Social Bridges in Urban Purchase Behavior

XIAOWEN DONG*, Massachusetts Institute of Technology
YOSHIHIKO SUHARA*, Massachusetts Institute of Technology
BURÇIN BOZKAYA, Sabanci University
VIVEK K. SINGH, Rutgers University
BRUNO LEPRI, Fondazione Bruno Kessler
ALEX 'SANDY' PENTLAND, Massachusetts Institute of Technology
(*equal contribution)

The understanding and modeling of human purchase behavior in city environment can have important implications in the study of urban economy and in the design and organization of cities. In this paper, we study human purchase behavior at community level and argue that, people who live in different communities but work at close-by locations could act as "social bridges" between the respective communities and that they are correlated with similarity in community purchase behavior. We provide empirical evidence by studying millions of credit card transaction records for tens of thousands of individuals in city environment during a period of three months. More specifically, we show that the number of social bridges between communities is a much stronger indicator of similarity in their purchase behavior than traditionally considered factors such as income and socio-demographic variables. Our findings also suggest that such an effect varies across different merchant categories, that presence of female customers in social bridges is a stronger indicator compared to that of their male counterparts, and that there seems to be a geographical constraint for this effect, all of which may have implications in the studies of urban economy and data-driven urban planning.

CCS Concepts: •**Information systems** → **Data mining;** •**Computing methodologies** → *Machine learning approaches;* •**Applied computing** → *Law, social and behavioral sciences;*

Additional Key Words and Phrases: Purchase behavior, social bridge, physical environment, credit card transaction

ACM Reference Format:
Xiaowen Dong*, Yoshihiko Suhara*, Burçin Bozkaya, Vivek K. Singh, Bruno Lepri and Alex 'Sandy' Pentland, 2016. (*equal contribution) Social bridges in urban purchase behavior. *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 39 (March 2010), 30 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Understanding purchase patterns of city residents can provide valuable insights for the study of the economic dimension of urban areas, thus having important implications in the design and organization of cities and in the study of urban economy. Traditional studies have utilized gravity-based spatial interaction models such as the Huff model [Huff 1964; Bozkaya et al. 2010], or discrete choice models [McFadden 1973], to characterize individual purchase preferences and behaviors. These approaches treat individual purchases separately and do not explicitly consider homophily in social networks and how different people or communities influence each other, which could how-

Author's addresses: X. Dong, Y. Suhara and A. Pentland, Media Lab, Massachusetts Institute of Technology; B. Bozkaya, School of Management, Sabanci University; V. K. Singh, School of Communication and Information, Rutgers University; B. Lepri, Mobile and Social Computing Lab, Fondazione Bruno Kessler.

DOI: 0000001.0000001

ever play an important role in financial decision-making processes. At the same time, there is a rich literature in marketing and economic research that studies the influence on purchase behavior played by socio-demographic characteristics such as age, gender, education and occupation as well as income [Zeithaml 1985; Dholakia 1999; Prasad and Aryasri 2011] and by social interactions [Arndt 1967; Algesheimer et al. 2005]; however, these studies are often based on surveys, and are usually focused on a specific type of products, making them not scalable.

In order to better understand how physical proximity plays a role in individuals' purchase behavior, we focus in this paper on (i) finding factors beyond traditional socio-economic indicators that would better explain similarity in purchase behavior of different groups of people, and (ii) testing the effect of such factors on daily purchase patterns of large population sizes in a city environment.

Several studies have suggested that in a modern era that sees increasing remote communication, the physical environment and social learning due to physical proximity still plays an important role in exchanging ideas [Wu et al. 2008; Eagle et al. 2009; Hristova et al. 2014; Toole et al. 2015]. We therefore conjecture that individuals living in different communities but working at close-by locations could act as *social bridges* that link the two communities, based on the assumption that, due to the exposure to a similar work environment, they could have better chance to exchange information by merely observing or possibly interacting with each other at or near their work places, and thus potentially promote similarity between the behavior of the rest of the residents in their respective communities. A high level illustration of the proposed idea is shown in Fig. 1. Moreover, the recent availability of large-scale financial transaction data [Krumme et al. 2013; Sobolevsky et al. 2014; Lenormand et al. 2015; Singh et al. 2015] has provided us with an excellent opportunity to study human purchase behavior and test our conjecture at large scale.

We provide empirical evidence towards our conjecture by studying the correlations between the presence of social bridges and similarity in community purchase behavior. Specifically, by analyzing millions of credit card transaction records about ten thousands of individuals in two major cities of an Organisation for Economic Co-operation and Development (OECD) country and by showing that, the number of social bridges between different communities is strongly correlated with similarity between the purchase behavior of people from those communities. In particular, the proposed metric based on social bridges is a much stronger indicator of similar purchase behavior among communities than traditional factors such as income, age, gender, and other socio-demographic variables, even after controlling for possible confounding factors such as population and geographical distance. We further test our findings against a null model, i.e., the Huff model traditionally used for modeling purchase behavior, and we show that the observed patterns cannot simply be explained by such a model based on geographical relationship and store popularity.

In addition, our results suggest that the effect of social bridges varies across different merchant categories, and that there exists a gender difference in the effect played by social bridges, i.e., the presence of female customers in social bridges is a stronger indicator compared to that of their male counterparts. Finally, by changing the distance threshold based on which we define the social bridges, we observe interesting results that suggest a possible geographical constraint for such an effect due to physical proximity.

To the best of our knowledge, our study constitutes one of the first attempts to study the correlation between physical proximity and purchase behavior of city residents, using large-scale financial transaction records. The obtained results solidify, by testing them at large scale, traditional studies and theories in marketing and economics about the role of physical environment and/or social learning in understanding purchase be-

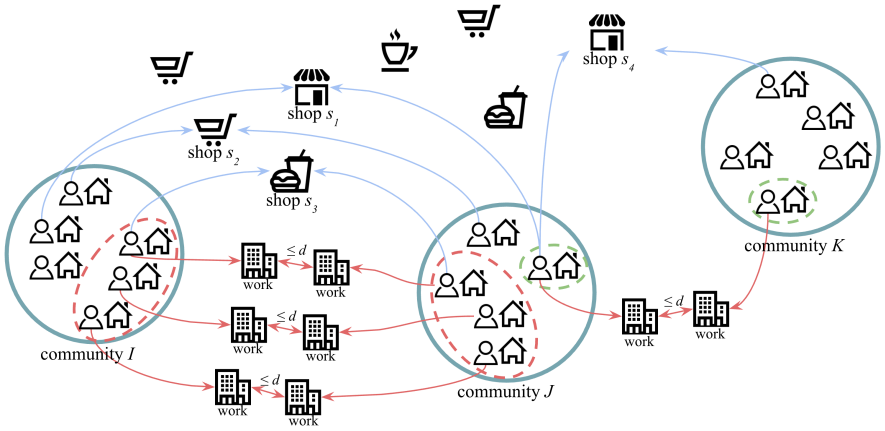


Fig. 1: Social bridges link purchase behavior of different communities of city residents. In this example, communities I and J have three social bridges between them, which are formed by three pairs of customers (highlighted by the red dash circle) having close-by work locations. People in these two communities, not necessarily the ones that form the bridges, purchase at a certain number of stores in common (in this case s_1, s_2 and s_3). Customers in communities J and K share less co-visits (in this case only s_4) which is suggested by the lack of social bridges (formed by people highlighted by the green dash circle) connecting the two communities. The icons used in this figure are obtained from https://iconmonstr.com.

havior. It further enables applications such as prediction of co-visitation patterns and stratification of urban population based on their purchase behavior. As an example, we show that a metric based on social bridges is more effective compared to those based on traditional factors in predicting co-visitation patterns of different urban communities. We therefore believe that our work opens new possibilities in leveraging big financial data for the analysis of human purchase behavior with implications in the studies of urban economy and data-driven urban planning, which would surely contribute to the field of urban computing and urban intelligence.

2. RELATED WORK

Human purchase behavior has traditionally been of interest in the marketing research field [Bawa and Ghosh 1999; Clemente 2002; Adjei et al. 2010; Goel and Goldstein 2013]. For example, Zelthami [Zeithaml 1985] has investigated the relationship between income as well as socio-demographic information (age, gender, working status and marital status) and supermarket shopping behavior (time, frequency, amount spent and attitudes), and used such relationship to segment customers into certain categories such as working females or housewife females. Prasad et al. [Prasad and Aryasri 2011] have shown that customers' socio-demographic information, family size and distance travelled to the store have significant association with retail format choice decisions, while Carpenter et al. [Carpenter and Moore 2006] have provided a general understanding of grocery consumers' retail format choice in the US marketplace. It is also well known that gender difference affects shopping behavior [Teller and Thomson 2012; Hart et al. 2007]. It has been shown that males often shop on a needs-driven basis, while females shop for the intrinsic pleasure [Hart et al. 2007], and female customers are more sensitive to social interaction in terms of shopping patronage

[Teller and Thomson 2012]. Finally, from a computational viewpoint, a gravity-based Huff model [Huff 1964] and a discrete choice model [McFadden 1973] have been utilized to study individual purchase preferences. These works form the foundation of our understanding of human purchase behavior and are certainly important references to the present paper.

There exists a large amount of works on the effect of co-location and face-to-face interactions on individuals' behavior. For example, there have been many studies using location-based social networks to understand social relationships [Li and Chen 2009; Cho et al. 2011; Leung et al. 2011; Scellato et al. 2011; Bouros et al. 2014; Pang and Zhang 2015b; 2015a]. It has been shown in [Chang and Sun 2011; Cho et al. 2011] that geographical proximity and visiting the same places (co-visits) are strong indicators for being friends in location-based social networks, and certain categories (e.g., food, nightlife and residence) of a place where two people meet are strong predictors of being friends [Brown et al. 2013]. Physical proximity has also been shown to promote the chance of face-to-face conversations and offline relationships [Wyatt et al. 2011; Chin et al. 2012].

Moreover, it has been pointed out by several studies that co-location and face-to-face interactions are often associated with homophily in human behavior. Dong et al. [Dong et al. 2011] have used mobile phones for tracking co-location and proximity interactions and have showed that spatio-temporal activities such as physical exercise, residential sector, and on-campus activities are the most important factors to form social relationships. Madan et al. [Madan et al. 2011] have observed increased co-location and physical proximity between students with the same political orientation during the 2008 US presidential election campaign. Hristova et al. [Hristova et al. 2014] have observed evidence of homophily with regards to political opinions, music tastes, health habits, etc., within the online and offline social networks of college students. Toole et al. [Toole et al. 2015] have found that individuals' visitation patterns are far more similar to and predictable by social contacts than by strangers, and that these measures are positively correlated with tie strength. These works have inspired us to define social bridges based on physical proximity and study its effect on shaping community purchase behavior.

The role played by social interactions and social learning [Bandura and McClelland 1977] has also been studied in the context of purchase behavior. The word-of-mouth (WOM) marketing, as a specific type of relationship marketing strategy, has attracted much attention [Arndt 1967; Brown et al. 2005]. Arndt [Arndt 1967] has reported an experiment designed to investigate the short-term sales' effects of product-related conversations, and showed that exposure to favorable comments increase the acceptance of a new product. It has also been suggested that physical proximity causes social learning and similarity in customers' shopping behavior [Algesheimer et al. 2005; Bikhchandani et al. 1998]. Algesheimer et al. [Algesheimer et al. 2005] have studied social influence between the customers of a brand community and have confirmed that community social interactions increase brand-related purchase behavior, while Bikhchandani et al. [Bikhchandani et al. 1998] have argued that reports of the actions or endorsements of one set of economic decision-makers often influence the reactions and purchases of others, and the integration of such learning/cascades effects with other factors could lead to a better understanding of the decision-making processes. In summary, these studies have shown that social interactions and social learning can increase the similarity of individual shopping behavior. Although these works have pointed out the importance of social influence in purchase behavior, they are mainly based on surveys and field-studies, and are often focused on a specific type of products. In comparison, our work is one of the first attempts to verify the effect of physical environment and possible social learning due to physical exposure on general daily

shopping behavior at the community level, using a large-scale data set of credit card transactions.

The recent availability of large-scale data sets, such as taxi trajectories, geo-localized check-ins and credit card transactions, opens new possibilities for studying individual financial behavior, city dynamics and urban economy at finer and unprecedented granularity [Zheng et al. 2014; Krumme et al. 2013; Sobolevsky et al. 2014; Lenormand et al. 2015; Singh et al. 2015; Fu et al. 2014a; Fu et al. 2014b; Karamshuk et al. 2013]. For example, Krumme et al. [Krumme et al. 2013] have studied the predictability of consumers' merchant visitation patterns, and have shown that shopping behavior is highly predictable at long time scales; while Singh et al. [Singh et al. 2015] have used three behavioral features (diversity, loyalty and regularity), based on spatio-temporal patterns of credit card usage, to predict individual financial well-being, namely, overspending, late payment, and being in financial trouble as documented by an administrative action of the bank.

Moreover, mobility patterns have been studied through geo-localized bank transaction data. Lenormand et al. [Lenormand et al. 2015] have shown that human mobility patterns vary between different socio-demographic groups of the population, while Sobolevsky et al. [Sobolevsky et al. 2014] have demonstrated that the flow of individual economic activity in a country is geographically cohesive and consistent with existing administrative regions. They have also pointed out different mobility patterns between local residents and tourists. In a more recent paper, Sobolevsky et al. [Sobolevsky et al. 2016] have also studied the characteristics of different cities through the mobility signatures defined by the spending behavior of their residents. Such mobility patterns have also been used to demonstrate the sensitivity of personal credit card transaction data and aroused discussions about privacy-related issues. For example, it has been shown by de Montioye et al. [de Montjoye et al. 2015] that only a few spatio-temporal points in an anonymized financial data set may be enough to re-identify an individual with little external information.

Finally, several research efforts have adopted a data-driven approach to address questions in urban economy, using methodologies mainly from the computer science perspective, for instance, feature extraction followed by probabilistic inference or supervised learning techniques. As examples, Fu et al. have proposed a sparse pairwise ranking model as well as a ClusRanking method based on geographical dependencies for ranking and predicting real estate values [Fu et al. 2014a; Fu et al. 2014b], and Karamshuk et al. have studied the predictive power of features extracted from geographic information and user mobility for optimal store location for maximum popularity [Karamshuk et al. 2013]. The main differences between these papers and our approach are as follows. First, they aim for addressing specific learning problems such as ranking, where the main contributions lie in feature extraction and novel probabilistic approaches. In comparison, our goal is to identify and test the effect of a simply defined metric, based on physical exposure, on the similarity of purchase behavior. Second, they mainly focus on objectives (such as rank) for individual entities (such as real estate neighborhood or place), where we study a collective behavior between different communities, in this case, the co-visitation patterns as we shall see later.

Despite a growing literature in analyzing quantitative behavioral data, current studies mostly treat individual data records independently and do not incorporate the modeling of social influence and social interactions into the investigation of purchase behavior. This is exactly the motivation of our paper, which contributes to this vibrant research field by providing new insights on the role of physical environment and/or social learning due to physical proximity in understanding the purchase behavior of a community of city residents.

3. DATA AND METHODS

In this section, we first introduce our data set and the associated statistics, along with some data processing steps. We then describe the framework and method we propose to study similarity between communities in purchase behavior.

3.1. Data

We consider more than ten million credit card transaction records provided by a major financial institution in an OECD country about hundreds of thousands of individuals during a period of three months. Each record in the data set corresponds to one credit card transaction along with customer and store IDs, as well as the time (day, hour and minute) of the transaction and the spending amount in local currency. Additional information about the customers and stores are also made available. For customers, we have access to:

— customer age;
— customer gender;
— customer marital status;
— customer education level;
— customer work style (employed by private sector, self-employed, etc.);
— customer income (estimated by the financial institution);
— customer home location;
— customer work location.

For stores, we have access to:

— store location (approximately 40% of the stores are geocoded);
— store category.

The customer-level data are appropriately anonymized such that each customer is represented by a pseudo-unique number and any unique identifier has been removed. The data are analyzed under legal restriction against re-identification, in a way that fully conforms to privacy laws of the country. Data that are sufficient to reproduce the results described in this paper will be made available online.

In our study, we consider city-wide data, hence excluding inter-city factors such as commuting or long distance purchases. We focus on the two largest cities in the country, which we denote as City A and City B. Both cities are densely populated, but with slightly different socio-demographic characteristics (see Table I). For each city, we consider customers that both live and work in the greater city area, where most of their activities take place, and their transactions at stores located within the same area.

Since we are interested in studying the correlation between physical proximity and/or social learning due to exposure to a similar environment, and people's decisions of making purchases, we focus on five merchant categories[1] that mostly correspond to onsite and discretionary purchases and for which we have most data available:

(1) "amusement and entertainment";
(2) "clothing stores";
(3) "retail stores" (including subcategory "grocery stores, supermarkets");
(4) "personal service providers";
(5) "miscellaneous stores" (including subcategory "eating places, restaurants").

We further filter out customers who have less than 20 transactions in total in these five categories during the period of three months, whom we do not consider as regular

[1]A complete list of merchant categories is presented in Appendix.

credit card users. After the filtering process, 44% of the customers in City A and 46% of the customers in City B have been kept, which however account for 80% of the transactions in both cities.

After the data processing steps described above, we are left with 49 thousand customers in City A who have made 2.3 million purchases at 110 thousand stores, and 9 thousand customers in City B who have made 0.4 million purchases at 30 thousand stores. These are the two data sets that we use for the analyses described in this paper. Table I shows some statistics about the socio-demographic characteristics in the two cities. As we can see, City A has more female, young (below 30), single and college-educated customers, as well as a slightly higher median income in local currency.

Table I: Descriptive statistics of the credit card transaction data used in this study.

	City A	City B
# Customers	49K	9K
# Stores	110K	30K
# Transactions	2.3M	0.4M
% Female Customers	37.3%	31.9%
% Young (Below 30) Customers	20.5%	16.1%
% Single Customers	31.4%	22.7%
% College-Educated Customers	51.1%	47.5%
% Employed Customers	92.9%	92.1%
Median Income	2400	2100

3.2. Methods

In this section, we first define the concept of social bridges between urban communities. We then introduce a number of behavioral indices for measuring similarity in community purchase behavior, which we use to evaluate the effect of social bridges.

3.2.1. Social bridges between communities. In order to study the similarity in purchase behavior between different communities of city residents and how this similarity is associated to physical proximity, we first need to define such communities. In the country under investigation, communities can naturally be defined as fine-scale administrative neighborhoods in the city. These are neighborhoods of varying areas from 0.05 square kilometers in the city center to 50 in the periphery of the city area, whose residents normally share to some extent common socio-demographic characteristics. More specifically, we consider each neighborhood as a community whose purchase behavior is defined as that of the people who live in that community. In our data set, there are around 800 communities in City A and 500 communities in City B.

Fig. 2 shows the histograms of some statistics of the communities in City A (left column) and City B (right column) based on which we produce the results in the paper. As we can see, these statistics are consistent with the general statistics in Table I in terms of the differences between the two cities. More specifically, in addition to the different numbers of customers and transaction records, communities in City A has more young and female customers and a higher income than City B.

We propose to define *social bridges* between each pair of communities I and J in order to capture the chance of physical proximity and/or social learning taking place between people from the respective communities. Specifically, we define a social bridge between a pair of communities I and J, for every pair of individuals i and j that live respectively in I and J and have work locations L_i and L_j within a distance threshold d. Therefore, the number of social bridges between I and J is:

$$\text{bdg}(I, J) = |\{i, j\}|, \text{ s.t. } i \in I,\ j \in J,\ D(L_i, L_j) <= d, \tag{1}$$

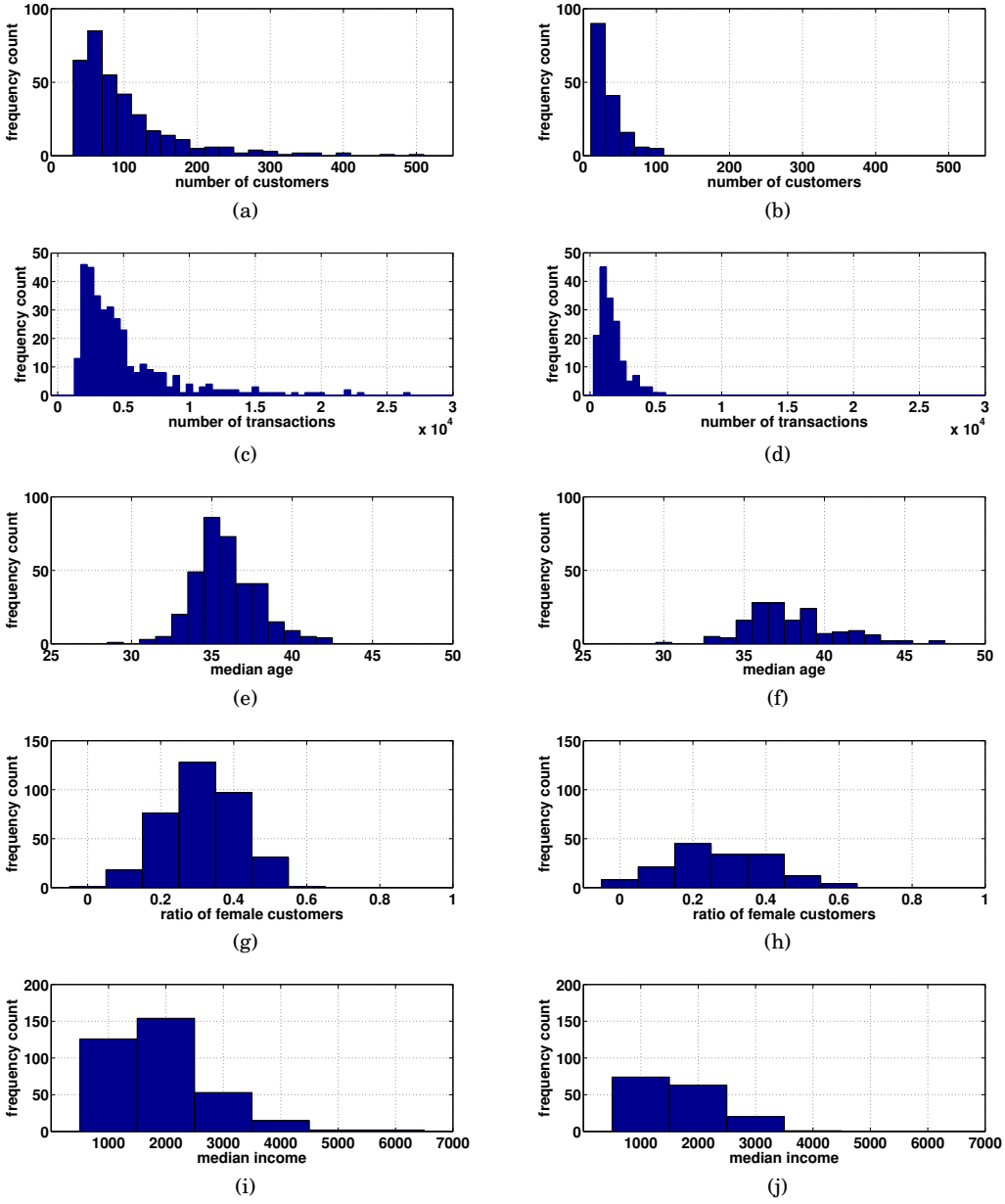


Fig. 2: Histograms of some statistics of the communities in City A (left column) and City B (right column).

where $D(L_i, L_j)$ represents the distance between L_i and L_j. Since people normally spend a considerable amount of time at work, it is our assumption that individuals who work at near-by locations, defined by a distance threshold d, would have reasonable chance to observe and interact with each other due to constant and repeated exposure promoted by physical proximity[2].

Mathematically speaking, given a bipartite graph where the two disjoint sets of vertices are the individuals living respectively in I and J, and the edges indicate whether the work location of each customer i from I and the work location of each customer j from J are within distance d, then the number of social bridges between I and J, $\text{bdg}(I, J)$, is the number of edges in this bipartite graph. In our approach, it is possible for an individual i from community I to form more than one bridge with individuals from J, as long as the pair's work locations are within the distance threshold. The number of social bridges between I and J is thus a number between 0 and the product of number of customers in I and J. Fig. 1 shows an illustrative example where there exist three bridges between the two communities I and J and one between I and K.

Clearly, the choice of d plays an important role in our model. If d is set to 0, then we only construct a bridge between two individuals i and j that work in the same office building, possibly indicating that they are colleagues. On the other hand, if d is set to be large enough to cover the whole city area, then every customer i from I would form a social bridge with every customer j from J. In Section 4, we first present results on chosen values of d for City A and City B, and then study the influence of different values of this parameter on the results, which leads to an interesting observation about a possible geographical constraint of the effect of social bridges.

3.2.2. Behavioral indices for community purchase behavior. For each pair I and J of the communities in the city, we measure the similarity/dissimilarity of purchase behavior of I and J in terms of the following three behavioral indices. Notice that in the computation of the following indices, for each customer we exclude (i) visits during working hours (i.e., from 10am to 6pm during weekdays), and (ii) visits at stores that are located in his/her home and work neighborhoods. The motivation of such treatment is as follows. Traditional purchase behavior model such as the Huff model suggests that people tend to shop more often near their home or work locations. Co-visits that take place in these locations, in particular those made by people around their co-working locations during working hours, can introduce bias in our analysis. By excluding these purchases, the resulting measures would capture similarity in purchase behavior of customers in I and J at stores outside the immediate vicinity of their home and work locations during their spare time, which is largely due to personal preferences.

The first behavioral index is the number of unique co-visited stores by customers in I and J during the period of three months:

$$\text{covisit}(I, J) = |C_I \cap C_J|, \tag{2}$$

where C_I and C_J are the sets of unique stores visited by customers in I and J, respectively, and $|\cdot|$ denotes the cardinality of a discrete set. This metric measures the purchase similarity of communities I and J in terms of purchase choices.

Second, we compute the similarity of temporal distributions of purchases made by customers in I and J. To this end, for each community I, we first compute a 48-

[2]Notice that we may also use the so-called "third places" to define social bridges instead of work locations. However, third places are more difficult to define accurately especially without fine-grained location information, and they often provide short-time exposure instead of the constant and repeated exposure such as that happening in work locations.

dimension vector:

$$T_I(n) = \begin{cases} T_{\text{weekday}}(n), & n = 1, 2, ..., 24, \\ T_{\text{weekend}}(n-24), & n = 25, 26, ..., 48, \end{cases} \tag{3}$$

where $T_{\text{weekday}}(t)$ counts the total number of purchases in the t-th hour on weekdays, and $T_{\text{weekend}}(t)$ counts that in the t-th hour on weekends. We then measure the similarity of T_I and T_J as follows:

$$\text{tsim}(I, J) = \exp(-\text{KL}(T_I, T_J)), \tag{4}$$

where $\text{KL}(\cdot, \cdot)$ denotes the symmetric Kullback-Leibler (KL) divergence [Kullback and Leibler 1951] defined by Johnson et al. [Johnson and Sinanovic 2000], and the exponential function is used for normalizing the divergence to be between 0 and 1. The higher the $\text{tsim}(I, J)$, the more similar the temporal distributions of purchases of I and J. Alternatively, we can also use the cosine similarity of the two 48-dimension vectors to measure the closeness between the temporal distributions.

Finally, we compute, between each pair of I and J, the sum of absolute differences in median spending amount in the five merchant categories introduced previously:

$$\text{mdiff}(I, J) = \sum_{c \in C} |M_I^{(c)} - M_J^{(c)}|_1, \tag{5}$$

where C={"amusement and entertainment", "clothing stores", "retail stores", "personal service providers", "miscellaneous stores"} denotes the set of the five categories we consider, $M_I^{(c)}$ and $M_J^{(c)}$ are the median spending amount of all the transactions made by customers in I and J, respectively, at stores in category c, and $|\cdot|_1$ denotes the L^1-norm. Therefore, $\text{mdiff}(I, J)$ can be thought of as a measure of dissimilarity in purchase behavior between I and J in terms of the spending level of customers at stores in the five selected categories.

To summarize, the proposed behavioral indices are example choices for measuring the similarity/dissimilarity of community purchase behavior in three aspects, namely, $\text{covisit}(I, J)$ for purchase choice, $\text{tsim}(I, J)$ for temporal distribution, and $\text{mdiff}(I, J)$ for spending level. We would however like to remark that, $\text{tsim}(I, J)$ and $\text{mdiff}(I, J)$ are concerned with temporal distribution and spending level that might be more constrained by time and financial situation, and are therefore experimentally less interesting. In comparison, $\text{covisit}(I, J)$ reflects how the choices of stores of different communities of city residents overlap in general, which we consider as the most important and robust index of the three. Therefore, while we present results on all three indices in Section 4.1, we will focus in Section 4.2 on the discussion of the results obtained investigating the relationship between $\text{bdg}(I, J)$ and $\text{covisit}(I, J)$.

3.2.3. Evaluation of effect of social bridges. As explained in Section 3.2.1, we assume that the number of bridges, $\text{bdg}(I, J)$, between communities I and J would capture the chance of physical proximity and/or social learning taking place between people in I and J due to the exposure to a similar physical (work) environment. We are therefore interested in testing the correlation between the number of social bridges between communities and the similarity of their purchase behavior, as evaluated by the three behavioral indices defined in Section 3.2.2.

We first study general trends of the three indices as $\text{bdg}(I, J)$ increases between communities in Section 4.1. We then focus on the more interesting index of $\text{covisit}(I, J)$, and test its relationship with $\text{bdg}(I, J)$ by a regression analysis, where we consider $\text{bdg}(I, J)$ as independent variable and $\text{covisit}(I, J)$ as dependent variable. There are two important remarks on our regression analysis. First, notice that we are interested

in computing the correlation between bdg(I, J) and covisit(I, J), which are both defined as dyadic relationships between communities. Defining two graphs where nodes represent communities and weighted edges represent bdg(I, J) and covisit(I, J), we therefore first vectorize the upper triangular part of the adjacency matrices of the two graphs, and we apply an Ordinary Least Squares (OLS) regression on the two resulting data vectors. Because the entries in each data vector are not independent due to the definitions of social bridges and co-visits, to avoid overestimating statistical significance in this situation, we apply the so-called Quadratic Assignment Procedure (QAP) [Krackhardt 1987; 1988] to test the statistical significance of the obtained β coefficients. The steps of the QAP consist of a random shuffling of the vertices in the graph for the dependent variable, followed by a re-application of the OLS regression. By repeating the QAP for a large number of times (100 in our experiments), we obtain β coefficients that correspond to the null hypothesis that there does not exist significant relationship between the independent and dependent variables; therefore, if the original β coefficient lies at an extreme percentile of the distribution under the null hypothesis, we could reject the null hypothesis and confirm the significance of the observed relationship.

Second, in the regression model we look at the relationship between the independent variable (number of social bridges) and the dependent variable (number of co-visits), while controlling for the effect of possible confounding factors including:

(1) the product of the numbers of individuals in I and J (hence the maximum number of possible bridges);
(2) the inverse of the squared geographical distances between I and J;
(3) similarities in socio-demographic variables including age, gender, marital status, education level and work style;
(4) income.

The motivation of such a multiple OLS regression analysis is as follows. First, as suggested by Pan et al. [Pan et al. 2013], the probability of forming social ties is closely related to the population density and the geographical distance. Furthermore, the number of co-visits between communities could also be related to the population and the distance between them. Second, both social bridges and co-visits are potentially influenced by income and socio-demographic characteristics of the communities. We therefore would like to take out the effects associated to such factors in order to study the effect of social bridges[3]. In addition to geographical distance between a pair of communities, we have also considered the travel time from one to the other by car, which is computed using the ArcGIS software and the road network in the two cities, as a factor, thus taking into account the influence of transportation infrastructure. We found that geographical distance is strongly correlated with travel time by car ($r^2 = 0.81$ in City A and $r^2 = 0.94$ in City B), and therefore do not include travel time as a control variable.

In summary, we apply an OLS regression model where we consider number of co-visits as dependent variable, and number of social bridges as well as other confounding

[3]Notice that in the multiple OLS regression model we directly control for the effect of possible confounding factors, which is in spirit similar to a bivariate analysis where we compute the partial correlation between the independent variable and the dependent variable while controlling for the effect of these factors.

factors as independent variables:

$$\begin{aligned}\text{covisit}(I,J) \sim \beta_0 + \beta_1\ \text{bdg}(I,J) + \beta_2\ \text{pop}(I) * \text{pop}(J) + \beta_3\ 1/\text{dist}(I,J)^2 \\ + \sum_{k=4}^{8} \beta_k\ \text{demo}_{k-3}(I,J) + \beta_9\ \text{inc}(I,J),\end{aligned} \tag{6}$$

where $\text{pop}(I)$ and $\text{pop}(J)$ are the numbers of individuals in I and J, respectively, $\text{dist}(I,J)$ is the geographical distance between I and J, and $\{\text{demo}_k(I,J)\}_{k=1}^{5}$ and $\text{inc}(I,J)$ are the similarities in five socio-demographic variables and income between I and J described in Section 4.2. The β coefficient associated with each independent variable is a quantitative measure on how much that variable contributes to the dependent variable while controlling for all others. In addition to the statistical test within the OLS regression model, we use a multiple regression QAP (MRQAP), which is an extension of the standard QAP in the case of multiple regression, to further validate the significance of the obtained β coefficients.

4. RESULTS

In this section, we present the results obtained verifying the effect of social bridges on shaping community purchase behavior. We first present in Section 4.1, Section 4.2 and Section 4.3 results with given choices of d in City A and City B, and then we investigate in Section 4.4 the influence of this parameter on our results. Finally, we present in Section 4.5 an example application of the effect of social bridges in a prediction task.

4.1. Social bridges and behavioral indices

In Fig. 3 we show the relationship between the number of social bridges between pairs of communities and the three behavioral indices in Section 3.2.2, for 352 communities with more than 50 customers in City A (first row), and for 158 communities with more than 20 customers in City B (second row). Communities with small number of customers are filtered out to make sure that computations are done with a reasonable amount of data. The thresholds 50 and 20 are chosen according to the mean value of customers in the communities in City A and City B, respectively. The majority of the customers has been kept after such filtering process (81% in City A and 75% in City B). The distance thresholds we use are $d \simeq 0.1$ km for City A and $d \simeq 0.2$ km for City B (see Section 4.4 for a discussion on the choice of d).

In Fig. 3, the x-axis represent bins in logarithmic scale that we use to separate the data into different buckets, and the y-axis shows both the mean and the 95% confidence interval (represented by an error bar) of the data in each bucket. We see that as the number of social bridges between I and J, $\text{bdg}(I,J)$, increases, on average (i) the number of unique co-visited stores by customers in communities I and J, $\text{covisit}(I,J)$, also increases; (ii) the temporal distributions of their purchases, $\text{tsim}(I,J)$, become similar; and (iii) the difference in median spending amount in five categories, $\text{mdiff}(I,J)$, decreases. It is worth noting that even though City A and City B have customers with different socio-demographic characteristics, and the volumes of data for the two are clearly different in our data set, the results generally follow similar trends. This show that a large number of social bridges between communities of city residents seems to lead to similar purchase behavior of people from those communities in general.

4.2. Regression analysis between social bridges and purchase similarity (co-visits)

In this section, we test the correlation between social bridges and purchase similarity. We first describe the experimental settings, and then present results on the regres-

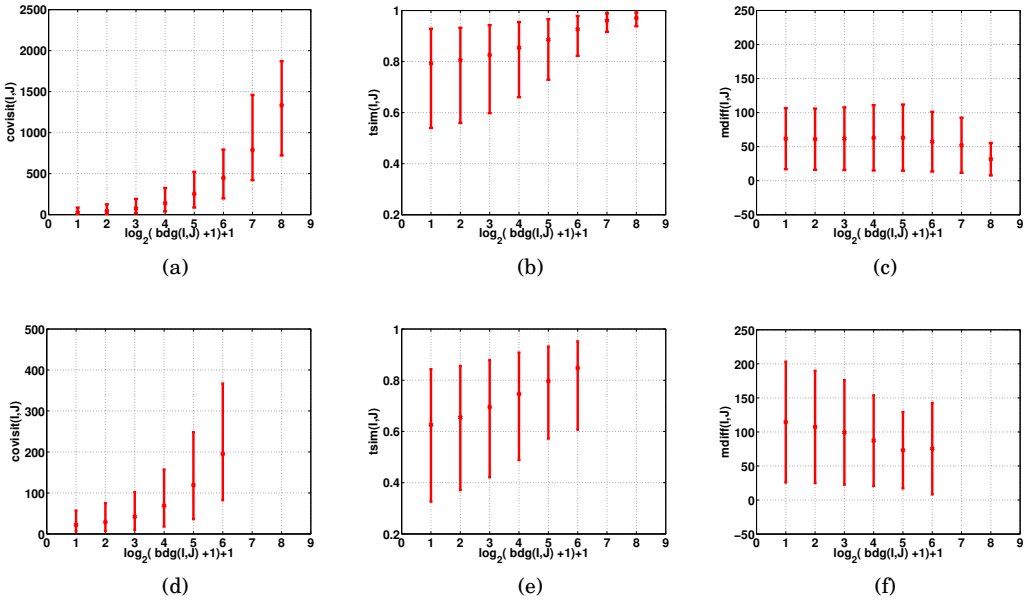


Fig. 3: The relationship between the number of social bridges between pairs of communities in City A (first row) and City B (second row), and the three behavioral indices: (a)(d) number of unique co-visited stores, (b)(e) similarity of temporal distributions of purchases, and (c)(f) sum of differences in median spending amount in local currency in five categories. The error bars show both the mean and the 95% confidence interval of the binned data.

sion analysis. We show the effect of social bridges due to different time and space constraints, role and gender of customers, and categories of merchants.

4.2.1. Settings for regression analysis. We use an OLS regression model where we consider purchase similarity (measured by the number of co-visits, covisit(I, J)) as dependent variable, and number of social bridges as independent variable along with the possible confounding variables introduced in Sec. 3.2.3[4]. This allows us to directly compare the effect of social bridges with other factors that have been traditionally considered to model similarity in purchase behavior, namely, social-demographic and income variables of different communities of city residents.

To this end, we first compute, for each community and for income and five socio-demographic variables including age, gender, marital status, education level and working style, a discrete distribution of each variable using predefined buckets. For example, for age, we compute the number of customers in three ranges, namely, [0-30], [30-60], and [60-90], to assign a 3-dimensional vector to each community; similarly, for income, we look at the number of customers in the [0-33], [33-66], and [66-100] percentiles of the whole income range, to represent each community as a 3-dimensional vector. For the other four variables, we construct such distributions by using directly different categories in each variable. Next, we compute the similarity of the distributions of each variable for every pair of communities I and J, using the normalized KL

[4]We have also considered normalized versions of co-visits and social bridges in an OLS regression model, and presented the results in Table VIII and Table IX in Appendix.

divergence (same as in Eq. (4)). This allows us to create six similarity graphs, one for each variable, and to consider their vectorized forms as independent variables in our regression model (see Eq. (6)).

4.2.2. Regression results. Table II(a) and Table II(b) show the β coefficients[5] and the 95% confidence intervals for the independent variables, as well as root-mean-square error (RMSE) and adjusted R-squared for the regression model, for each pair of 352 communities in City A and each pair of 158 communities in City B, respectively. As we can see, in both cases, the large β coefficients indicate that the number of bridges between I and J is a strong indicator for similar purchase behavior, even after controlling for possible confounding variables such as population, distance, income and socio-demographic variables. We further conduct a robustness check of our results using the jackknife resampling technique and the results are presented in Fig. 9 in Appendix.

Table II: OLS regression model between purchase similarity (i.e., number of co-visits) and number of social bridges, while controlling for population, distance, socio-demographic variables and income.

(a) City A

Indicator	β Coefficient	Confidence Interval
# Social Bridge	0.760 ***	[0.754, 0.766]
Population	0.102 ***	[0.095, 0.108]
Distance	0.094 ***	[0.090, 0.097]
Age	0.038 ***	[0.034, 0.042]
Gender	0.015 ***	[0.011, 0.019]
Marital Status	0.017 ***	[0.013, 0.021]
Education	0.046 ***	[0.042, 0.051]
Working Style	0.015 ***	[0.011, 0.019]
Income	0.034 ***	[0.030, 0.039]
Num. Obs.		61776
RMSE		0.465
Adj. R^2		0.784

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(b) City B

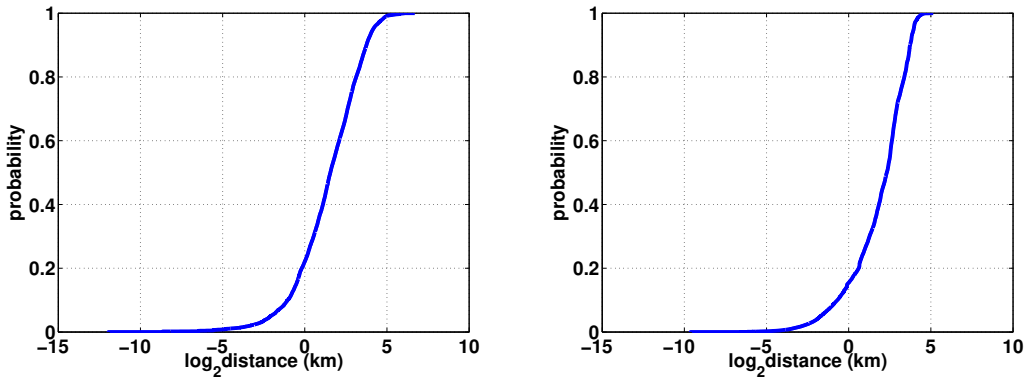
Indicator	β Coefficient	Confidence Interval
# Social Bridge	0.410 ***	[0.393, 0.426]
Population	0.288 ***	[0.272, 0.305]
Distance	0.167 ***	[0.156, 0.179]
Age	0.060 ***	[0.048, 0.072]
Gender	0.155 ***	[0.143, 0.167]
Marital Status	0.023 ***	[0.011, 0.035]
Education	-0.008	[-0.021, 0.005]
Working Style	0.031 ***	[0.019, 0.043]
Income	0.085 ***	[0.072, 0.099]
Num. Obs.		12403
RMSE		0.643
Adj. R^2		0.586

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Even though the causal direction of this relation cannot be claimed in this case, that is, it is not clear whether it is a large number of bridges that leads to similar purchase behavior or the other way around, these results demonstrate that the number of social bridges is a strong statistical indicator of purchase similarity. This shows that physical exposure, in this case due to work location similarity, can be more effective than traditionally considered factors in shaping purchase choices. The difference between the proposed and traditional factors is particularly significant in City A, possibly due to a more vibrant city environment reflected by its socio-demographic characteristics as compared to City B.

To verify that the co-visitation patterns are not simply due to the proximity of co-visited stores to co-working locations, we analyze the distribution of distance between co-visited stores and co-working locations. We define a co-working location as the middle point of the work locations of two people who form a social bridge. Since co-visits are defined on the community level, in order to analyze the distance between co-visited stores and co-working locations, we proceed as follows. For each co-visited store shared

[5]Notice that the statistical significance shown in Table II, Table IV, Table V and Table VI are based on the OLS regression analysis. We note that the p-value cannot be the only criteria for identifying significant correlation in large-scale data sets [Lin et al. 2013], hence we also report confidence intervals and adjusted R-squared. We further validate the significance of the β coefficients using the MRQAP described in Sec. 3.2.3. The results associated with Table II are presented in Fig. 8.



(a) (b)

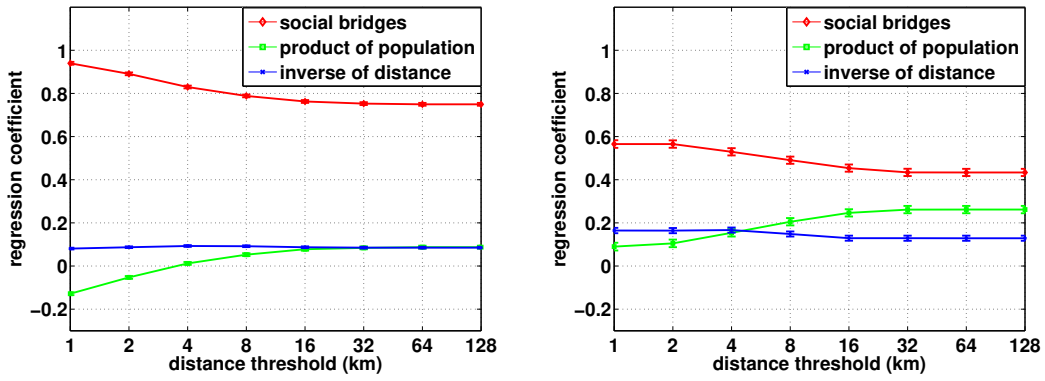
Fig. 4: Cumulative distribution function of distance between co-visited store and the closest co-working location, for (a) City A and (b) City B.

by communities I and J, we compute its distance from the closest co-working location, associated with any social bridge between I and J. Fig. 4 shows the cumulative distribution function of such distance for all co-visited stores (for which we have location information) for all community pairs. We see that co-visits indeed take place sufficiently far away from co-working locations. In fact, 62% of the co-visits take place more than 2 km away from the closest co-working location for City A, and 74% for City B. This verifies that the co-visitation patterns are not simply due to the proximity of the co-visited stores to co-working locations.

To further show that the observed relationship in Table II is robust against the distance between co-visits and co-working locations, we conduct a series of regression analyses where we exclude co-visits that take place beyond certain thresholds on the distance between co-visited store and the closest co-working location. Fig. 5 shows the change of regression coefficients for three variables, i.e., number of social bridges, product of population, and inverse geographical distance, as functions of the distance threshold used for defining the co-visits in this way. As we can see, the coefficient for the number of social bridges decays slightly as distance threshold increases (probably due to the fact that in our data set long-distance co-visits are less often), but always remains strong and significant.

Finally, to show that the observed relationship in Table II is robust against the time window of co-visits, we conduct a series of regression analyses where we compute co-visits that take place in five different time windows: 1) Weekday 12am-10am; 2) Weekday 6pm-12am; 3) Weekend 12am-10am; 4) Weekend 10am-6pm; and 5) Weekend 6pm-12am. The results for the regression analyses are presented in Table III. We see that for all the five time windows, the effect of social bridges remains strong and significant.

4.2.3. Comparison between bridge and non-bridge customers. To better understand the regression results presented in Table II, it is important to look at how the actual co-visits between communities I and J are contributed by customers of different roles in their communities. Based on the definition of social bridges, in each community pair I and J, there are two types of customers: (i) customers who form bridges between I and J, whom we call "bridge customers", and (ii) the rest of the members in their respective



(a) City A (b) City B

Fig. 5: β coefficients in the OLS regression model between number of co-visits and number of social bridges, product of population, and inverse geographical distance, as functions of the distance threshold used for defining the co-visits.

Table III: OLS regression model between purchase similarity (i.e., number of co-visits) in different time windows and number of social bridges.

(a) City A

Co-Visits Windows	β Coefficient	Confidence Interval	**Adj. R^2**
Weekday 12am-10am	0.780 ***	[0.774, 0.786]	0.797
Weekday 6pm-12am	0.801 ***	[0.794, 0.807]	0.754
Weekend 12am-10am	0.757 ***	[0.750, 0.764]	0.706
Weekend 10am-6pm	0.800 ***	[0.792, 0.806]	0.717
Weekend 6pm-12am	0.751 ***	[0.745, 0.758]	0.720

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(b) City B

Co-Visits Windows	β Coefficient	Confidence Interval	**Adj. R^2**
Weekday 12am-10am	0.400 ***	[0.379, 0.417]	0.439
Weekday 6pm-12am	0.393 ***	[0.376, 0.411]	0.521
Weekend 12am-10am	0.359 ***	[0.340, 0.377]	0.465
Weekend 10am-6pm	0.360 ***	[0.340, 0.381]	0.340
Weekend 6pm-12am	0.338 ***	[0.320, 0.357]	0.487

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

communities, whom we call "non-bridge customers". An illustrative example is shown in Fig. 1, where the bridge customers between I and J are highlighted by the red dash circle and those between J and K are highlighted by the green dash circle[6]. Our conjecture is that, due to the exposure to a similar work environment, bridge customers could exchange information revealing community preferences. We are therefore interested in asking the following questions: (i) Do bridge customers indeed have more co-visits? (ii) Is the number of social bridges correlated with more co-visits even between non-bridge customers?

To answer these questions, we consider, for each pair of communities I and J, two types of co-visits: (i) the co-visits made by bridge customers, and (ii) the co-visits made by non-bridge customers. Both cases are illustrated in the example in Fig. 1. For example, for community pair I and J, ratios of bridge customers are $\frac{3}{6} = 0.5$ and $\frac{3}{5} = 0.6$ for I and J, respectively, and the ratio of co-visits by bridge customers and that by non-bridge customers are $\frac{1}{3} = 0.33$ and $\frac{2}{3} = 0.67$, respectively. Fig. 6(a)(b) show the histogram of ratio of bridge customers for all pairs of communities (two ratios per pair) in City A and City B, respectively. We see that, for each pair of communities, the ratio of bridge customers are relatively small. Fig. 6(c)(d) further show the ratio of co-visits by bridge customers versus ratio of co-visits by non-bridge customers for each pair of

[6]Notice that the definition of bridge and non-bridge customers depends on the specific pair of communities and may vary from one pair to another.

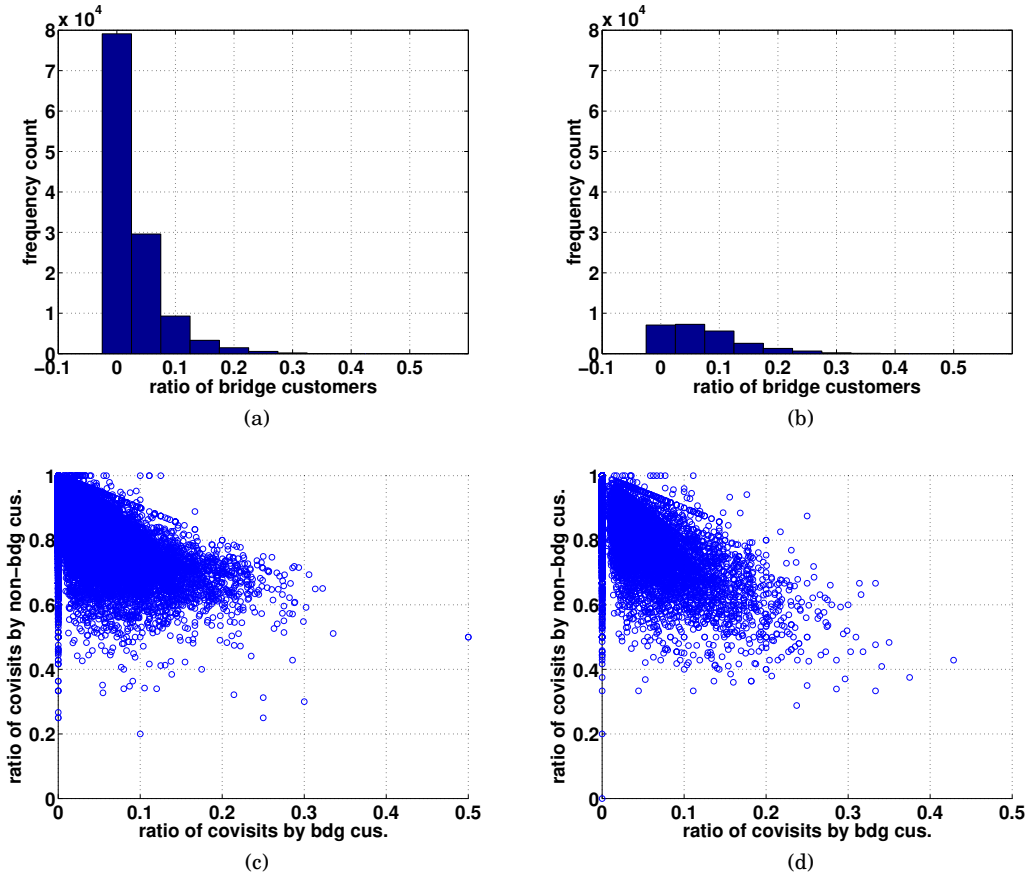


Fig. 6: (a,b) Histogram of ratio of bridge customers for all pairs of communities (two ratios per pair) in (a) City A and (b) City B; (c,d) Ratio of co-visits by bridge customers versus ratio of co-visits by non-bridge customers for each pair of communities in (c) City A and (d) City B.

communities in City A and City B, respectively. We see that a larger portion of co-visits are made by non-bridge customers.

We then test separately their correlations with the number of social bridges between I and J, and the results are shown in Table IV. First, we see that, compared with co-visits by all the customers, the number of co-visits made by bridge customers is even more strongly correlated with the number of social bridges, which is demonstrated by a larger β coefficient. This matches our intuition that physical proximity and/or social learning is more likely to be associated with similar purchase choices. More interestingly, the correlation between the number of bridges and the number of co-visits made by non-bridge customers still remains reasonably strong for City A. As for City B, the relationship is only moderate but still positive and statistically significant. Given that the working locations of non-bridge customers form two separate spatial clusters in the city with a minimum distance of d between any pair of points from the respective clusters, these customers do not seem to have a significant spatial overlap during weekdays, and the fact that they tend to co-visit more stores in case of more

bridges might provide empirical evidence for the effect of social bridges. The exclusion of stores in home and work neighborhoods in the calculation of co-visits, as described in Sec. 3.2.2, further guarantees that these results are not due to people's bias towards stores close to their home and work locations.

Table IV: OLS regression model between purchase similarity (i.e., number of co-visits) of different customer groups and number of social bridges, while controlling for population, distance, socio-demographic variables and income.

(a) City A

Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
By All	0.760 ***	[0.754, 0.766]	0.784
By Bridge Cus.	1.005 ***	[0.999, 1.011]	0.766
By Non-Bridge Cus.	0.653 ***	[0.646, 0.660]	0.705

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(b) City B

Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
By All	0.410 ***	[0.393, 0.426]	0.586
By Bridge Cus.	0.717 ***	[0.700, 0.734]	0.558
By Non-Bridge Cus.	0.238 ***	[0.220, 0.256]	0.490

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

4.2.4. Factors of category of merchant and gender of customer. Another interesting aspect to analyze is whether the effect of social bridges differs for co-visits at different types of stores. In Table V, we show results of a regression analysis where we test separately relationships between the number of social bridges between I and J, and co-visits at the four most common subcategories of the stores within the five broad categories in Sec. 3.1, namely, groceries/supermarkets, eating places/restaurants, family clothing stores, and drug stores/pharmacies. It is interesting to see that, in both cities, the effect of social bridges is strongest for restaurants but weak for supermarkets, where that for clothing stores and drug stores/pharmacies is generally intermediate. This is consistent with our intuition that we are more likely to exchange information about restaurants while for groceries/supermarkets we usually stick to the most convenient choices.

Table V: OLS regression model between purchase similarity (i.e., number of co-visits) at different subcategories of stores and number of social bridges, while controlling for population, distance, socio-demographic variables and income.

(a) City A

Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
Supermarkets	0.610 ***	[0.603, 0.618]	0.693
Restaurants	0.812 ***	[0.805, 0.818]	0.776
Clothing Stores	0.623 ***	[0.615, 0.631]	0.631
Drug Stores	0.724 ***	[0.716, 0.732]	0.589

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(b) City B

Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
Supermarkets	0.291 ***	[0.274, 0.309]	0.537
Restaurants	0.445 ***	[0.426, 0.465]	0.399
Clothing Stores	0.330 ***	[0.312, 0.347]	0.539
Drug Stores	0.286 ***	[0.263, 0.310]	0.182

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Finally, in the context of purchase behavior, an interesting aspect to look at is the effect of gender. To compare the difference between females and males, we analyze two different types of social bridges, namely, female-female bridge and male-male bridge, and two types of co-visits, namely, co-visits by non-bridge female customers and those by non-bridge male customers. We then correlate the number of different types of bridges with the number of different types of co-visits, and the results are presented in Table VI. Interestingly, we see in both cities that bridges formed by female customers lead to stronger correlations with co-visitation patterns of both non-bridge females and non-bridge males in their respective communities. This seems to suggest that female customers are more effective in terms of exchanging store information, and are more

influenced by the physical environment and by the exposure to peers' behavior than their male counterparts. Our results are in line with the ones obtained by [Hart et al. 2007; Teller and Thomson 2012] that have pointed out a gender difference in purchase behavior.

Table VI: OLS regression model between purchase similarity (i.e., number of co-visits) of different customer groups and number of social bridges of different gender combinations, while controlling for population, distance, socio-demographic variables and income.

(a) City A

Bridge Types	Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
Female-Female	**By Non-Bridge Female**	0.527 ***	[0.520, 0.533]	0.625
Female-Female	**By Non-Bridge Male**	0.404 ***	[0.398, 0.411]	0.615
Male-Male	**By Non-Bridge Female**	0.360 ***	[0.352, 0.368]	0.543
Male-Male	**By Non-Bridge Male**	0.393 ***	[0.385, 0.400]	0.604

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(b) City B

Bridge Types	Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
Female-Female	**By Non-Bridge Female**	0.327 ***	[0.311, 0.343]	0.340
Female-Female	**By Non-Bridge Male**	0.106 ***	[0.091, 0.120]	0.468
Male-Male	**By Non-Bridge Female**	-0.073 ***	[-0.092, -0.055]	0.261
Male-Male	**By Non-Bridge Male**	0.044 ***	[0.028, 0.060]	0.460

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

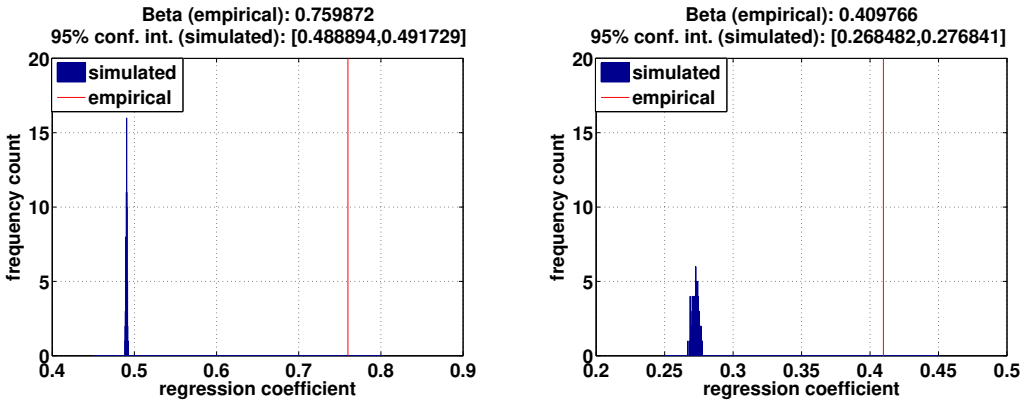
4.3. Comparison with simulation from the Huff model

Purchase choices in urban environment are constrained by popularity and location of the merchants. To further validate the effect of social bridges on co-visitation and show that the observed co-visitation patterns are not simply due to these natural constraints, we consider in this section a null model, namely, the Huff model [Huff 1964], for individual purchase preferences. The basic version of the Huff model takes the following form:

$$p_{is} = \frac{u_{is}}{\sum_{s \in S} u_{is}} = \frac{A_s^{\alpha_1} / D_{is}^{\alpha_2}}{\sum_{s \in S} (A_s^{\alpha_1} / D_{is}^{\alpha_2})}, \tag{7}$$

where p_{is}, the probability for customer i to visit store s, depends on the utility function $u_{is} = A_s^{\alpha_1} / D_{is}^{\alpha_2}$. In Eq. (7), A_s is a measure of the attractiveness of s, D_{is} is the distance between customer i and store s, S is a set of stores, and α_1 and α_2 are two constants. As we can see, the Huff model is a gravity-based model in which the probability for customer i to visit store s is based on the popularity of s and the distance between i and s. Using the Huff model as a null model, we are thus interested in simulating individual purchases so that we can calculate simulated co-visitation patterns between the communities.

We proceed as follows. For each store s that we have location information, we define A_s as the total number of visits it has received during the period of our data, and D_{is} as the distance between the home location of customer i and the location of store s. Following the commonly used Huff-model parameter estimation method [Nakanishi and Cooper 1982], we use an OLS regression model to fit the parameters α_1 and α_2 by minimizing the approximation error between the simulated utility function u_{is} and the empirical $\hat{u}_{is}$ computed from the data. Taking the logarithm of Eq. (7), the parameters α_1 and α_2 can be considered the coefficients of A_s and D_{is} in a linear model. The OLS



(a) (b)

Fig. 7: The distribution of simulated β coefficient (blue) for the factor of number of social bridges in the regression analysis, compared with the empirical $\hat{\beta}$ coefficient (red) in Table II: (a) City A; (b) City B.

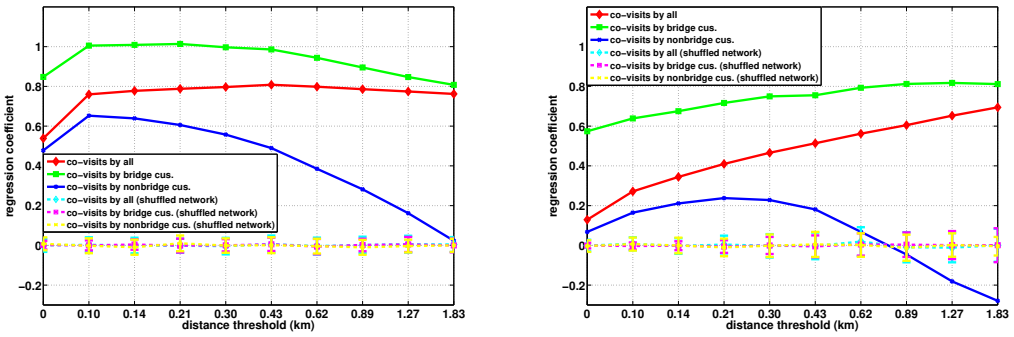
regression model chooses these parameters that minimize the sum of the squares of the gap between the simulated utility value u_{is} and the empirical utility value $\hat{u}_{is}$, namely, the actual visiting count of user i to merchant s. Based on u_{is}, we can then compute the simulated probability p_{is} using Eq. (7).

As a next step, we simulate purchase choices of each individual in the data set. Specifically, for customer i who has made r_i purchases, we simulate his/her purchases using a multinomial distribution with parameters r_i and p_{is}. That is, we sample r_i merchants with replacement based on the probability distribution $p_{i\cdot}$. We consider the sampled results as simulated purchases of customers. We then combine these simulated purchases with empirical purchases at stores for which we do not have location information, and compute the simulated number of co-visits between each pair of communities I and J. Finally, we apply the same regression analysis as in Sec. 4.2 to compute a simulated β coefficient for the factor of number of social bridges, and compare it with the empirical $\hat{\beta}$ coefficient in Table II. As a simulated β coefficient depends on the simulated purchase counts, we repeat the purchase count simulation and regression analysis 100 times to calculate a distribution of the simulated β coefficient.

Fig. 7 shows the distribution of simulated β coefficient (blue) for the factor of number of social bridges in the regression analysis, compared with the empirical $\hat{\beta}$ coefficient (red) in Table II. Since we include empirical purchases at stores without location information in the computation of our simulated number of co-visits, the β coefficient of the null model (Huff model) are in some sense "unfairly" close to the $\hat{\beta}$ coefficient in the empirical results. However, we still observe in Fig. 7 that there exists a significant difference between the regression coefficients in the two cases, which shows that, for both cities, the relationship between the number of social bridges and co-visits are not simply driven by the Huff model.

4.4. Influence of d and geographical constraint of social bridge effect

As described before, the distance threshold d used for the definition of social bridges is a critical parameter in our analysis. Intuitively, a small d means that the social bridges are only constructed within a small area, while a large d means that they can be con-



(a) City A (b) City B

Fig. 8: β coefficients in the OLS regression model between number of co-visits and number of social bridges, as a function of the distance threshold d used for defining the social bridges.

structed even between people who work far away from each other. To simplify, this threshold defines for each customer a circle around his/her work location, with radius d, within which the customer is expected to have the chance to observe or interact with another customer (e.g., while going to the same or close-by places for lunch or coffee, or while taking public transportation). As suggested by Pan et al. [Pan et al. 2013], the chance of a pair of individuals to form social ties decays exponentially as the distance between them increases. Thus, it would be interesting to investigate how the results in Table IV change if we gradually increase the distance threshold d.

In Fig. 8, we show the β coefficients in our regression model as a function of the distance threshold d used for defining social bridges. The values for d are chosen to be logarithmically equi-spaced. The red, green and blue solid curves represent the β coefficients in regression models built considering (i) all the customers, (ii) only the bridge customers, and (iii) only the non-bridge customers, respectively. The cyan, magenta and yellow dash lines represent the corresponding β coefficients after a network shuffling in a MRQAP.

First, our test shows that the obtained β coefficients are not due to correlation between the numbers of social bridges for different pairs of communities. Second, we see that, as d increases from 0, the correlations between the three different types of co-visits and the number of bridges go up. One possible explanation for this behavior is that, as we slightly relax the distance threshold, the criteria for creating social bridges becomes less restrictive, as we start considering people who work sufficiently close-by but not at exactly the same location (e.g., the same office building). This makes sense as physical exposure is not necessarily limited in the office building, and by increasing d slightly we expect to form more bridges between people who have a reasonable chance to observe each other or to interact. As d keeps increasing, the green curve remains quite stable, and the gap between the green and red curves decreases as more and more customers switch their roles from non-bridge customers to bridge customers. Interestingly, the blue curve starts to drop significantly beyond a certain distance. This suggests that after some point, due to the increasing number of bridges, the overall effect of bridge customers in promoting behavioral change in their respective communities seems to decrease. Assuming that every bridge customer is equally good at and willing to spread information in his/her own community, the distance range corre-

sponding to the region around the peak of the blue curve can be thought of as a possible geographical constraint for the social bridge effect.

4.5. Predicting co-visitation patterns using social bridges

The correlation between social bridges and co-visitation patterns of different urban communities enables a number of practical applications. In this section, we show an example application, where we aim at predicting co-visitation patterns of different communities using the proposed metric based on social bridges.

To this end, we formulate a three-class classification problem, where we divide all the community pairs into three groups, according to the three-quantiles of the number of social bridges between all these pairs. This results in three equal-sized groups of community pairs that correspond to small, medium, and large amount of co-visitation. We then consider each of the independent variables in our OLS regression model and the combination of them as features in a classification problem. We train the classifiers based on 20% of randomly selected community pairs (training set), and test the performance on the remaining community pairs (testing set) in terms of prediction accuracy. For classification, we use the scikit-learn library [Pedregosa et al. 2011] with the RBF kernel, where the optimal model parameters are found by a 5-fold cross-validation with grid-search.

We show in Table VII the prediction accuracy for different features (indicators), averaged over 20 random splits of the whole data set into training and testing sets. As we can see, the metric based on social bridges is more effective than any of those based on the traditional factors in terms of predicting co-visitation patterns of different communities, especially for City A. Moreover, in both cases, adding the metric based on social bridges to all other features leads to an improved prediction performance, which demonstrates its meaningfulness in such tasks.

Table VII: Accuracy of prediction of co-visitation patterns between urban communities: (a) City A; (b) City B.

(a) City A

Indicator	Accuracy	Confidence Interval
# Social Bridge	65.10%	[65.06% 65.14%]
Population	55.76%	[55.72% 55.79%]
Distance	48.52%	[48.44% 48.59%]
Age	42.64%	[42.58% 42.70%]
Gender	37.84%	[37.79% 37.88%]
Marital Status	38.28%	[38.24% 38.32%]
Education	40.19%	[40.14% 40.24%]
Working Style	40.82%	[40.74% 40.91%]
Income	35.61%	[35.51% 35.72%]
All except # Social Bridge	67.40%	[67.32% 67.48%]
All	71.75%	[71.70% 71.81%]

(b) City B

Indicator	Accuracy	Confidence Interval
# Social Bridge	55.72%	[55.55% 55.90%]
Population	53.28%	[53.16% 53.40%]
Distance	48.08%	[47.96% 48.21%]
Age	42.25%	[42.10% 42.39%]
Gender	43.73%	[43.60% 43.87%]
Marital Status	39.28%	[39.10% 39.46%]
Education	43.56%	[43.34% 43.77%]
Working Style	42.85%	[42.72% 42.97%]
Income	40.70%	[40.49% 40.91%]
All except # Social Bridge	65.30%	[65.15% 65.46%]
All	66.16%	[65.98% 66.34%]

5. DISCUSSION AND CONCLUSION

Our findings suggest that social bridges between communities of city residents, here defined based on physical proximity of work locations of individuals living in different areas of the city, may account for similarity in purchase behavior. More precisely, we show that the proposed metric based on social bridges is a much stronger indicator of similarity in purchase behavior than traditional factors such as income, age, gender and other socio-demographic variables, even after controlling for possible confounding factors such as geographical distance and population size. Furthermore, we show that

the observed effect cannot simply be explained by a traditional model on purchase behavior, i.e., the Huff model. Therefore, we argue that such similarity might be due to community preferences that are revealed by physical exposure, which is captured by the definition of social bridges.

Our results also show that the effect of social bridges varies across different merchant categories, and that there exists a gender difference in the effect played by social bridges (i.e., the presence of female customers in social bridges is a stronger indicator compared to that of their male counterparts). Finally, results based on different distance thresholds suggest a possible geographical constraint for the effect played by social bridges. Our findings altogether provide evidence that our metric based on social bridges might capture a form of social learning due to physical proximity or exposure to a similar work environment. This is similar in spirit to the concept of "the familiar stranger" [Milgram 1977; Paulos and Goodman 2004; Sun et al. 2013], which suggests that people that observe each other repetitively are more likely to interact than would be perfect strangers due to a background of shared experiences.

Our work bears similarity to works in the computer science community on network structure and influence. For example, it is interesting to notice that the bridge customers, who span both residential communities by working at close-by locations, can also be considered as the structural hole spanners defined in [Lou and Tang 2013]. It is also possible to apply network-based models [Zhang et al. 2012] to analyze influence between customers by constructing a geographical network among them where nodes represent customers and edges represent whether they live or work at close-by locations. However, while these papers aim at developing and analyzing algorithms for computing similarities or quantifying influence between entities, such as the ones for finding structural holes in [Lou and Tang 2013] or computing penalized hitting time in [Zhang et al. 2012], our main objective is to study the relationship between a physical exposure network and a behavioral similarity network, namely, to test and identify the existence of statistical correlation between a simply defined metric based on physical exposure and co-visitation patterns at an aggregate (community) level.

There is evidence in the literature that word-of-mouth and physical exposure are well-known powerful sources of behavior propagation, but their effectiveness in modern cities remains unknown. We therefore test in this paper the existence and strength of such correlation by looking for correlation between physical exposure and shopping behavior. We believe that the strong correlations found in this paper would have practical importance because current methods used in urban planning, policy-making, and marketing mostly rely on demographic information, and our results may provide a different source of information and approaches for such purposes. For instance, for urban planning and policy-making, we can imagine that actions or decisions by planners and policy-makers leading to (re-)location and/or (re-)design of shopping venues, malls, strips, etc. near major workplaces (or new ones to be built) may take into account the location and magnitude of the most prominent social bridges, thus further strengthening the interactions between communities. The concept of social bridges can also be exploited to revive the low economic activity in an area, by analyzing the potential traffic from different communities to the area for economic purposes. Finally, in marketing, companies can take advantage of the concept by analyzing the neighborhoods where they perform poorly and by increasing marketing efforts at or around major workplaces from where, according to the analytics, the largest bridge impact will be transmitted to such neighborhoods.

Furthermore, our work suggests a new way to think about spreading awareness. This is different from traditional notions which rely purely on social or purely on geographic contact to spread awareness. This work suggests that a combination of the two approaches (geo+social through indirect bridges) might work well in many scenarios.

For example, convincing individuals who work in city center might be the appropriate way to convince others about vaccination rather than focusing on just those who have many local connections within the suburb. Such an approach for awareness and spreading of ideas is applicable to the spread of products, services, and ideological viewpoints. From an urban planning perspective, the mixing of people in city centers as opposed to mixed-usage dwellings in the city have different implications regarding the spread of such ideas and viewpoints.

In observational studies, there often exist unobserved confounding variables. The effect of social bridges may be due to word-of-mouth and physical exposure, but also to other unobserved variables that might lead to co-working locations and co-visited stores. While some of them are related to or mediated by demographic/income information, for instance school district, housing price and partly exposure to similar online and TV advertisements, and are thus partially controlled for in our approach, others may remain untested. However, we found out that for a given pair of communities, the ratio of bridge customers is relatively small and a larger portion of the co-visits are made by non-bridge customers. This, together with the results on the correlation between social bridges and co-visits by non-bridge customers, seems to limit the effect of general unobserved variables (that contribute to both co-working and co-purchase) in favor of the effect of social bridges.

One may also argue based on the correlation results that people who visit the same stores might interact with each other, or get exposed to the same job posting information through word-of-mouth during the shopping experience, which in turn leads them to pursue the same job or jobs at close-by locations. However, we believe that the effect of social bridges on purchase similarity is more plausible here, because: (i) it is usually much more difficult to change jobs than visiting different stores; (ii) constant and repeated exposure such as that happening in work locations is more likely to be effective than short-time exposure in stores; (iii) purchase similarity of non-bridge customers who do not work at close-by locations provide evidence supporting our hypothesis.

It is worth noting that our results do not imply a causal relation between social bridges and similarity in purchase behavior. However, even without a causal link, the social bridge effect may have applications in behavior prediction and stratification, campaign targeting, and urban resource allocation. For example, given the shopping preferences of a certain community, we could estimate the likelihood of other communities having similar preferences, based on the concept of social bridges instead of traditional factors such as demographics or geographical distance. As an example, we show that social bridges can be used for predicting co-visitation patterns of different urban communities in a way more effective compared to using traditional factors. Another scenario is stratification of urban neighborhoods by applying a clustering procedure to the social bridge graph. The fact that the definition of social bridges only relies on location information also means that these applications are possible with other data sources such as mobile phone records or geo-localized social media data that are publicly available.

As for causal inference, the results based on the similarity between purchase behavior of non-bridge customers could serve as a first step towards designing causal inference frameworks to verify social influence between different communities in terms of their purchase behavior. More work is needed to understand and model, for example, how social learning or possible interactions between bridge customers lead to a certain level of exchange of purchase or store related information, and how such information could be propagated to other members in their respective communities. With additional longitudinal data, one idea is to study causal relation and behavioral change by defining explicit events of influence. For example, we are currently studying the spreading of customers of newly opened stores in the city, and quantify the effect of

social bridges on such propagation compared to traditional factors. It would also be interesting to investigate how long it takes for the effect of social bridges to show its signs, once a new person starts working somewhere or moves into a neighborhood. Such analyses would certainly have implications in the studies of urban economy and data-driven urban planning.

The data set of credit card transaction records used in our study is a random sample of about 10% of the full customer base of the financial institution. However, the sampling strategy is designed in such a way that the resulting sample set is a representative set of the full customer base. One limitation of using credit card transaction data is that credit card holders may only represent a certain fraction of the population, and people may prefer to pay by cash in several situations. Furthermore, the data set only covers a period of three months, which might seem limited when studying long-term and persistent behavior. Despite these limitations, however, the general consistency between the two cities of different demographics suggests that our results are likely to generalize.

ACKNOWLEDGMENTS

X. Dong was supported by a Swiss National Science Foundation Mobility fellowship while completing this work. The authors are grateful to the financial institution that provided the credit card transaction data for this research.

REFERENCES

M. T. Adjei, S. M. Noble, and C. H. Noble. 2010. The influence of C2C communications in online brand communities on customer purchase behavior. *Journal of the Academy of Marketing Science* 38, 5 (2010), 634–653.

R. Algesheimer, U. M. Dholakia, and A. Herrmann. 2005. The social influence of brand community: Evidence from European car clubs. *Journal of Marketing* 69, 3 (2005), 19–34.

J. Arndt. 1967. Role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research* 4, 3 (1967), 291–295.

A. Bandura and D. C. McClelland. 1977. Social learning theory. (1977).

K. Bawa and A. Ghosh. 1999. A model of household grocery shopping behavior. *Marketing Letters* 10, 2 (1999), 149–160.

S. Bikhchandani, D. Hirshleifer, and I. Welch. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives* 12 (1998), 151–170.

P. Bouros, D. Sacharidis, and N. Bikakis. 2014. Regionally influential users in location-aware social networks. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 501–504.

B. Bozkaya, S. Yanik, and S. Balcisoy. 2010. A GIS-based optimization framework for competitive multi-facility location-routing problem. *Networks and Spatial Economics* 10, 3 (2010), 297–320.

C. Brown, A. Noulas, C. Mascolo, and V. Blondel. 2013. A place-focused model for social networks in cities. In *Proceedings of the International Conference on Social Computing (SocialCom)*. 75–80.

T. J. Brown, T. E. Barry, P. A. Dacin, and R. F. Gunst. 2005. Spreading the word: Investigating antecedents of consumers positive word-of-mouth intentions and behaviors in a retailing context. *Journal of the Academy of Marketing Science* 33, 2 (2005), 123–138.

A. Cameron and P. K. Trivedi. 2005. Microeconometrics : Methods and applications. *Cambridge University Press* (2005).

J. M. Carpenter and M. Moore. 2006. Consumer demographics, store attributes, and retail format choice in the US grocery market. *International Journal of Retail & Distribution Management* 34, 6 (2006), 434–452.

J. Chang and E. Sun. 2011. Location3: How users share and respond to location-based data on social networking sites. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

A. Chin, B. Xu, H. Wang, and X. Wang. 2012. Linking people through physical proximity in a conference. In *Proceedings of the 3rd International Workshop on Modeling Social Media*. 13–20.

E. Cho, S. A. Myers, and J. Leskovec. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1082–1090.

M. N. Clemente. 2002. *The marketing glossary: key terms, concepts and applications*. clementebooks.

Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. Pentland. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347 (2015), 536–539.

R.R. Dholakia. 1999. Going shopping: Key determinants of shopping behaviors and motivations. *International Journal of Retail & Distribution Management* 27, 4 (1999), 154–165.

W. Dong, B. Lepri, and A. Pentland. 2011. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (MUM)*. 134–143.

N. Eagle, A. Pentland, and D. Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)* 106, 36 (2009), 15274–15278.

Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, and J. Yuan. 2014a. Sparse real estate ranking with online user reviews and offline moving behaviors. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. 120–129.

Y. Fu, H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z.-H. Zhou. 2014b. Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1047–1056.

S. Goel and D. G. Goldstein. 2013. Predicting individual behavior with social networks. *Marketing Science* 33, 1 (2013), 82–93.

C. Hart, A. M. Farrell, G. Stachow, G. Reed, and J. W. Cadogan. 2007. Enjoyment of the shopping experience: Impact on customers' repatronage intentions and gender influence. *The Service Industries Journal* 27, 5 (2007), 583–604.

D. Hristova, M. Musolesi, and C. Mascolo. 2014. Keep your friends close and your Facebook friends closer: A multiplex network approach to the analysis of offline and online social ties. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*. 206–215.

D. L. Huff. 1964. Defining and estimating a trading area. *Journal of Marketing* 28, 3 (1964), 34–38.

D. H. Johnson and S. Sinanovic. 2000. Symmetrizing the Kullback-Leibler distance. *IEEE Transactions on Information Theory* (2000).

D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. 2013. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 793–801.

D. Krackhardt. 1987. QAP partialling as a test of spuriousness. *Social Networks* 9 (1987), 171–186.

D. Krackhardt. 1988. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social Networks* 10 (1988), 359–381.

C. Krumme, A. Llorente, M. Cebrian, A. Pentland, and E. Moro. 2013. The predictability of consumer visitation patterns. *Scientific Reports* 3, 1645 (2013).

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 1 (Mar 1951), 79–86.

M. Lenormand, T. Louail, O. G. Cantú-Ros, M. Picornell, R. Herranz, J. M. Arias, M. Barthelemy, M. S. Miguel, and J. J. Ramasco. 2015. Influence of sociodemographic characteristics on human mobility. *Scientific Reports* 5, 10075 (2015), 1–15.

K. W.-T. Leung, D. L. Lee, and W.-C. Lee. 2011. CLR: A collaborative location recommendation framework based on co-clustering. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 305–314.

N. Li and G. Chen. 2009. Analysis of a location-based social network. In *Proceedings of the 2009 International Conference on Computational Science and Engineering*, Vol. 4. 263–270.

M. Lin, H. C. Lucas Jr, and G. Shmueli. 2013. Too big to fail: Large samples and the p-value problem. *Information Systems Research* 24, 4 (2013), 906–917.

T. Lou and J. Tang. 2013. Mining structural hole spanners through information diffusion in social networks. In *Proceedings of the 22nd International Conference on World Wide Web*. 825–836.

A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland. 2011. Pervasive sensing to model political opinions in face-to-face networks. In *Proceedings of the 9th International Conference on Pervasive Computing*. 214–231.

D. McFadden. 1973. Conditional logit analysis of qualitative choice behavior. *P. Zarembka (ed.), Frontiers in econometrics, Academic Press: New York* (1973), 105–142.

S. Milgram. 1977. The individual in a social world: Essays and experiments. *Reading, Mass.: Addison-Wesley Pub. Co.,* (1977).

M. Nakanishi and L. G. Cooper. 1982. Technical note-simplified estimation procedures for MCI models. *Marketing Science* 1, 3 (1982), 314–322.

W. Pan, G. Ghoshal, C. Krumme, M. Cebrian, and A. Pentland. 2013. Urban characteristics attributable to density-driven tie formation. *Nature Communications* 4, 1961 (Jun 2013).

J. Pang and Y. Zhang. 2015a. Exploring communities for effective location prediction. In *Proceedings of the 24th International Conference on World Wide Web*. 87–88.

J. Pang and Y. Zhang. 2015b. Location Prediction: Communities Speak Louder than Friends. In *Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN)*. 161–171.

E. Paulos and E. Goodman. 2004. The familiar stranger: Anxiety, comfort, and play in public places. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 223–230.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.

C. J. Prasad and A. R. Aryasri. 2011. Effect of shopper attributes on retail format choice behaviour for food and grocery retailing in India. *International Journal of Retail & Distribution Management* 39, 1 (2011), 68–86.

S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. 2011. Socio-spatial properties of online location-based social networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*. 329–336.

V. K. Singh, B. Bozkaya, and A. Pentland. 2015. Money walks: Implicit mobility behavior and financial well-being. *PLoS ONE* 10, 8 (2015), e0136628.

S. Sobolevsky, I. Sitko, R. Tachet des Combes, B. Hawelka, J. M. Arias, and C. Ratti. 2014. Money on the Move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in Spain. In *Proceedings of the 2014 IEEE International Congress on Big Data*. 136–143.

S. Sobolevsky, I. Sitko, R. Tachet des Combes, B. Hawelka, J. M. Arias, and C. Ratti. 2016. Cities through the prism of peoples spending behavior. *PLoS ONE* 11, 2 (2016), e0146291.

L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang. 2013. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences (PNAS)* 110, 34 (2013), 13774–13779.

C. Teller and J. A. Thomson. 2012. Gender differences of shoppers in the marketing and management of retail agglomerations. *The Service Industries Journal* 32, 6 (May 2012), 961–980.

J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González. 2015. Coupling human mobility and social ties. *Journal of The Royal Society Interface* 12, 105 (2015), 20141128.

L. Wu, B. Waber, S. Aral, E. Brynjolfsson, and A. Pentland. 2008. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. In *Proceedings of the International Conference on Information Systems (ICIS)*.

D. Wyatt, T. Choudhury, J. Bilmes, and J. A. Kitts. 2011. Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. *ACM Transactions on Intelligent Systems and Technology* 2, 1 (2011), 7:1–7:41.

V. A. Zeithaml. 1985. The new demographics and market fragmentation. *Journal of Marketing* 49, 3 (1985), 64–75.

C. Zhang, L. Shou, K. Chen, G. Chen, and Y. Bei. 2012. Evaluating geo-social influence in location-based social networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 1442–1451.

Y. Zheng, L. Capra, O. Wolfson, and H. Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology* 5, 3 (2014), 38:1–38:55.

Appendix

List of merchant categories

The complete list of merchant categories available in our data set corresponds to most Merchant Category Codes (MCC) listed in ISO 18245 for retail financial services, which includes:

(1) Agricultural Cooperatives
(2) Air Conditioning, Heating, and Plumbing Contractors
(3) Airlines
(4) Amusement and Entertainment
(5) Automobiles and Vehicles
(6) Automotive/Vehicle Rentals
(7) Business Services
(8) Carpentry Contractors
(9) Clothing Stores
(10) Concrete Work Contractors
(11) Contractors, Special Trade-notelsewhere classified
(12) Electrical Contractors
(13) General Contractors-Residential and Commercial
(14) Government Services
(15) Horticultural and Landscaping Services
(16) Hotels and Motels
(17) Insulation, Masonry, Plastering, Stonework, and Tile Setting Contractors
(18) Mail Order/Telephone Order Providers
(19) Marriot
(20) Miscellaneous Stores
(21) Miscellaneous Publishing and Printing
(22) Personal Service Providers
(23) Professional Services and Membership Organizations
(24) Property manager
(25) Repair Services
(26) Retail Stores
(27) Roofing and Siding, Sheet Metal Work Contractors
(28) Service Providers
(29) Specialty Cleaning, Polishing,and Sanitation Preparations
(30) Transportation
(31) Typesetting, Plate Making, and Related Services
(32) Utilities
(33) Veterinary Services
(34) Wholesale Distributors and Manufacturers

Within these broad categories, there exists a more detailed list of subcategories of merchants. We have selected four of these subcategories within the five broad categories in Sec. 3.1 for the analysis in Sec. 4.2.4.

Regression analysis on normalized social bridge and co-visit indexes

In this section, we conduct analysis on normalized social bridge and co-visit indexes. Specifically, the normalized social bridge index is defined as the number of social bridges divided by the product of the population of the two communities in the data set, and the normalized co-visit index is defined as the Jaccard index.

Since the union of visits by two communities, which is used as a normalizing factor in the Jaccard index, is strongly correlated with the product of their population in the

data set ($r = 0.86$ for City A and $r = 0.84$ for City B), both normalized indexes have essentially taken into account the population factor. We therefore remove the population favor in the regression analysis, and the results are presented below in Table VIII and Table IX (equivalent to Table II and Table IV of the main manuscript). Overall, we see that results are consistent with those in Table II and Table IV of the main manuscript: In case of City A, there is a strong correlation between the normalized social bridge and co-visit indexes. For City B, the social bridge effect is moderate but remains positive and statistically significant.

In the regression analysis presented in the main manuscript, we take into account the population factor by considering it as an independent variable. The normalized social bridge and co-visit indexes described above are an alternative to achieve the same objective. Since normalized indexes can often be defined in a number of different ways (e.g., another normalized version of social bridge index can be the ratio of the number of customers forming social bridges from the two communities to the total number of customers in the two communities), we choose to keep the current regression framework with the original variables while considering population as a confounding variable in the main manuscript.

Table VIII: OLS regression model between purchase similarity (i.e., number of co-visits) and number of social bridges, while controlling for other variables.

(a) City A

Indicator	β Coefficient	Confidence Interval
# Social Bridge	0.505 ***	[0.500, 0.512]
Distance	0.205 ***	[0.199, 0.211]
Age	0.063 ***	[0.056, 0.069]
Gender	0.034 ***	[0.027, 0.040]
Marital Status	0.054 ***	[0.047, 0.060]
Education	0.149 ***	[0.142, 0.157]
Working Style	0.106 ***	[0.100, 0.112]
Income	0.005	[-0.003, 0.012]
Num. Obs.		61776
RMSE		0.743
Adj. R^2		0.448

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(b) City B

Indicator	β Coefficient	Confidence Interval
# Social Bridge	0.221 ***	[0.207, 0.236]
Distance	0.289 ***	[0.274, 0.303]
Age	0.076 ***	[0.061, 0.091]
Gender	0.208 ***	[0.192, 0.223]
Marital Status	0.057 ***	[0.043, 0.071]
Education	0.070 ***	[0.053, 0.086]
Working Style	0.117 ***	[0.103, 0.132]
Income	0.139 ***	[0.122, 0.156]
Num. Obs.		12403
RMSE		0.801
Adj. R^2		0.359

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table IX: OLS regression model between purchase similarity (i.e., number of co-visits) of different customer groups and number of social bridges, while controlling for other variables.

(a) City A

Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
By All	0.505 ***	[0.500, 0.512]	0.448
By Bridge Cus.	0.498 ***	[0.491, 0.504]	0.377
By Non-Bridge Cus.	0.464 ***	[0.457, 0.470]	0.414

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(b) City B

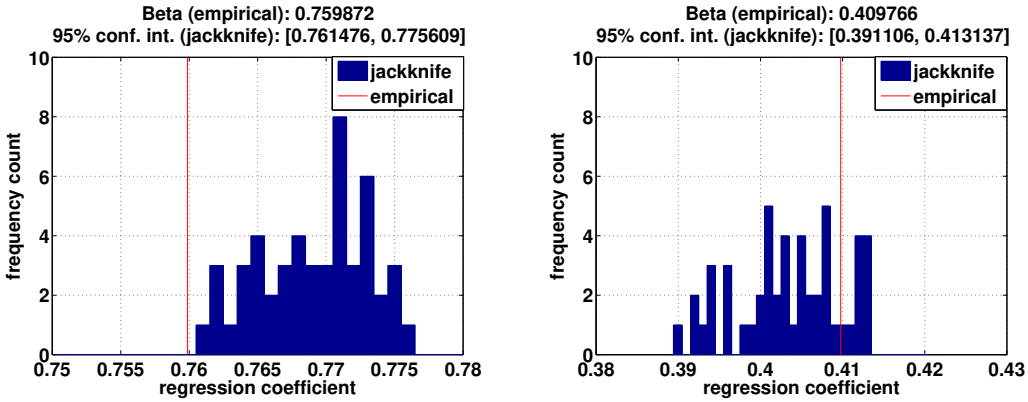
Co-Visits Types	β Coefficient	Confidence Interval	**Adj. R^2**
By All	0.221 ***	[0.207, 0.236]	0.359
By Bridge Cus.	0.344 ***	[0.328, 0.360]	0.270
By Non-Bridge Cus.	0.141 ***	[0.126, 0.156]	0.318

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Robustness check using jackknife resampling

As a robustness check of our results in Table II, we have computed the jackknife estimate of the regression coefficient for the variable of social bridges. The jackknife is a resampling technique for variance estimation [Cameron and Trivedi 2005]. To compute the jackknife estimate of a parameter, a random subset of the data is repeatedly left out in the analysis and estimates of the parameter of interest from multiple such trials are averaged. Specifically, we randomly remove 5% of the active customer-store pairs in our data set to compute co-visits, and apply the same regression analysis to

obtain the coefficient for the variable of social bridges. We then repeat such procedure for 50 times and compute the 95% confidence interval of the regression coefficient, and the results are presented in Fig. 9. We see that the confidence interval of the jackknife estimate of the regression coefficient for social bridges is very close to the empirical value in Table II, which indicates its robustness against perturbation of the data.



(a) City A (b) City B

Fig. 9: Jackknife estimate of the β coefficients in the OLS regression model between number of co-visits and number of social bridges.