Social Connections in Urban Consumer Behaviour

B Meenal

January 22, 2025

Social Connections in Urban Consumer Behaviour

Meenal Badki
UT Dallas

Abstract
Understanding and depicting human purchasing patterns in urban areas can have a profound effect on urban economics research and can shape city planning and development. This study investigates consumer purchasing behavior at the community level, proposing that individuals living in different communities but working in close proximity can act as "social bridges," connecting these communities and exhibiting a relationship with shared patterns in their purchasing activities.We present empirical evidence through the analysis of a large dataset consisting of credit card transactions from a diverse sample of urban individuals over a three-month period. Specifically, we illustrate that the presence of social connections connecting communities acts as a notably more influential indicator of congruence in their consumption patterns than conventional considerations such as income and socio-demographic factors. Our findings suggest that the impact of this phenomenon varies across different types of merchants, with female customers in social networks proving to be a more reliable predictor than their male counterparts. Furthermore, a geographical constraint seems to influence the manifestation of this effect, highlighting important considerations for studies on urban economies and data-driven urban planning.
Information systems > Data mining; Computing methodologies > *Machine learn- ing approaches;* **Applied computing >** *Law, social and behavioural sciences*

Keywords: Purchase behaviour, social bridge, physical environment, credit card transaction

1. Introduction
Acquiring a deeper understanding of the purchasing patterns of urban residents can provide valuable insights into the economic dynamics of metropolitan areas, impacting urban planning, city development, and the analysis of urban economic phenomena. Traditional research has utilised gravity-centered spatial interaction frameworks such as the Huff model [Huff 1964; Bozkaya et al.]. The year 2010 or the discrete choice models established by McFadden in 1973 are utilized to elucidate individual preferences and behaviours in making purchases. These approaches examine individual transactions independently and do not specifically consider the influence of homophily within social networks or the interconnected relationships between individuals or groups, which can have a substantial influence on financial decision-making processes. Concurrently, a substantial body of academic literature in the fields of marketing and economics has been dedicated to investigating the influence of socio-demographic variables such as age, gender, education, occupation, and income on consumer behaviour. This existing research also emphasises the significance of social interactions in shaping consumer behaviour. However, these investigations often utilise survey methods and tend to focus on specific product categories, leading to challenges in scalability.
In order to enhance understanding regarding the impact of physical proximity on consumer purchasing behaviour, this study focuses on (i) delineating determinants beyond traditional socio-economic measures that offer improved clarity on the commonalities in purchasing conduct across diverse demographic segments, and (ii) assessing the ramifications of these determinants on the everyday purchasing patterns of extensive populations residing in urban environments. Several studies have demonstrated that, notwithstanding the increasing prevalence of remote communication in the present era, the physical proximity and social interactions that occur in close physical proximity continue to play a crucial role in facilitating the exchange of ideas (Wu et al., ...). 2008; Eagle et al. 2009; Hristova et al. 2014; Toole et al. 2015]. Therefore, we propose that individuals residing in separate communities but working in close proximity have the potential to serve as social bridges that link the two communities together. This proposition is based on the idea that their mutual workplace setting increases the probability of information exchange through informal observation or potential engagements at or around their place of work, ultimately encouraging similarities in the behaviours of other residents within their respective communities. A comprehensive illustration of the proposed concept is provided in Figure [1]. Moreover, the recent availability of comprehensive financial transaction data [Krumme et al. 2013; Sobolevsky et al. 2014; Lenormand et al. 2015; Singh et al. The year 2015 presented a valuable opportunity to study human purchasing behaviour and assess our

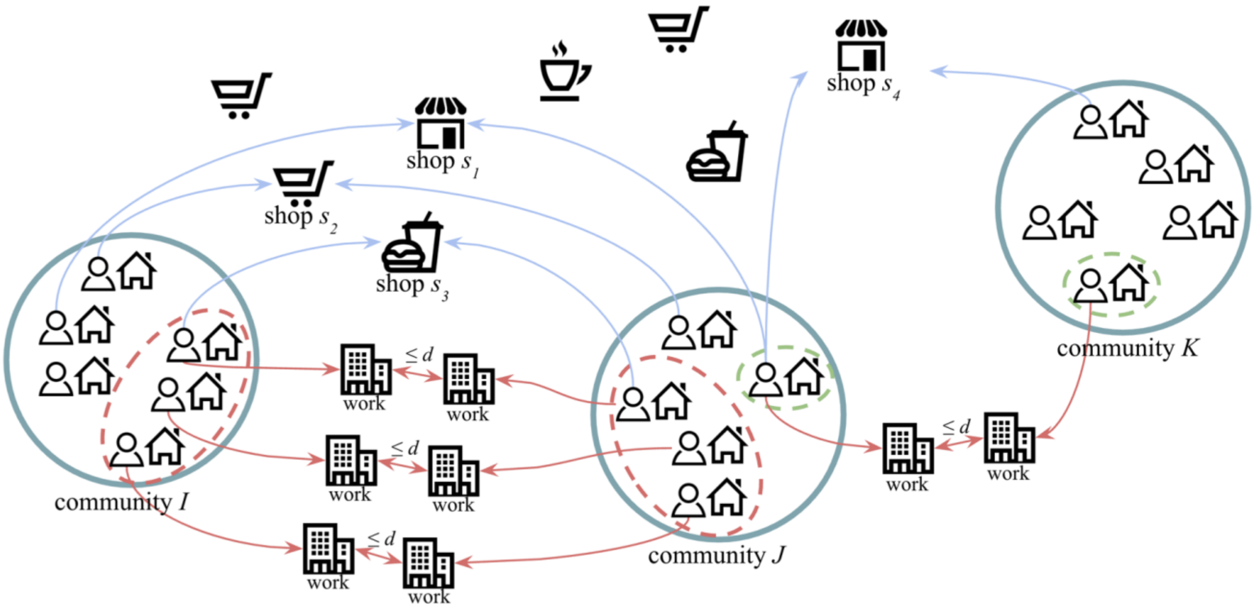
hypotheses on a significant level.

We provide empirical evidence for our hypothesis through an analysis of the correlations between the existence of social bridges and the resemblances in community purchasing behaviours. Specifically, by analysing extensive datasets comprising credit card transactions of numerous individuals in two metropolitan areas within a member nation of the Organisation for Economic Co-operation and Development (OECD), it is evident that the degree of social connections across different communities has a substantial association with the alignment in consumer behaviour patterns of individuals within those communities. In particular, the proposed measure centered on social connections emerges as a notably superior indicator of analogous buying patterns among localities as opposed to traditional variables such as income, age, gender, and other socio-demographic characteristics, even after considering potential influencing factors like population size and spatial proximity. We further evaluate our findings in comparison to a null model, specifically the Huff model typically used to simulate consumer purchasing patterns. Our analysis confirms that the observed trends cannot be solely attributed to the spatial proximity of stores and their popularity, as encapsulated in the null model. Additionally, our research reveals variations in the effect of social connections across different merchant sectors, with a notable divergence in the significance of such connections based on gender. Particularly, the presence of female clientele within social connections emerges as a more impactful factor compared to the presence of male clientele.

To the best of our understanding, our study constitutes one of the first attempts to investigate the correlation between proximity and purchasing behaviour among city dwellers, analysing a comprehensive dataset of financial transactions. The findings from extensive testing provide further validation to established academic research and theories in the fields of marketing and economics that examine the impact of physical surroundings and social learning in shaping consumer purchasing decisions. Moreover, it facilitates the generation of predictions regarding patterns of co-visitation and the segmentation of city dwellers according to their purchasing behaviours. For example, we illustrate that a metric centering on social connections proves more effective than conventional parameters in forecasting patterns of co-visitation across diverse urban neighbourhoods. We show that focusing on social connections is better than traditional methods for predicting how people will visit different urban neighbourhoods. Social connections as a metric is better for predicting patterns of co-visitation in urban neighbourhoods than traditional parameters. For instance, we demonstrate that a measurement that focuses on social connections is more successful than traditional parameters when it comes to predicting how people from various urban neighbourhoods visit each other.

We show that a social connections-based metric is better at predicting co-visitation patterns in various urban areas compared to traditional parameters. For instance, we demonstrate that a metric focusing on social relationships is more efficient than traditional parameters in predicting trends of reciprocal visits within a variety of urban communities. For instance, we show that a way to measure social connections works better than regular measurements to predict how people visit different neighbourhoods in the city. Therefore, our study provides novel prospects for leveraging extensive financial data in the analysis of consumer purchasing behaviours. This has significant implications for research in urban economics and data-driven urban planning, potentially contributing to advancements in the fields of urban computing and urban intelligence. Historically, there has been a focus on human purchasing behaviour in the realm of marketing research, as evident in the works of Bawa and Ghosh (1999), Clemente (2002), and Adjei et al. According to Goel and Goldstein (2013), as previously noted in 2010. As an illustration, Zelthami [Zeithaml 1985] investigated the correlation between income and socio-demographic variables (such as age, gender, employment status, and marital status) and shopping behaviours associated with supermarkets (including time allocation, shopping frequency, expenditure habits, and attitudes), leveraging this association to classify consumers into distinct segments like employed women or stay-at-home individuals. Prasad and colleagues. [Prasad and Aryasri's (2011) study elucidated that customers' socio-demographic characteristics, household size, and proximity to the retail outlet play a significant role in shaping their preferences for specific retail formats, while Carpenter et al. Carpenter and Moore (2006) have provided an extensive analysis of the decision-making process of American grocery shoppers when choosing retail formats. It is widely acknowledged in the academic literature that variations in gender have an impact on consumer shopping patterns (Teller and Thomson, 2012; Hart et al.). 2007]. Studies have shown that men generally engage in shopping out of necessity, while women tend to shop for recreational purposes (Hart et al., year). 2007]. Furthermore, women consumers demonstrate a

greater sensitivity to social interactions in the context of brand allegiance, as highlighted by Teller and Thomson's (2012) study. In the context of computation, researchers have utilised a gravity-based Huff model [Huff 1964] and a discrete choice model [McFadden 1973] to investigate the purchasing preferences of individuals. These research articles lay the foundation for our understanding of consumer purchasing patterns and serve as important references for the present study.
Numerous studies have concentrated on investigating the influence of physical proximity and face-to-face communication on individuals' behavior. For example, a plethora of scholarly investigations have employed location-based social networks as a tool for examining 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; and 2015a]. Studies conducted by Chang and Sun (2011) and Cho et al. have illustrated... According to Brown et al. (2011), proximity and shared visitation to various locations carry substantial importance in determining friendships within location-based social networks. Additionally, the types of venues where individuals frequently gather, such as those related to food, nightlife, and housing, serve as robust determinants of social connections. 2013]. Moreover, research has shown that proximity plays a crucial role in fostering face-to-face interactions and tangible social relationships (Wyatt et al., [year]). 2011; Chin et al. 2012].
Studies have shown that people often share similarities due to being in the same place and interacting face-to-face. Dong et al. [Dong et al. In 2011, they used mobile phones to track where people were together and how close they were, showing that activities like exercising, living in the same area, and participating in campus events play a big role in forming social connections. Madan and colleagues. [Madan et al. In 2011, it was observed that students with similar political beliefs were more likely to spend time together during the 2008 US presidential election campaign. Hristova and colleagues. [Hristova et al. In 2014, researchers found that college students tend to have similar interests like politics, music, and health practices in both online and offline social networks. Toole and others. [Toole et al. In 2015, it was found that people's visitation habits are more similar and predictable based on social connections rather than strangers, with these patterns being influenced by the strength of relationships. These studies encouraged us to form relationships through proximity and study how they affect shopping habits within the community.
Researchers have also looked into how social interactions and learning from others can affect people's buying decisions. Word-of-mouth marketing, seen as a type of relationship marketing, has attracted a lot of attention from scholars (Arndt 1967; Brown et al.). 2005]. Arndt (1967) did a study to look at how product information affects consumers quickly.
Social connections influence the shopping behavior of different groups living in cities. In this situation, communities I and J have three social connections in common, formed by three pairs of customers who work near each other (highlighted by red dashed circles). People from both communities, not just those who make connections, shop at some of the same stores (specifically s1, s2, and s3) together. Clients in communities J and K have fewer meetings together because there are no social connections between them. The symbols in this image come from https://iconmonstr.com.



Conversations concerning sales indicate that positive comments increase the probability of new product acceptance. Another proposition suggests that physical proximity fosters social learning and influences similarities in shoppers' behavior [Algesheimer et al.]. 2005; Bikhchandani et al. 1998]. Algesheimer and colleagues. [Algesheimer et al. A 2005 study found that interacting with others in a brand community positively impacts customer purchasing decisions, as highlighted by Bikhchandani et al. [Bikhchandani et al. In 1998, it was posited that the behaviors or approvals of a particular set of economic decision-makers often influence the responses and transactions of other individuals, and integrating these learning and cascade effects with a range of variables could facilitate a deeper understanding of the decision-making mechanisms. In summary, these investigations suggest that engaging in social interactions and social learning may contribute to the similarity observed in the shopping behaviors of individuals. While the aforementioned studies underscore the importance of social influence in consumer purchasing decisions, they predominantly utilize surveys and field investigations, often focusing on specific product categories.

By contrast, our research represents one of the early attempts to evaluate the impact of the physical setting and potential social learning arising from physical interactions on the collective shopping behaviors within a community, utilizing an extensive dataset of credit card transactions. The advent of large-scale data collections, such as taxi routes, location-specific check-ins, and credit card purchases, offers novel possibilities for investigating individual financial actions, urban dynamics, and municipal economics with enhanced intricacy and unparalleled depth. 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]. An example of this can be found in the work of Krumme and fellow researchers. [Krumme et al. The study conducted by 2013 examined the predictability of consumer merchant visitation patterns and provided evidence that shopping behavior exhibits high levels of predictability over prolonged durations; conversely, Singh et al. [Singh et al. In a study published in 2015, the researchers employed a methodology that involved the analysis of three behavioral factors - diversity, loyalty, and regularity - extracted from the spatio-temporal data of credit card transactions. These factors were utilized to predict an individual's financial well-being, focusing on indicators such as excessive spending, delinquent payments, and occurrences of financial distress as documented through the bank's administrative procedures.

Furthermore, the investigation of mobility patterns has been conducted through the analysis of geo-referenced banking transaction data. Lenormand and others. [Lenormand et al. In a study conducted in 2015, it was found that variations in human mobility patterns exist across different socio-demographic groups, as evidenced by the research conducted by Sobolevsky et al. [Sobolevsky et al. The study conducted in 2014 indicated that economic activity within a country exhibits spatial coherence and corresponds with prevailing administrative boundaries. They have also elucidated several patterns of mobility observed among both local inhabitants and tourists. In a more recent study, Sobolevsky et al. The characteristics of different cities were further studied in 2016 through an analysis of mobility signatures derived from the spending behaviors of their residents. These movement trends have been

leveraged to underscore the vulnerability of individual credit card transaction information, prompting discussions regarding privacy issues. As exemplified by de Montioye and colleagues. [de Montjoye et al. The study conducted in 2015 revealed that a limited set of spatio-temporal coordinates within an anonymized financial dataset could be adequate for the re-identification of individuals with only a small amount of external information.
In conclusion, numerous research endeavors have adopted a data-centric approach to address issues in urban economics, employing techniques predominantly grounded in computer science, such as feature extraction followed by probabilistic reasoning or supervised learning methods. An example of this is demonstrated by Fu and colleagues. Proposed a sparse pairwise ranking model along with a ClusRanking methodology that incorporates geographical dependencies to evaluate and predict real estate prices [Fu et al. 2014a; Fu et al. 2014b]. Meanwhile, Karamshuk et al. Assessed the predictive capacity of characteristics obtained from geographic data and user mobility in identifying the optimal store site for maximizing popularity levels [Karamshuk et al.]. 2013]. The main differences between these studies and our approach are outlined as follows. In the early stages, their aim is to address specific learning challenges such as ranking, with a predominant emphasis on feature extraction and pioneering probabilistic techniques in their primary contributions. On the contrary, our goal is to acknowledge and assess the influence of a precisely established metric, based on physical exposure, on the resemblance in purchasing patterns. Additionally, their main focus is directed towards individual entities, such as neighborhoods or specific locations, with an emphasis on achieving goals like rank. In contrast, our analysis delves into collective behaviors across multiple communities, particularly examining the patterns of co-visitation that will be elaborated upon in subsequent sections.
Despite the increasing number of research studies that analyze quantitative behavioral data, the current literature predominantly treats individual data entries in isolation and lacks the incorporation of social influence and interactions in studying consumer purchasing behavior. The primary objective of our study is to contribute to the evolving research field by providing novel insights into the impact of both the physical environment and social learning on the purchasing decisions of an urban population.

3. DATA AND METHODS

In this part, we initially present our data set and its related statistics, along with a few data processing procedures. We subsequently outline the framework and approach we suggest for examining the similarities in purchasing behaviour among communities.

3.1 Data

We conducted an analysis of more than ten million credit card transaction records provided by a prominent financial institution in an OECD country, encompassing hundreds of thousands of individuals over a period of three months. Each record within the dataset represents a distinct credit card transaction, containing details such as customer and store identification, as well as the timestamp (day, hour, and minute) of the transaction and the expenditure sum in the native currency.
Supplementary information pertaining to the clientele and retail establishments is also included. For customers, we have access to:
- customer age;
- customer gender;
- customer marital status;
- customer education level;
- customer working 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 data has been effectively anonymised at the individual customer level, whereby each customer is denoted by a pseudo-unique identifier, and all distinct identifiers have been removed. The data is

scrutinised in accordance with legal parameters related to re-identification, thereby ensuring full adherence to the privacy regulations of the country.
Data sufficient for reproducing the outcomes reported in this manuscript will be made available online. Our study primarily centers on municipal data, thereby excluding external influences such as inter-city commuting patterns or remote transactions. We focus on the two most populous urban centers in the country, designated as City A and City B. Despite sharing a high population density, each city displays distinct socio-demographic characteristics, as outlined in Table I. We consider customers in each urban center who both live and work within the broader metropolitan area, where their primary activities are concentrated, as well as their transactions at retail establishments located within that geographical vicinity.
As we aim to examine the relationship between physical closeness and/or social learning from being in a similar environment, and people's purchasing choices, we concentrate on five merchant categories1 that primarily align with onsite and discretionary buying, for which we possess the most data:
(1) “amusement and entertainment”
(2) “clothing stores”
(3) “retail stores” (including subcategory “grocery stores, supermarket”)
(4) “personal service providers”
(5) “miscellaneous stores” (including subcategory “eating places, restaurants”)
We also exclude customers with fewer than 20 transactions across these five categories in the three-month period, as we do not view them as regular credit card users. Following the filtering process, 44% of the customers in City A and 46% of the customers in City B have been retained, which still represent 80% of the transactions in both cities.
Following the previously outlined data processing steps, we have 49 thousand customers in City A who completed 2.3 million transactions at 110 thousand shops, and 9 thousand customers in City B with 0.4 million purchases at 30 thousand stores. These are the two data sets utilised for the analyses outlined in this paper. Table I presents various statistics regarding the socio-demographic traits of the two cities. It is evident that City A has a larger number of young (under 30), single, female, and college-educated clients, along with a somewhat higher median income in the 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 part, we initially outline the idea of social connections among urban communities. We subsequently present several behavioural indicators to assess the similarity in community purchasing patterns, which we utilise to analyse the impact of social bridges.
3.2.1. Community connections fostering social relationships. To investigate the correlation between purchasing behaviors within different demographic groups residing in urban areas and their proximity, it is imperative to identify these specific communities. In the context of the observed nation, localities may be delineated as compact administrative units situated within urban areas. The sizes of these residential districts vary significantly, spanning from 0.05 square kilometers in the central urban region to 50 square kilometers on the periphery, where inhabitants generally display common socio-demographic characteristics. More specifically, we consider each locality as a social unit, with its buying patterns influenced by the inhabitants of that particular area. The dataset includes an estimated 800 neighbourhoods in City A and 500 neighbourhoods in City B.
Figure 2 depicts the histograms illustrating a range of statistical data pertaining to the communities in

City A (on the left column) and City B (on the right column), serving as the foundation for the findings discussed in this manuscript. These statistics are consistent with the broader data outlined in Table I concerning the disparities observed between the two municipalities. Moreover, in addition to the fluctuating customer numbers and transaction records, the demographics of City A's communities indicate a higher representation of younger and female consumers, coupled with elevated income levels in comparison to City B. Our recommendation is to create social connections that link each pair of communities I and J, in order to enhance the probability of both physical proximity and shared social knowledge among individuals hailing from these communities. Specifically, we create a social linkage that connects two distinct communities, I and J, by pairing individuals i and j from these communities. These individuals are situated in I and J, respectively, with their workplaces, Li and Lj, falling within a defined proximity threshold, d.

Consequently, the quantity of social connections linking I and J is:

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

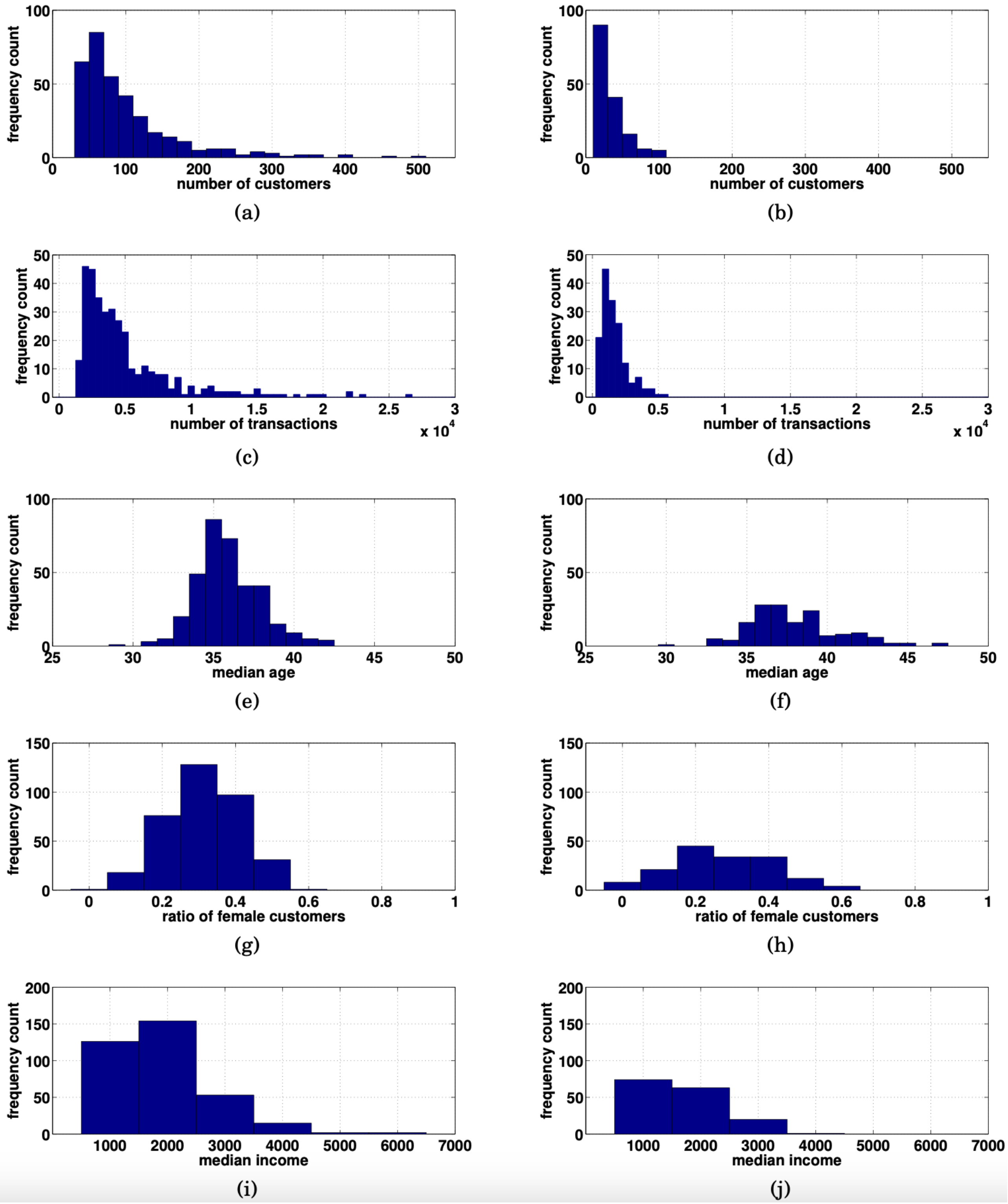


Figure [2]: Histograms of some statistics of the communities in City A (left column) and City B (right column).

where considering individuals' substantial time allocation towards work-related activities, we posit that individuals working at proximate locations, defined by a specific distance threshold d, are more likely to interact with one another due to the regular and frequent exposure facilitated by their close physical proximity, as represented by the distance metric D(Li, Lj).

Viewed through a mathematical lens, in a bipartite graph where the sets of vertices represent two distinct groups of individuals, denoted as sets I and J, the relationships between customers i from set I and customers j from set J are represented by edges indicating whether their work locations are within a specified distance d. The count of social bridges connecting sets I and J, symbolized as bdg(I,J), corresponds to the overall quantity of edges present in this bipartite graph. In our approach, a member denoted as 'i' belonging to community 'I' has the capability to form numerous associations with individuals from group 'J', contingent upon the condition that their places of employment fall within the prescribed distance threshold. The level of social connections between entities I and J is indicated by a numerical value that varies from 0 to the multiplication of the total number of customers in I and J. The illustration in Figure 1 depicts a scenario in which three interactions exist between communities I and J, along with one interaction between I and K.

Clearly, the inclusion of variable d holds substantial importance within our model. In the case where d is equal to 0, a connection is exclusively established between individuals i and j who occupy the same office premises, thereby implying a potential collegial association between them. On the contrary, should the parameter d be calibrated to adequately cover the entirety of the urban area, it would lead to the establishment of social connections between every customer i belonging to set I and every customer j belonging to set J. In Section 4, we first present results pertaining to specific d values in cities A and B, followed by an examination of the influence of varying this parameter on the results. This analysis leads to the identification of a noteworthy observation regarding a potential spatial constraint on the influence of social connections.

3.2.2. Behavioural indices for community purchase behaviour.

We evaluate the similarity or dissimilarity in purchase behaviours between every pair of communities I and J within the city by analysing the following three behavioural indices. It should be emphasised that in the computation of subsequent indices, visits occurring during standard working hours (specifically, from 10:00 AM to 6:00 PM on weekdays) and visits to establishments located within the customer's residential and occupational vicinities are excluded from the analysis. The rationale underpinning this intervention is as follows. Traditional models of consumer behaviour, such as the Huff model, suggest that individuals tend to shop more often in close proximity to their residence or place of work. The occurrence of co-visits in these locations, particularly by individuals in close proximity to their co-working spaces during working hours, could introduce bias into our evaluation. Through the exclusion of these particular purchases, the resultant parameters would capture the convergence in consumer behaviour between individuals in I and J when venturing to locations beyond their residence and professional environment during recreational periods, predominantly shaped by individual preferences.

The initial behavioural metric is the count of distinct stores co-visited by customers in I and J over a three-month duration:

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

where Ci and Cj represent the collections of distinct stores frequented by customers in I and J, respectively, and | · | indicates the size of a discrete set. This metric assesses the similarity in purchasing decisions of communities i and j regarding their purchase selections.

Secondly, we calculate the similarity of the temporal patterns of purchases made by customers in communities I and J. For this purpose, for each community I, we initially calculate a 48-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 Tweekday(t) counts the total number of purchases in the t-th hour on weekdays, and Tweekend(t) counts that in the t-th hour on weekends. We then measure the similar- ity of TI and TJ as follows:

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

where the symmetric Kullback-Leibler (KL) divergence, denoted as KL(·, ·) and originally formulated by Kullback and Leibler in 1951, is defined according to the prescription provided by Johnson and co-authors. According to the study by Johnson and Sinanovic in 2000, the normalization of the divergence is achieved by employing the exponential function to constrain its value within the range of 0 to 1. As the tsim(I,J) value increases, the resemblance between the temporal purchase distributions of I and J becomes more noticeable. In contrast, an alternative approach involves quantifying the similarity between the temporal distributions by utilizing the cosine similarity metric applied to the two 48-dimensional vectors.

In conclusion, we derive the sum of absolute differences in median expenditures within each I and J pair across the aforementioned five merchant categories.

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

where C={“entertainment and amusement”, “apparel shops”, “retail outlets”, “individual “service providers”, “miscellaneous stores”} represents the collection of the five categories we examine, M(c) and M(c) indicate the median expenditure of all the transactions conducted by clients in I and J, correspondingly, at outlets in category c, and |·|1 indicates the L1-norm. Thus, mdiff(I,J) can be considered a metric for the differences in purchasing behaviour between I and J based on the spending levels of customers at stores across the five chosen categories.

In conclusion, the proposed behavioral indicators can be utilized as illustrative metrics for evaluating the comparability/diversity of community shopping patterns along three different aspects: covisit(i,j) for purchase selection, tsim(i,j) for temporal dispersion, and mdiff(i,j) for spending patterns. However, it is important to acknowledge that tsim(i,j) and mdiff(i,j) are associated with temporal distribution and expenditure levels, which may be constrained by temporal and financial factors, consequently lowering their experimental appeal. On the contrary, the measure of covisit(I,J) denotes the degree to which the store choices made by diverse members of the community across the urban area converge as a whole, which we consider to be the most substantial and dependable metric among the three. Therefore, while we present results for all three indices in Section 4.1, our focus in Section 4.2 will be on analyzing the relationship between bdg(i,j) and covisit(i,j).

3.2.3 Evaluation of effect of social bridges.

As detailed in Section 3.2.1, we posit that the number of connections, denoted as bdg(I,J), between communities I and J can be indicative of the probability of proximity and/or knowledge exchange between individuals in I and J due to sharing a similar physical (work) environment. Accordingly, our objective is to investigate the correlation between the degree of social interconnectedness within communities and the similarity in their purchasing behaviors, as evaluated through the three behavioral metrics delineated in Section 3.2.2.

We first analyze the general patterns observed across the three indices as the variable bdg(I,J) increases within communities in Section 4.1. Subsequently, we focus on the index covisit(I,J) and explore its relationship with bdg(I,J) using regression analysis, where bdg(I,J) is considered the independent variable and covisit(I,J) as the dependent variable. Two notable observations can be made regarding our regression analysis. Firstly, it is important to note that our primary interest lies in determining the correlation between bdg(I,J) and covisit(I,J), two forms of dyadic relationships existing within communities. We initiated the analysis by creating two graphs with nodes representing communities and weighted edges signifying the values of bdg(I,J) and covisit(I,J). Subsequently, we transformed the upper triangular portions of the adjacency matrices from both graphs into vectors, followed by applying Ordinary Least Squares (OLS) regression to the resulting data vectors. Given that the elements within each dataset are correlated due to the principles of social connections and shared visitation patterns, we utilize the Quadratic Assignment Procedure (QAP) [Krackhardt 1987; 1988] to evaluate the statistical significance of the derived coefficients and mitigate the potential inflation of statistical significance. The QAP procedure involves the random permutation of the graph's vertices in relation to the dependent variable, followed by a recalibration of the OLS regression. Through repeated implementations of the QAP (100 instances in our study), we obtain coefficients that are consistent with the null hypothesis, which asserts the absence of a statistically significant association between the independent and dependent variables. Therefore, when the original coefficient is situated in an outlier percentile of the null hypothesis distribution, it allows us to reject the null hypothesis and confirm the statistical significance of the observed association.

Secondly, in the regression model, we examine the connection between the independent variable (number of social bridges) and the dependent variable (number of co-visits), while accounting for the influence of potential confounding variables such as:
(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 justification for undertaking this multiple ordinary least squares regression analysis is outlined below. At the outset, as demonstrated by Pan et al. In the study conducted by Pan et al. (year), it was found that... In the year 2013, the probability of forming social ties is intricately linked to both population density and geographical proximity. Moreover, the occurrence of co-visits among communities could potentially be correlated with the size of their population as well as the distance that separates them.
Secondly, income and socio-demographic factors within the communities could potentially impact both social connections and co-visits. We consequently wish to eliminate the influences linked to these factors to analyze the impact of social bridges. Along with the geographical separation between two communities, we have also taken into account the travel time by car between them, calculated using ArcGIS software and the road networks of the respective cities, thereby considering the impact of transportation infrastructure. We discovered that geographic distance is highly related to car travel time ($r^2 = 0.81$ in City A and $r^2 = 0.94$ in City B), thus we do not treat travel time as a control variable.
In conclusion, we utilise an OLS regression model in which the dependent variable is the number of co-visits, while the number of social bridges and other confounding factors serve as independent variables:

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

where pop(I) and pop(J) denote the counts of individuals in I and J, respectively, dist(I,J) represents the geographical distance separating I and J, and {demok(I,J)}(limits from 5 to k=1) along with inc(I,J) indicate the similarities in five socio-demographic factors and income between I and J as outlined in Section 4.2. The coefficient linked to each independent variable quantifies the extent to which that variable affects the dependent variable, while accounting for the influence of all other variables.
Alongside the statistical test in the OLS regression model, we utilise a multiple regression QAP (MRQAP), an extension of the standard QAP for multiple regression cases, to further confirm the significance of the acquired coefficients.

4 Results

In this portion, we showcase the findings that confirm the influence of social connections on molding community buying behaviour. Initially, we showcase in Section 4.1, Section 4.2, and Section 4.3 results based on selected values of d for City A and City B, followed by an exploration in Section 4.4 of how this parameter affects our findings. Ultimately, in Section 4.5, we showcase an example application demonstrating the impact of social bridges in a predictive task.

4.1. Social bridges and behavioural indices

In Fig. 3, we demonstrate the correlation between the number of social connections connecting pairs of communities and the three behavioral metrics outlined in Section 3.2.2. This analysis is conducted on a sample of 352 communities in City A, each consisting of more than 50 customers (first row), and 158 communities in City B with over 20 customers (second row). Communities with a constrained customer base are overlooked to guarantee that computations are conducted with an adequate dataset volume. The values of 50 and 20 are chosen in accordance with the mean number of customers in the localities of City A and City B, correspondingly. The screening procedure has successfully maintained a majority of the client base, with 81% in City A and 75% in City B. The distance thresholds utilised are

0.1 km for City A and 0.2 km for City B, with further elaboration on the derivation of these values available in Section 4.4.
In Fig. 3 The x-axis is utilised to represent logarithmically scaled bins used for data categorisation, while the y-axis shows the mean and the 95% confidence interval (represented by error bars) of the data in each bin. As the social ties between entities I and J, denoted as bdg(I,J), increase, it is noticeable that there is a corresponding increase in various aspects: (i) the number of unique stores visited by individuals in communities I and J collectively, covisit(I,J), tends to rise; (ii) the temporal patterns of their purchasing behaviours, tsim(I,J), tend to align more closely; and (iii) the disparity in median expenditures across five distinct categories, mdiff(I,J), tends to decrease. It is pertinent to note that despite the differing socio-demographic characteristics of the customer base in City A and City B, as well as the substantial variance in data volume between the two cities in our dataset, the findings generally demonstrate similar patterns. This suggests that strong social ties within urban neighbourhoods may influence the consumer behaviour of individuals within those communities.

4.2 Analysis of regression concerning social bridges and purchasing similarity (co-visits)
In this part, we examine the relationship between social connections and similarities in purchases. Initially, we outline the experimental conditions, followed by a discussion of the results from the regression analysis. We illustrate the impact of social connections stemming from varying time and spatial limitations, the roles and genders of clients, and types of retailers.

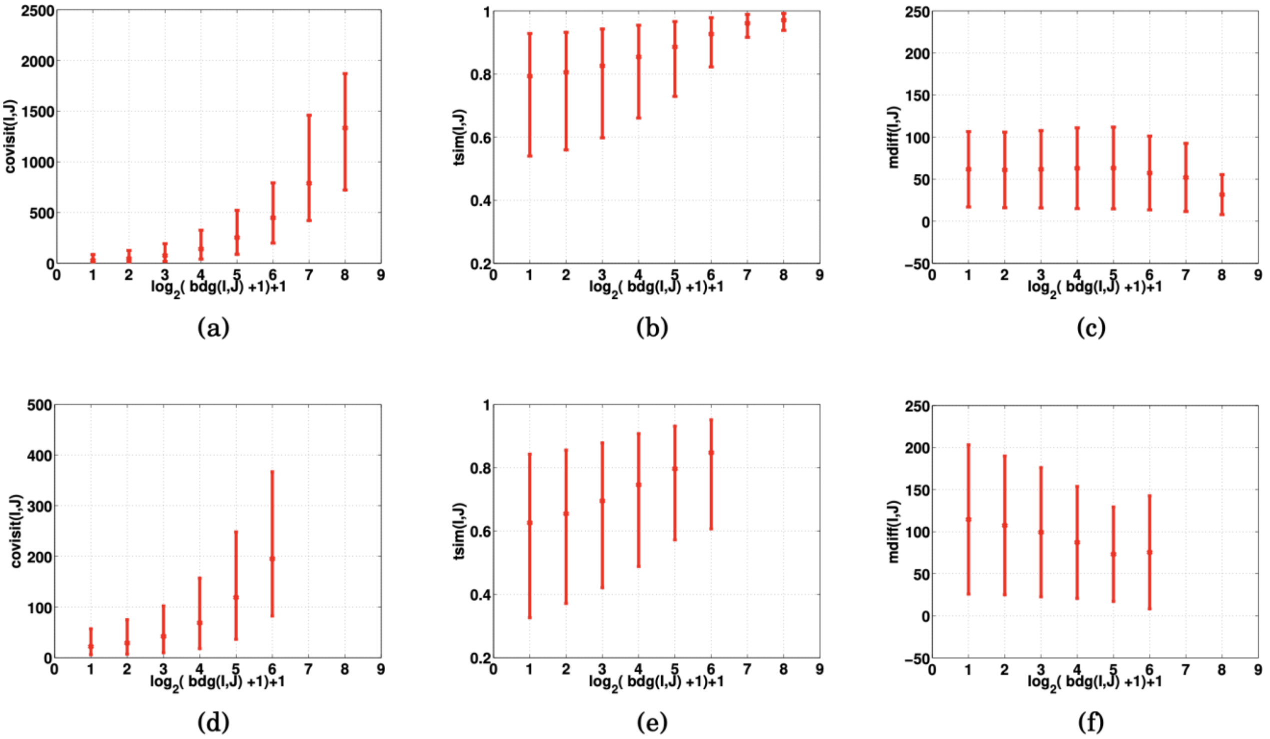


Figure [3]: The connection between the quantity of social bridges linking community pairs in City A (first row) and City B (second row), along with the three behavioral indices: (a)(d) count of distinct co-visited stores, (b)(e) similarity in the timing of purchase distributions, and (c)(f) total differences in median spending amounts in local currency across five categories. The error bars indicate the average and the 95% confidence interval of the grouped data.

4.2.1 Settings for regression analysis.
We employ an OLS regression model in which we regard purchase similarity (quantified by the number of co-visits, covisit(I,J)) as the dependent variable, with the number of social bridges serving as the independent variable, along with potential confounding variables mentioned in Sec. 3.2.34. This enables us to directly contrast the influence of social bridges with other elements that have historically been regarded as models for similarity in purchasing behaviour, specifically, social-demographic and income factors among various communities of city inhabitants.
For this purpose, we initially calculate a discrete distribution for each variable, including income and five socio-demographic factors—age, gender, marital status, education level, and working style—for

each community, utilising set buckets. For instance, regarding age, we calculate the count of customers across three intervals: [0-30], [30-60], and [60-90], which helps us assign a 3-dimensional vector to each community. Likewise, for income, we examine the number of customers within the [0-33], [33-66], and [66-100] percentiles of the entire income spectrum to depict each community as a 3-dimensional vector. For the remaining four variables, we create these distributions by directly utilising various categories within each variable. Subsequently, we calculate the similarity between the distributions of each variable for every pair of communities I and J, utilising the normalised KL-divergence (as in Eq. It seems you have provided a placeholder (4)) and not a specific text to paraphrase. Could you please provide the text you want to be paraphrased? This enables us to construct six similarity graphs, one corresponding to each variable, and treat their vectorised representations as independent variables in our regression model (refer to Eq. (6)).

4.2.2 Regression results.

Tables II(a) and II(b) present the beta-coefficients(5) and the 95% confidence intervals for the independent variables, along with the 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 observed, in both scenarios, the high coefficients suggest that the number of bridges between I and J serves as a significant predictor of similar purchasing behaviour, even when accounting for potential confounding factors like population, distance, income, and socio-demographic variables. Additionally, we perform a robustness check of our findings utilising the jackknife resampling method, with the outcomes displayed in Fig. 9 in the Appendix.

Table II: OLS regression analysis examining the relationship between purchase similarity (i.e., count of co-visits) and the number of social bridges, while accounting for population, distance, socio-demographic factors, and income.

(a) City A

Indicator	Beta coefficients	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	Beta coefficients	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

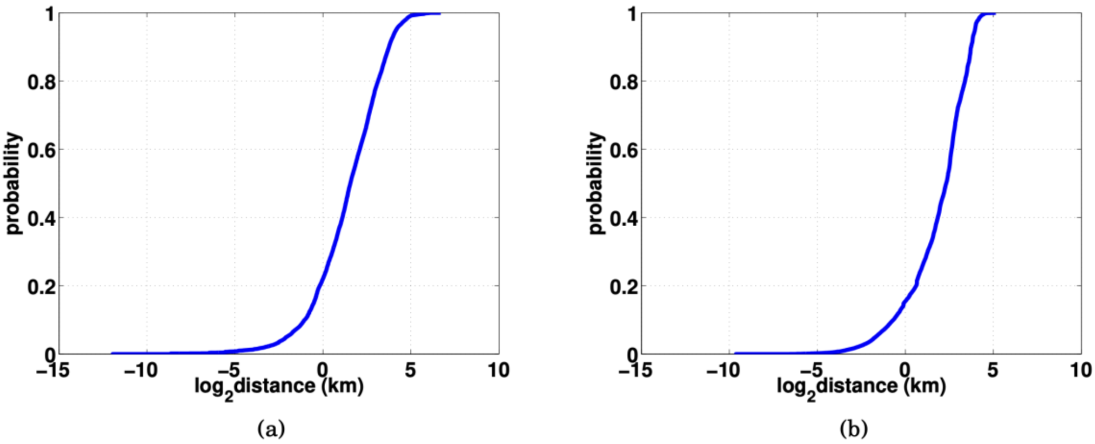
****p < 0.001, **p < 0.01, *p < 0.05*

Although the causal direction of this relationship cannot be established in this instance—meaning it is uncertain whether a high number of bridges causes similar purchasing behaviour or vice versa—these findings indicate that the quantity of social bridges is a significant statistical predictor of purchase similarity. This indicates that physical proximity, here linked to the similarity of work locations, may be more influential than typically recognised elements in influencing buying decisions. The distinction between the suggested and conventional factors is especially notable in City A, likely because of a more dynamic urban setting indicated by its socio-demographic traits compared to City B.

To ensure that the co-visitation trends are not merely a result of the closeness of co-visited shops to co-working sites, we examine the distance distribution between co-visited stores and co-working locations. A co-working location is defined as the midpoint between the workplaces of two individuals who create a social connection. As co-visits are determined at the community level, to examine the proximity of co-visited shops and co-working sites, we proceed in this manner. For every co-visited store shared by communities I and J, we calculate its distance to the nearest co-working space linked to any social bridge between I and J. Figure 4 illustrates the cumulative distribution function of this distance for all co-visited stores (for which we possess location data) across all community pairs. It is evident that co-visits occur at a considerable distance from co-working sites. Indeed, 62% of the joint visits occur over 2 km from the nearest co-working site for City A, while for City B, this figure is 74%. This confirms that the patterns of co-visitation are not merely a result of the closeness of the co-visited shops to co-working spaces.

Figure [4]: Cumulative distribution function of the distance from co-visited store to the nearest co-working space, for (a) City A and (b) City B.

To further demonstrate that the relationship noted in Table II remains strong regardless of the distance between co-visits and co-working sites, we perform a series of regression analyses by excluding co-visits that occur beyond specific thresholds regarding the distance from the co-visited store to the nearest co-working location. Fig. 5 illustrates the variation of regression coefficients for three variables, namely the count of social bridges, population product, and inverse geographical distance, as they relate to the distance threshold applied for defining co-visits in this manner. As observed, the coefficient for the count of social bridges diminishes marginally as the distance threshold rises (likely because in our dataset, long-distance co-visits occur less frequently), yet consistently stays robust and significant.

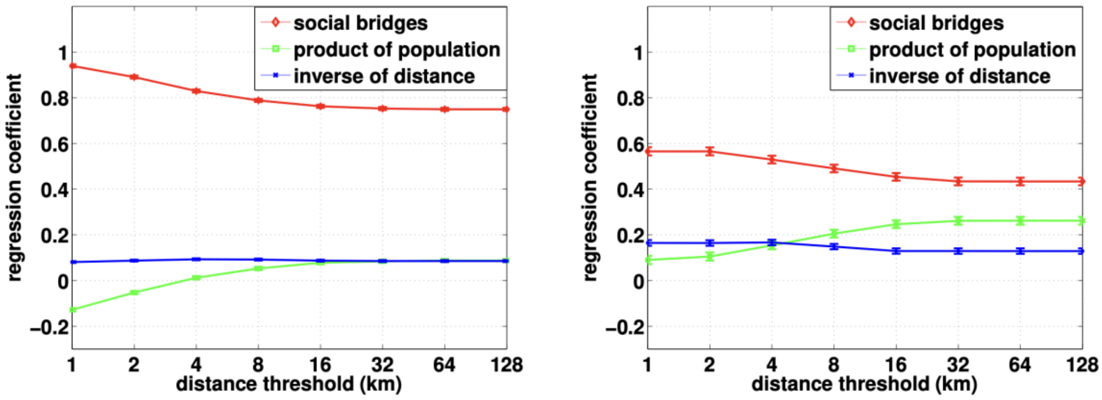


To demonstrate that the relationship noted in Table II remains consistent across varying co-visit time windows, we perform a series of regression analyses, calculating co-visits during five distinct time periods: 1) Weekday 12am-10am; 2) Weekday 6pm-12am; 3) Weekend 12am-10am; 4) Weekend 10am-6pm; and 5) Weekend 6pm-12am. The outcomes of the regression analyses are shown in Table III. It is evident that across all five time frames, the influence of social bridges continues to be robust and considerable.

4.2.3 Comparison between bridge and non-bridge customers.

To enhance the understanding of the regression results presented in Table II, it is essential to analyze the impact of different customer roles within communities I and J on the actual co-visit patterns between these communities. Based on the principle of social bridges, there are distinct customer categories within each pair of communities I and J: (i) "bridge customers" who facilitate connections between I and J, and (ii) "non-bridge customers" who do not engage in this bridging role. This scenario is illustrated in the accompanying figure. First, the connections between I and J are denoted by the red dashed circle, while the connections between J and K are represented by the green dashed circle. It is postulated that bridge customers, due to their shared exposure to a common work environment, may exchange information that reveals communal preferences. We are therefore interested in investigating the following queries: (i) Is there evidence to suggest that bridge customers engage in more co-visits? (ii) Can a relationship be observed between the quantity of social connections and a rise in joint visits, particularly within the subset of customers without any social connections?

Figure [5] :beta-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.



(a) City A (b) City B

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

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

****p < 0.001, **p < 0.01, *p < 0.05*

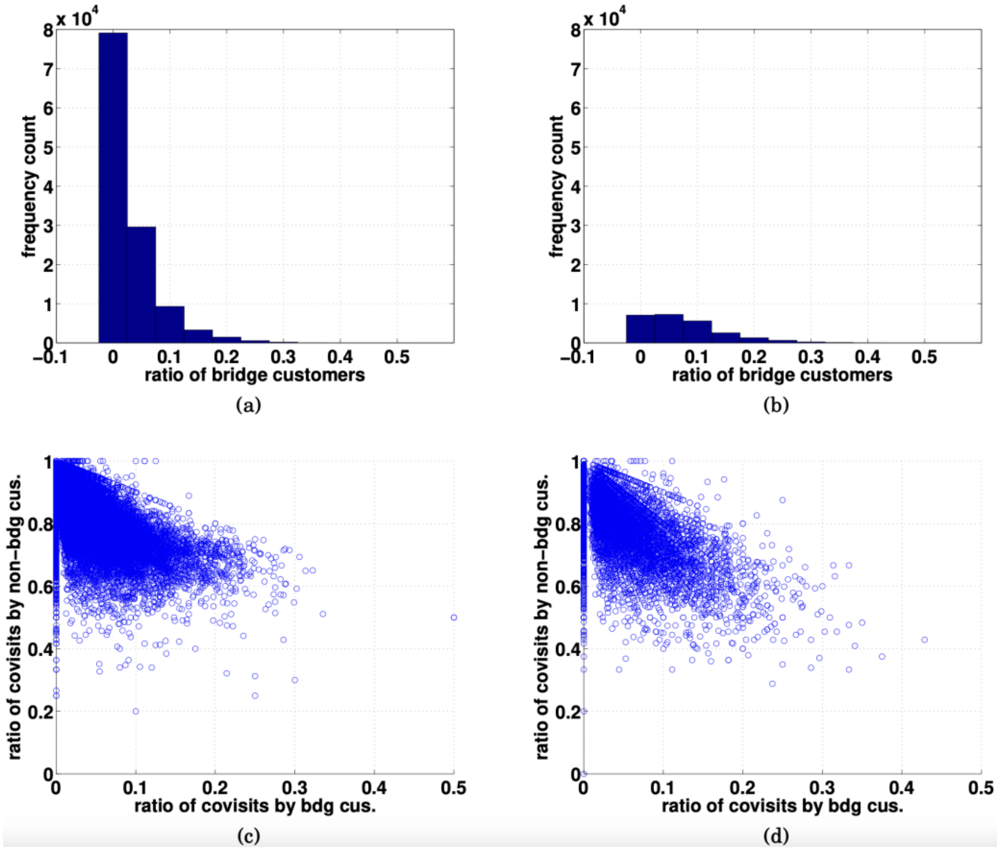
(b) City B

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

****p < 0.001, **p < 0.01, *p < 0.05*

To investigate these questions, we analyze two types of co-visits for each pair of communities I and J: (i) co-visits carried out by bridge customers, and (ii) co-visits conducted by customers who are not utilizing the bridge. Both occurrences are illustrated in the example provided in Figure. 1. In the case of communities I and J, the fraction of bridge customers is 0.5 and 0.65, respectively, while the corresponding ratios of co-visits by bridge customers compared to non-bridge customers are 0.33 and 0.67 for I and J, respectively. Fig. Sections 6(a) and 6(b) present visual representations of the distribution in the ratio of bridge users between all pairs of communities, resulting in two distinct ratios for each community pair within City A and City B. It is evident that the percentage of bridge customers is relatively minimal for each pair of communities. Fig. In addition, section 6(c)(d) showcases a comparison between the co-visit ratios of bridge customers and non-bridge customers for each community pair within City A and City B, respectively. It is apparent that a larger proportion of joint visits are carried out by clientele who do not utilize bridges.

Figures 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.



We then proceed to analyze their respective relationships with the number of social connections between actors I and J, and the results are outlined in Table IV. Initially, it is noted that co-visits by bridge customers exhibit a more pronounced correlation with the quantity of social bridges, evidenced by a larger coefficient, in contrast to co-visits from all customers. This is consistent with our hypothesis that proximity and/or social learning are positively associated with convergent purchasing choices. Significantly, the relationship between the overall number of bridges and the co-visitation by non-bridge clients remains consistently strong within City A. In City B, the correlation is moderate in strength but remains positive and displays statistical significance. Given that the work zones of non-commuting clients form two separate spatial groupings within the urban area, each being at least a distance of d apart, it seems that there is limited spatial intersection among these clients during the workweek. Moreover, their proclivity to patronize a greater number of establishments in the presence of extra bridges could be seen as empirical evidence supporting the impact of social bridges. The exclusion of retail establishments located in residential and commercial zones from the evaluation of co-visit patterns, as detailed in Section. Furthermore, it can be ensured that these outcomes remain unaffected by the personal preferences of individuals regarding the proximity of shops to their homes and workplaces.

Table IV: OLS regression model examining the relationship between purchase similarity (i.e., co-visit frequency) among various customer groups and the number of social bridges, while accounting for population, distance, socio-demographic factors, and income.

(a) City A

Co-visits Types	beta-coefficient	Confidence Interval	**Adj. R^2**

By All	0.760***	[0.754, 0.766]	0.784
By Bridge Customers	1.005***	[0.999, 1.011]	0.766
By Non-Bridge Customers	0.653***	[0.646, 0.660]	0.705

***p < 0.001, **p < 0.01, *p < 0.05

(b) City B

Co-visits Types	beta-coefficient	Confidence Interval	**Adj. R^2**
By All	0.410***	[0.393, 0.426]	0.586
By Bridge Customers	0.717***	[0.700, 0.734]	0.558
By Non-Bridge Customers	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 notable aspect to investigate concerns the potential variation in the effect of social connections on joint visits to different categories of retail outlets. In Table V, the results of a regression analysis are displayed, investigating the relationships between the quantity of social bridges connecting I and J, and co-visits within the four most dominant subcategories of stores within the five overarching categories described in Section. Section 3.1 focuses on a variety of retail establishments, including grocery stores/supermarkets, dining establishments/restaurants, family apparel shops, and pharmaceutical stores/drugstores. It is important to highlight that in both urban centers, the impact of social networks is most pronounced within the restaurant sector, showing relatively less influence on supermarkets. The influence on clothing stores and drug stores/pharmacies tends to vary moderately between the two extremes. This observation is consistent with our natural inclination to exchange information about dining establishments, while displaying a preference for convenience in the context of grocery stores and supermarkets.

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	beta-coefficient	Confidence Interval	**Adj. R^2**
Supermarkets	0.610***	[0.603, 0.618]	0.784
Restaurants	0.812***	[0.805, 0.818]	0.766
Clothing Stores	0.623***	[0.615, 0.631]	0.705
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	beta-coefficient	Confidence Interval	**Adj. R^2**
Supermarkets	0.291***	[0.274, 0.309]	0.586
Restaurants	0.445***	[0.426, 0.465]	0.558
Clothing Stores	0.330***	[0.312, 0.347]	0.490
Drug Stores	0.286***	[0.263, 0.310]	0.182

***p < 0.001, **p < 0.01, *p < 0.05

An important area of consideration with respect to purchasing behavior pertains to assessing the impact of gender. In order to analyze the variances between genders, our study focuses on two specific social connections: female-female connections and male-male connections, along with two classifications of joint visits: visits by female customers not involved in facilitating connections, and visits by male customers not involved in facilitating connections. We then establish a connection between the number of different bridge types and the number of distinct co-visit types, and present the outcomes in Table VI. It is noteworthy that in both urban centers, it is evident that the relationships forged by female clients exhibit stronger correlations with the co-visitation patterns of both non-bridging females and non-bridging males within the localities. This suggests that female shoppers demonstrate greater proficiency in disseminating store-related information and display a heightened sensitivity towards environmental factors and peer behaviors in contrast to male shoppers. Our results are consistent with the findings presented in the study conducted by Hart and colleagues. Previously documented in the literature (Teller & Thomson, 2007; 2012), a discrepancy in buying habits based on gender was emphasised.

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	beta-coefficient	Confidence Interval	**Adj. R^2**
Female-Female	By Non-Bridge Female	0.527***	[0.520, 0.533]	0.625
Restaurants	By Non-Bridge Male	0.404***	[0.398, 0.411]	0.615
Clothing Stores	By Non-Bridge Female	0.360***	[0.352, 0.368]	0.543
Drug Stores	By Non-Bridge Male	0.393***	[0.385, 0.400]	0.604

***p < 0.001, **p < 0.01, *p < 0.05

(a) City B

Bridge Types	Co-visits types	beta-coefficient	Confidence Interval	**Adj. R^2**
Female-Female	By Non-Bridge Female	0.327***	[0.311, 0.343]	0.340
Restaurants	By Non-Bridge Male	0.106***	[0.091, 0.120]	0.468
Clothing Stores	By Non-Bridge Female	-0.073***	[-0.092, -0.055]	0.261
Drug Stores	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

Buying options in a city setting are limited by the merchants' popularity and their locations. To further confirm the impact of social bridges on co-visitation and demonstrate that the apparent co-visitation patterns are not merely a result of these inherent limitations, we examine in this section a null model, specifically, the Huff model [Huff 1964], regarding individual purchase preferences. The fundamental version of the Huff model is represented as follows:

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

$u_{is} = A_s^{\alpha_1}/D_{is}^{\alpha_2}$ where pis, the probability for a customer *i* to visit store *s,* depends on the utility function which is mentioned. In Eq (7), As is a measure of the attractiveness of *s*, *Dis* is the distance between the customer *i* and store *s*, *S* is a set of stores, and alpha 1 and alpha 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 before *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.

Next, we replicate the purchasing decisions of every person in the dataset. In particular, for customer i, who has completed ri purchases, we model his/her purchases through a multinomial distribution characterised by parameters ri and pis. In other words, we draw ri merchants with replacement according to the probability distribution pi·. We view the sampled outcomes as simulated customer purchases. We then merge these simulated purchases with actual purchases at stores lacking location data, and calculate the simulated number of co-visits between each community pair I and J. Lastly, we conduct the same regression analysis as described in Sec. 4.2 to calculate a simulated coefficient for the variable of social bridges and contrast it with the empirical coefficient in Table II. Since the simulated β coefficient is influenced by the purchase counts generated, we conduct the purchase count simulation and regression analysis 100 times to determine a distribution for the simulated β coefficient.

Fig. 7 illustrates the distribution of the simulated β coefficient (blue) regarding the number of social bridges in the regression analysis, alongside the empirical β̂ coefficient (red) presented in Table II. As we factor in actual store purchases lacking location data into our calculation of the simulated co-visits, the coefficients of the null model (Huff model) are somewhat "unjustly" near the coefficients obtained from empirical findings. Nonetheless, we still see in Fig. 7 that there is a notable difference between the regression coefficients in the two scenarios, indicating that, for both cities, the connection between the quantity of social bridges and co-visits is not merely influenced by the Huff model.

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

As previously mentioned, the distance threshold *d* utilised for defining social bridges is a key factor in our analysis. Intuitively, a small d signifies that social bridges are built within a limited area, whereas a large d indicates that they can be established even among individuals who are far apart. In simpler terms, this threshold establishes a circle around each customer's work location, with a radius of *d*, where it is anticipated that the customer will be able to see or engage with another customer (for instance, while heading to the same or nearby locations for lunch or coffee, or while using public transit). As indicated by Pan et al. According to [Pan et al. 2013], the likelihood of two individuals establishing social connections diminishes exponentially with greater distance between them. Therefore, it would be intriguing to examine how the findings in Table IV vary when we progressively raise the distance threshold *d*.

In Fig. 8, we present the coefficients from our regression model based on the distance threshold d used to define social bridges. The values for d are selected to be evenly spaced on a logarithmic scale. The solid curves in red, green, and blue illustrate the coefficients in regression models developed with (i) all customers, (ii) solely the bridge customers, and (iii) exclusively the non-bridge customers, in that order. The dash lines in cyan, magenta, and yellow indicate the respective coefficients following a network shuffling in a MRQAP.

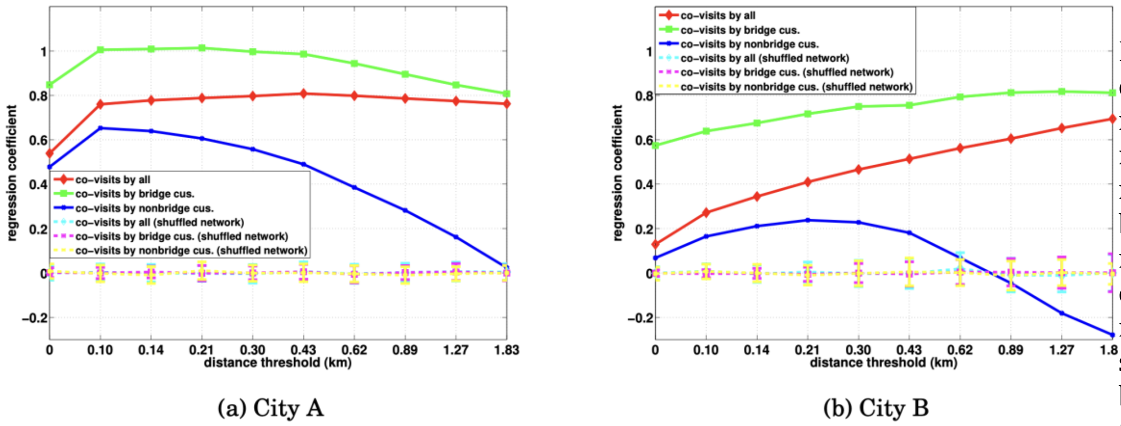


Figure [8]: β coeffficients in the OLS regression model between number of co-visits and number of social bridges, as a function of the distance threshold *d* used for defining the social bridges.

Initially, our test indicates that the derived coefficients are not a result of correlation between the counts of social bridges for various pairs of communities. Secondly, we observe that as d rises from 0, the relationships among the three distinct types of co-visits and the bridge count increase. One potential reason for this behavior is that, when we slightly ease the distance requirement, the standards for forming social bridges become less rigid, as we begin to include individuals who work close enough but not at the exact spot (e.g., within the same office building). This is logical because physical exposure isn't restricted solely to the office building, and by slightly increasing d, we anticipate creating more connections among individuals who are likely to see or engage with one another. As d continues to rise, the green curve stays relatively stable, and the distance between the green and red curves narrows as an increasing number of customers transition from non-bridge customers to bridge customers. Interestingly, the blue curve begins to decline considerably after a specific distance. This indicates that at a certain stage, because of the growing number of bridges, the overall impact of bridge users on encouraging behavioral change within their communities appears to diminish. If we consider that each bridge customer is equally skilled and eager to share information within their community, the distance range pertaining to the area near the top of the blue curve can be seen as a potential geographical limitation for the social bridge effect.

4.5. Predicting co-visitation patterns using social bridges

The connection between social bridges and co-visitation trends among various urban communities allows for several practical uses. In this part, we present an illustrative application that focuses on forecasting co-visitation trends among various communities utilizing the suggested metric grounded in social bridges.

For this purpose, we establish a three-class classification task, categorizing all community pairs into three groups based on the three quantiles of the quantity of social bridges connecting these pairs. This leads to three equally sized groups of community pairs that relate to small, medium, and large levels of co-visitation. We subsequently examine each independent variable in our OLS regression model along with their combination as features for a classification challenge. We train the classifiers using 20% of randomly chosen community pairs (training set), and evaluate their performance on the other community pairs (testing set) regarding prediction accuracy. For classification purposes, we utilize the scikit-learn library [Pedregosa et al. 2011] employing the RBF kernel, and the ideal model parameters are determined through a 5-fold cross-validation using grid-search.
In Table VII, we present the prediction accuracy for various features (indicators), averaged across 20 random divisions of the entire dataset into training and testing sets. As observed, the metric relying on social bridges proves to be more effective than those dependent on traditional factors for forecasting co-visitation patterns among various communities, particularly for City A. Additionally, in both scenarios, incorporating the social bridges metric with all other features enhances prediction performance, illustrating its significance in these tasks.

(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	42.62%	[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%]

(A) 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 research indicates that social connections among city residents, defined here by the physical closeness of their workplaces in various city areas, could explain similarities in purchasing habits. Specifically, we demonstrate that the suggested metric grounded in social bridges serves as a far more potent indicator of similarity in buying behavior compared to conventional factors like income, age, gender, and other socio-demographic elements, even when accounting for potential confounding variables such as geographical distance and population size. Additionally, we demonstrate that the effect noted cannot be solely accounted for by a conventional model of purchasing behavior, such as the Huff model. Consequently, we contend that this similarity may stem from community preferences indicated by physical exposure, as defined by the concept of social bridges.
Our findings further indicate that the influence of social bridges differs among various merchant categories, and there is a gender disparity in the impact of social bridges (i.e., the presence of female

customers in social bridges serves as a stronger indicator than that of male customers). Ultimately, findings derived from various distance thresholds indicate a potential geographical limitation on the influence exerted by social bridges. Our results collectively suggest that our measure rooted in social connections may reflect a type of social learning influenced by physical closeness or shared work surroundings. This resonates with the idea of "the familiar stranger" [Milgram 1977; Paulos and Goodman 2004; Sun et al. 2013], indicating that individuals who see each other often are more inclined to engage than complete strangers, owing to a foundation of common experiences.

Our research is comparable to studies in the computer science field regarding network structure and influence. For instance, it is noteworthy to observe that the bridge customers, who connect residential communities by being employed at nearby sites, can also be viewed as the structural hole spanners described in [Lou and Tang 2013]. Network-based models [Zhang et al. 2012] can also be utilized to examine the influence among customers by creating a geographical network, with nodes symbolizing customers and edges indicating whether they reside or work in proximity to each other. Nevertheless, although these studies focus on creating and assessing algorithms to compute similarities or measure influence among entities, as seen with structural holes in [Lou and Tang 2013] or penalized hitting time in [Zhang et al. 2012], our primary goal is to examine the link between a physical exposure network and a behavioral similarity network, specifically, to assess and confirm the presence of statistical correlation between a straightforward metric derived from physical exposure and co-visitation trends at a collective (community) level.

The literature indicates that word-of-mouth and physical exposure are recognized strong influencers of behavior spread, yet their efficacy in contemporary cities is still unclear. In this paper, we examine the presence and intensity of this correlation by investigating the relationship between physical exposure and shopping habits. We think that the significant correlations identified in this paper would hold practical value since existing techniques utilized in urban planning, policy formulation, and marketing primarily depend on demographic data. Our findings could offer an alternative source of information and strategies for these objectives. For example, in urban planning and policy-making, we can envision that the actions or choices made by planners and policy-makers regarding the (re-)location and/or (re-)design of shopping areas, malls, strips, etc., close to significant workplaces (or upcoming ones) may consider the placement and scale of the key social bridges, thereby enhancing the connections between communities. The idea of social bridges can likewise be utilized to rejuvenate the low economic activity in a region, by examining the potential influx from various communities to the area for economic reasons. Ultimately, in marketing, businesses can leverage the idea by examining the areas where they underperform and by enhancing marketing activities at or near significant workplaces from which, based on the data, the greatest bridge effect will be conveyed to those areas.

Additionally, our study proposes a fresh perspective on raising awareness. This contrasts with conventional ideas that depend solely on social interactions or exclusively on geographic connections to raise awareness. This study proposes that merging both methods (geo+social via indirect links) could be effective in numerous situations. For instance, persuading people employed in the city center may be a more effective approach to influence others regarding vaccination instead of solely targeting those with numerous local ties in the suburb. This method for raising awareness and disseminating concepts applies to the promotion of products, services, and ideological beliefs. From an urban planning standpoint, the integration of individuals in city centers compared to mixed-use residences in the city holds distinct consequences for the dissemination of these ideas and perspectives.

In observational research, there frequently are hidden confounding variables. The influence of social bridges could stem from word-of-mouth and physical presence, as well as other unrecognized factors that may result in shared workspaces and co-visited shops. Although some are linked to or influenced by demographic/income data, such as school district, housing costs, and to some extent, exposure to comparable online and TV ads, which we partially account for in our method, others might still go unexamined. Nevertheless, we discovered that for a specific pair of communities, the proportion of bridge customers is quite low, and a higher percentage of the co-visits are conducted by non-bridge customers. This, along with the findings on the relationship between social bridges and co-visits by non-bridge customers, appears to constrain the influence of overall unobserved variables (that affect both co-working and co-purchase) in favour of the impact of social bridges.

It is important to emphasize that our findings do not suggest a causal link between social bridges and similarities in buying behavior. Nonetheless, even in the absence of a causal relationship, the social bridge effect could be useful for predicting behavior and stratification, targeting campaigns, and allocating urban resources. For instance, by analyzing the shopping habits of a particular community, we could assess the chances of other communities sharing similar preferences, using the idea of social bridges rather than conventional factors like demographics or physical proximity. For instance, we demonstrate that social bridges can more effectively predict co-visitation patterns among various urban communities than traditional factors do. An alternative situation involves the stratification of urban neighborhoods through the use of a clustering method on the social bridge graph. The reliance of the definition of social bridges solely on location information indicates that these applications can also utilize other data sources like mobile phone records or publicly accessible geo-localized social media data.

Regarding causal inference, the outcomes derived from the purchase behavior similarities of non-bridge customers might act as an initial approach to developing causal inference frameworks that verify social influence across various communities in relation to their purchasing behavior. Further research is required to comprehend and simulate, for instance, how social learning or potential interactions among bridge customers result in a specific degree of sharing related to purchases or store information, and how such information might be disseminated to other individuals within their communities. With more longitudinal data, one suggestion is to examine causal relationships and behavioral changes by identifying specific events of influence. For instance, we are presently examining how customers from newly established stores in the city disperse and measure the impact of social networks on this spread relative to conventional influences. It would also be intriguing to explore how long it takes for the impact of social bridges to become evident after a new individual begins working in a place or relocates to a community. These analyses would undoubtedly impact the research on urban economics and data-informed urban planning.

The credit card transaction record data set utilized in our research comprises a random sample of approximately 10% of the complete customer base of the financial institution. Nonetheless, the sampling strategy is crafted to ensure that the resulting sample collection accurately represents the entire customer base. A drawback of utilizing credit card transaction data is that credit card users might only constitute a certain portion of the population, and individuals may choose to pay with cash in various circumstances. Moreover, the data set spans just three months, which may appear restricted when examining long-term and ongoing behavior. Nonetheless, the overall uniformity observed between the two cities with varying demographics indicates that our findings are probably applicable elsewhere.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.
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).
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.
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).
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.
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.
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.
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.
S. Goel and D. G. Goldstein. 2013. Predicting individual behavior with social networks. Marketing Science 33, 1 (2013), 82–93.
Impact on customers' repatronage intentions and gender influence. The Service Industries Journal 27, 5 (2007), 583–604.
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).
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, 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.
C. Hart, A. M. Farrell, G. Stachow, G. Reed, and J. W. Cadogan. 2007. Enjoyment of the shopping experience:
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. 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.
C. Krumme, A. Llorente, M. Cebrian, A. Pentland, and E. Moro. 2013. The predictability of consumer visitation patterns. Scientific Reports 3, 1645 (2013).
M. Lenormand, T. Louail, O. G. Cantu´-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.
S. Kullback and R. A. Leibler. 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 1 (Mar 1951), 79–86.
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. 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.
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.
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.

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.

Appendix

List of merchant categories

The thorough compilation of merchant categories found in our data set aligns with the majority of Merchant Category Codes (MCC) specified in ISO 18245 for retail financial services, which encompasses:

(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 general categories, there is a more specific list of merchant subcategories. We have chosen four of these subcategories from the five main categories in Sec. 3.1 for the examination in Sec. 4.2.4.

In this part, we perform an analysis of normalized social bridge and co-visit indices. In particular, the normalized social bridge index is defined as the ratio of the number of social bridges to the product of the populations of the two communities in the dataset, while the normalized co-visit index is characterized as the Jaccard index.

Given that the combination of visits from two communities, serving as a normalization factor in the Jaccard index, shows a strong correlation with their population product in the dataset ($r = 0.86$ for City

A and r = 0.84 for City B), both normalized indexes effectively account for the population factor. Consequently, we exclude the population favor from the regression analysis, and the outcomes are shown below in Table VIII and Table IX (corresponding to Table II and Table IV of the primary manuscript). In summary, the findings align with those presented in Table II and Table IV of the primary manuscript: For City A, a significant correlation exists between the normalized social bridge and co-visit indexes. In City B, the social bridge effect is moderate yet continues to be positive and statistically significant.

In the regression analysis included in the primary manuscript, we account for the population factor by treating it as an independent variable. The normalized social bridge and co-visit indexes mentioned earlier serve as an alternative to reach the same goal. Because normalized indexes can frequently be expressed in various manners (for instance, an alternative normalized form of the social bridge index could be the ratio of customers creating social bridges between the two communities to the overall number of customers in both communities), we opt to maintain the existing regression framework with the original variables while treating population as a confounding factor in the primary 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	Beta coefficients	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	Beta coefficients	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	beta-coefficient	Confidence Interval	**Adj. R^2**
By All	0.505***	[0.500, 0.512]	0.448
By Bridge Customers	0.498***	[0.491, 0.504]	0.377

By Non-Bridge Customers	0.464***	[0.457, 0.470]	0.414

***p < 0.001, **p < 0.01, *p < 0.05

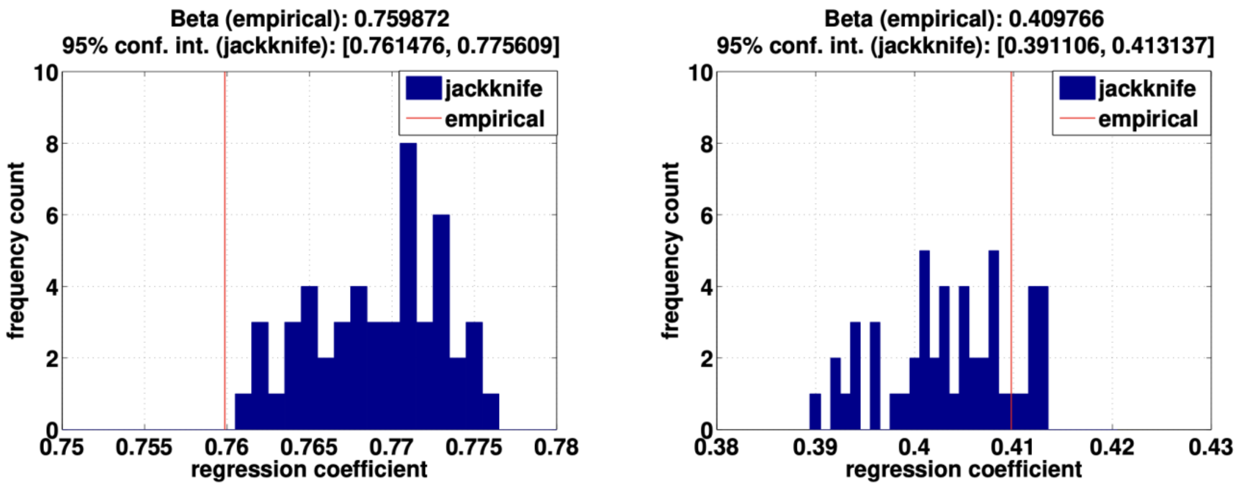
(b) City B

Co-visits Types	beta-coefficient	Confidence Interval	**Adj. R^2**
By All	0.221***	[0.207, 0.236]	0.359
By Bridge Customers	0.344***	[0.328, 0.360]	0.270
By Non-Bridge Customers	0.141***	[0.126, 0.156]	0.318

***p < 0.001, **p < 0.01, *p < 0.05

Robustness check using jackknife resampling

To ensure the reliability of our findings in Table II, we calculated the jackknife estimate of the regression coefficient for the social bridges variable. The jackknife is a method of resampling used for estimating variance [Cameron and Trivedi 2005]. To calculate the jackknife estimate of a parameter, a random sample of the data is consistently excluded from the analysis, and estimates of the desired parameter from various trials are averaged. In particular, we randomly eliminate 5% of the active customer-store pairs from our data set to calculate co-visits, and perform the same regression analysis on secure the coefficient for the social bridges variable. We subsequently carry out this procedure 50 times and calculate the 95% confidence interval of the regression coefficient, with the results displayed in Fig. 9. The confidence interval of the jackknife estimate for the regression coefficient of social bridges is nearly identical to the empirical value shown in Table II, suggesting its resilience to data fluctuations.



(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.