*, 5(1), (2012): 1-167.
35. OICA, personal car registrations and sales 2017, http://www.oica.net/wp-content/uploads/Sales-Passenger-cars-2017.pdf(accessed 2018-12-16).