Beyond Distance Decay: Discover Homophily in Spatially Embedded Social Networks

Yang Xu*[1], Paolo Santi[2,3], and Carlo Ratti[2]

[1] *Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University*
[2] *Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA*
[3] *Istituto di Informatica e Telematica del CNR, Pisa, Italy*

Abstract

Existing studies suggest "distance decay" as an important geographic property of online social networks. Namely, social interactions are more likely to occur among people that are closer in physical space. However, limited effort has been devoted so far to quantifying the impact of "homophily" forces on social network structures. In this study, we provide a quantitative understanding of the joint impact of geographic distance and people's socioeconomic characteristics on their interaction patterns. By coupling large scale mobile phone, income and housing price datasets in Singapore, we reconstruct a spatially embedded social network that captures the cellphone communications of millions of phone users in the city. By associating phone users with their estimated residence, we introduce two indicators (communication intensity and friendship probability) to examine the cellphone interactions among places with various housing price values. Our findings suggest that, after controlling for distance, similar places tend to have relatively higher communication intensity than dissimilar ones, confirming a significant "homophily" effect as a determinant of communication intensity. However, when the analysis is focused on the formation of social ties, the "homophily" effect is more nuanced. It persists at relatively short distances, while at higher distances a tendency to form ties with people in the highest social classes prevails. Overall, the results reported in this study have implications for understanding social segregation in cities. In particular, the physical separation of social groups in a city (e.g., residential segregation) will have a direct impact on shaping communication or social network segregation. The study highlights the importance of incorporating socioeconomic data to the understanding of spatial-social networks.

Key words: social network, homophily, distance decay, mobile phone data, segregation

1 Introduction

In the past two decades, Internet and telecommunication technologies have permeated almost every aspect of human life, transforming the ways people conduct their daily activities. One important dimension of human life that has changed dramatically is social interaction. Technological advancements have created a "virtual space" [1], where new forms of social communications are emerging and evolving. These new social channels — such as mobile phones, emails, and online social media — empower people to connect with others who are thousands of miles away. Unlike face-to-face communications that require people to go out and meet in physical space, social interactions in virtual space hardly demand any travel, and seem to be not constrained by geographic distance. On the other hand, the dynamics according to which social ties are created and evolve in an online network have only started to be unveiled, and it is reasonable to assume that pre-existing social ties based on face-to-face relationships have a strong influence on the formation of ties in online social

*Electronic address: yang.ls.xu@polyu.edu.hk

networks. So, although formally there is no geographic constraint on the formation of ties in online social networks, it is possible that geographic properties play a role also in online social networks due to their tight connections with pre-existing (physical) social networks [2, 3, 4].

Debates have emerged on how or whether new information technologies will change the geographic properties of human social interactions. The reflections on the "death of distance" [5] and "the end of geography" [6] are among the early works that call for a reconceptualization of space and place in the information age. Although relevant debates are still ongoing, it is widely acknowledged — sometimes as common sense — that human social interactions do not occur parallelly in physical and virtual space. Instead, they are continuously blending into each other. In other words, "only by maintaining linked, relational conceptions of both new information and communications technologies and space and place will we ever approach a full understanding of the inter-relationships between them" [6] (pp. 181).

Inspired by these reflections, scholars started exploring the geographic properties of various online and mobile social networks, with considerable focus on the "distance decay" effect. This distance effect has been observed across different types of datasets and networks, such as Facebook communities [7, 8], networks of bloggers [9], and telecommunications [10, 11, 12]. One consensus reached by these studies is that the intensity or probability of human communications between places decays with geographic distance. This observation suggests that the relationship between virtual and physical space is strong: social interactions are more likely to occur among people that are closer in physical space.

The research finding is not too surprising because we have fewer opportunities to know someone who is far away. In other words, the lack of physical interactions has a notable impact on the geographic dispersal of social networks. One important issue that has not been addressed, however, is whether the observed *distance decay* effect is purely a reflection of the decreasing opportunities for potential human interactions. As human beings, we tend to connect with similar others. The presence of "homophily" plays an important role in shaping social network structures [13]. Then, an interesting question worth investigating is *how the socioeconomic characteristics of people and their distribution in physical space — for example, in a city — would affect the geographic properties of online social networks.*

Answering this question has many implications for urban planning. Social segregation — a long standing research topic in geography and sociology [14, 15, 16, 17, 18] — is a good example. If online social networks exhibit a distance decay effect, then a city where rich and poor people are highly separated and meanwhile clustered are likely to suffer from a certain level of segregation. The implication is beyond the traditional understanding of residential segregation [19]. It would imply that the physical separation of social groups will directly contribute towards emergence of a "communication segregation" in the virtual space. Hence, a secondary question worth investigating is whether the socioeconomic configuration of a city is the sole driving force of online social segregations. To be more specific, *do people connect simply because they are close? When distance being equal, are people more likely to interact with similar others*?

To answer these questions, we perform a case study in Singapore by analyzing a large-scale mobile phone dataset that captures the communication patterns of 2.6 million people during a period of 50 days. A spatial-social network is established by embedding phone users into geographic space based on their estimated residence. We explore the geographic properties of the network by examining the communication intensity and probability of social ties among different places. By further integrating income data and a high-resolution housing price dataset, we examine whether places with similar socioeconomic characteristics tend to maintain higher levels of cellphone interactions. Different from previous studies that focus on the distance decay effect, this study aims to unravel the joint impact of geographic distance and homophily on the social network structure. We argue that people's social interactions are not only affected by their physical proximity, but also by forces of homophily in the society. The research findings have many implications for policy makings that aim to foster social integration in cities. The research framework can be applied or extended to better understand other types of spatial-social networks.

2 Theoretical Context

The notion of social network provides a structural representation of human relations and interactions. As a fundamental concept in social science, it has generated broad interests across many disciplines (e.g., sociology, geography, transportation, physics, and computer sciences). Social networks emerge as a reflection of personal relations in societies [20]. Such relations are created or maintained through different kinds of human activities, many of which would take place, or are empowered by what happened, in physical space. Therefore, studies of social networks often consider a spatial context of social structure, be it explicit or implicit.

Over the years, many studies have incorporated a spatial dimension into social network analysis [21, 9, 22, 23]. These efforts involve the conceptualization of social actors (i.e., people), relations (i.e., social ties), and their linkage with built environment. Studies of spatial-social networks often link social actors with physical locations — like one's home or neighborhood — such that contextual information can be leveraged to better understand the factors that contribute to or hamper the formation of social relationships.

Enabling a spatial view of social networks is essential. A typical example is the discovery of distance decay effect on social relations [24]. Namely, people tend to socialize more with ones that are close. As face-to-face interaction is a key form to maintain social relations, the distance effect can partly be explained by the travel cost that is incurred for conducting social activities. Thus, travel behavior or mobility is considered as an important dimension that explains the interplay between social relations and distance [3, 25, 26].

In the past two decades, many online (e.g., Facebook) and mobile social networks have emerged. Studies found that the chances people form relations in these networks are still notably affected by geographic distance [9, 7, 8]. Although online social interactions in principle can occur without a need of traveling in the physical space, the rediscovery of distance effect suggests that online social networks partly mirror pre-existing social ties, which are shaped by various constraints in the physical world. Thus, the spatialization of online social networks could provide additional insights into human interactions in an online-offline setting.

Beyond reflecting existing social structures, social networks have been found to influence future activities and travels [3]. For example, studies found that telecommunications between phone users are indicative of their co-location patterns, a reflection of their social interaction potentials [4, 27]. Therefore, mobile and online social networks would have an impact on travel behavior and physical activities, which in turn further affect network dynamics and evolution. Thus, there is a mutual effect between human travel and online social interactions.

Social relations are not only constrained by geographic distance. Studies suggest that homophily, or similarities in people's sociodemographic characteristics, are catalysts of social interactions [13]. From the perspective of travel behavior, great satisfaction might be obtained from interactions between people with similar background, and therefore, individuals are "willing to trade-off extra (travel) cost" (pp.142) for these interactions [24]. As one's social background can be largely explained by the underlying built environment (e.g., income, racial makeup), the homophily principle would imply stronger social connections among locations with similar characteristics.

Although the distance and homophily effects have been studied separately, their joint impact on social network structures remains underexplored. In particular, there is a lack of research on quantifying such impact at intra-urban scales and over social networks empowered by modern information and communications technologies (ICTs). This is partially due to the difficulty of coupling large-scale human interactions with fine-grained sociodemographic data. Filling this research gap, as this study attempts to do, has important implications for cities. For instance, if stronger social connections (e.g., telecommunications) are observed among similar places (e.g., by income or housing price) after geographic distance is controlled, this would imply more social travels in the past or more interaction potential in the future. In other words, socioeconomic configurations in cities would have an impact on human interactions in the online space, on their future travel behavior (in physical space), and more importantly, on socioeconomic segregation in the urban environment [28, 15, 18].

3 Research Design

3.1 From mobile phone data to city-scale social network

The CDR (Call Detail Record) dataset was collected from a major mobile phone operator in Singapore. The anonymized dataset tracks the communication patterns and location footprints of 4.4 million phone users during a period of 50 days in 2011. When a phone call or text message was initiated by a user, a record was generated by the telecommunication system, documenting the unique ID of the caller and the callee, the event type (i.e., call/SMS), the associated timestamp, and the cell towers that the users were connected to. Such information allows us to not only extract social network structure, but also infer the frequented locations of users (e.g., home), from which the social network can be embedded into geographic space.

Note that CDRs are passively generated and the observations for some users can be quite sparse. By measuring the number of days with records for each user, we find a large variation of the level of activity across the population (Figure 1A). To control the data sparsity issue, this study focuses on a subset of users who have at least 10 active days of phone usage. On the one hand, this choice allows us to focus more on local residents by filtering short-term subscribers such as tourists. On the other hand, it would ensure that the number of observation days for the retained users are high enough to support a reliable estimation of home location. After this filtering, we are left with a dataset of 2.6 million phone users.

We then extract the social network structure by measuring the communication patterns among cellphone users. Social ties (links) are established between users (nodes) who have at least one reciprocal contact during the study period. The associated tie strength is measured as the total number of calls/SMS. Note that we only retain reciprocal contacts because some of the one-way communications do not necessarily reflect human interactions (e.g., robocalls). Performing this step gives us a social network with a skewed degree distribution (Figure 1B). The average and median number of social contacts per person are 9.7 and 7.0, respectively.

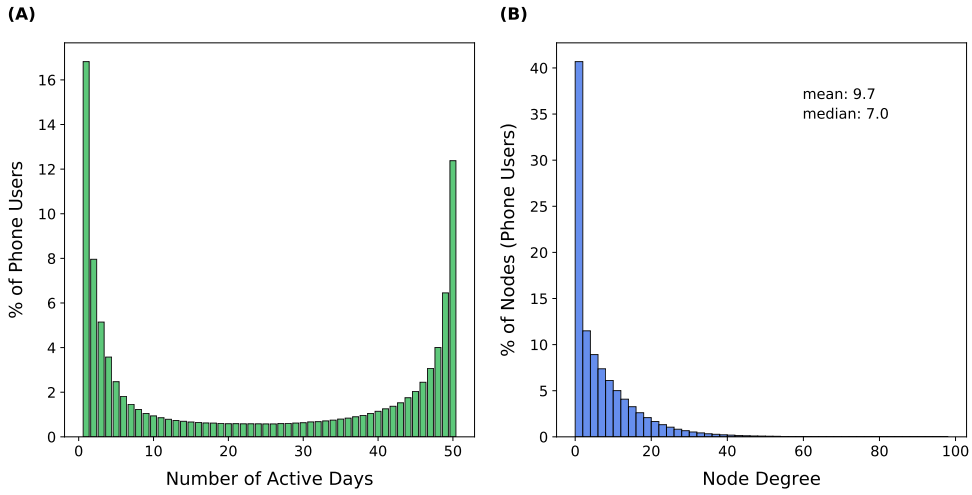

Figure 1: (A) The distribution of number of active days of users; (B) Degree distribution of the extracted social network.

3.2 Embedding social network into geographic space

We embed the social network into geographic space based on the spatial properties of the nodes. Previous studies have suggested that social ties are more likely to be observed at shorter distances as individuals tend to interact more with their spatial neighbors [7, 29]. One usual practice, therefore, is to assign a "home location" to users in order to embed nodes into geographic space [29]. This

study adopts this strategy by first estimating the home location of each phone user. The estimated location is then used as the node property for spatial embedding.

Inferring home location from CDR data has been investigated extensively [30, 31, 32, 27]. In this study, we adopt the method from [27], which estimates each user's home location as the most used cellphone tower before 06:00 and after 19:00. After performing the home location estimation, we compute the number of phone users in each planning area of Singapore[1] and correlate it with the official census population from the Singapore Department of Statistics (http://www.singstat.gov.sg). As shown in Figure 2A, we observe a strong correlation between the two variables, which indicates the robustness of the estimation. Following this step, we perform the spatial embedding by assigning users to their estimated home cellphone towers (Figure 2B).

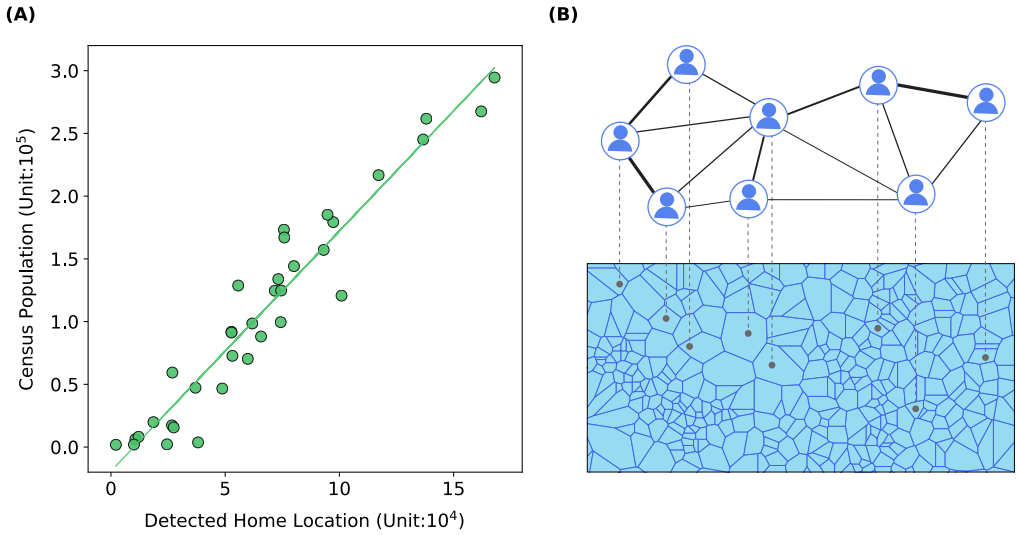

Figure 2: (A) By aggregating phone users based on their estimated home location, we compute the number of users at the level of Singapore's planning area. The numbers are highly correlated with the official census population (Pearson's $r = 0.96$, p-value < 0.001); (B) The social network is embedded into geographic space by assigning phone users to their home cellphone tower (Voronoi cells are used to approximate towers' service areas).

3.3 Extracting social-spatial properties of the embedded network

We introduce two indicators, namely the *normalized communication intensity* and *friendship probability*, to investigate the spatial properties of social network connections. Given any two cellphone towers L_i and L_j in the embedded network, the normalized communication intensity between them, $I(L_i, L_j)$, is defined as the total communication intensity among all the phone users that are concerned, normalized by the total possible friend pairs between these two locations:

$$I(L_i, L_j) = \frac{Strength(L_i, L_j)}{N(L_i) * N(L_j)} \tag{1}$$

Here $Strength(L_i, L_j)$ denotes the total number of calls/SMS exchanged between phone users assigned to L_i and L_j. Note that within-cell communications are not counted. $N(L_i)$ and $N(L_j)$ refer to the total number of phone users (nodes) assigned to L_i and L_j, respectively. In other words, $N(L_i) * N(L_j)$ describe the total possible social links that can be established between the two locations.

While $I(L_i, L_j)$ measures the likelihood of communications between a pair of locations, the value is largely affected by the tie strength, i.e., communication frequency between social links. Therefore

we compute another measure, the *friendship probability*, by considering only the presence of social ties:

$$F(L_i, L_j) = \frac{Ties(L_i, L_j)}{N(L_i) * N(L_j)} \quad (2)$$

Unlike other online social networks (e.g., Facebook) in which social ties are well defined, cellphone communications can only provide an indication of people's social relationships. To address this challenge, we define social links as cellphone user pairs with at least one reciprocal contact during the study period. Although this is not a perfect measure, a previous study based on large-scale telecommunication data suggests that a significant proportion of phone users with at least one reciprocal contact have shared the same place with each other at the same time [4]. These co-location patterns are indicative of coordination calls and face-to-face interactions. Therefore, the definition of social links here provides a reasonable proxy of existing social ties between phone users, and also excludes one-way communications (e.g., advertisement, robot calls) that do not reflect interpersonal relationships.

Thus, the $Ties(L_i, L_j)$ in equation 2 denotes the total number of social links between phone users assigned to L_i and L_j. Since the distance between cellphone towers can be easily computed, the two indicators (I and F) can thus be used to examine the spatial properties (e.g., distance decay) of social network connections.

3.4 Incorporating socioeconomic characteristics of locations

To evaluate the impact of locations' socioeconomic similarities on the social network structure, we consider another housing price dataset. The dataset, which is acquired from a private company in Singapore, includes information of thousands of residential properties collected between 2011 and 2012. Each record in the dataset documents the information of a single housing property, such as the property type (condo, landed or HDB[2]), geographic coordinates (latitude and longitude) and the total sale price of one housing unit. Since users are assigned to their home locations, the housing price dataset can be used to reflect the socioeconomic characteristics of places and the corresponding phone users.

We use the housing price dataset instead of income data collected by census because individual housing properties provide a fine-grained view of the socioeconomic characteristics of the places. The dataset can be well integrated with the spatial-social network at the cellphone tower level. The income data from census is usually reported at a coarser resolution and cannot capture well the spatial heterogeneity.

Before we use the dataset to label locations in the network, we perform a correlation analysis between housing price and income at the level of planning areas. The income data is acquired from the 2011 Household Interview Travel Survey (HITS). By extracting individuals from HITS who reported their monthly income (12,111 in total), we compute the average monthly income of respondents in each planning area. We then compute the average sale price of housing units in each area. We find that the two variables match each other relatively well except for a few outliers (Figure 3A). Through further exploration, we think this inconsistency is partially caused by the sampling bias during the household interview survey. For example, only two respondents were sampled in Southern Islands and both of them reported a monthly income of 500 SGD. However, the island is well known for its many luxury residential neighborhoods. After removing these three outliers, the Pearson's correlation coefficient between the two variables increase to 0.88 (p-value < 0.001), which suggests that housing price could well indicate the socioeconomic level of places (Figure 3B).

In this study, we use the average housing price to label each cellphone tower. Since cellphone towers are point features, we use Voronoi cells to approximate their service areas (Figure 2B). We then identify the housing properties that fall within each cell and compute the average housing price to label the corresponding cellphone tower. After this step, each cellphone tower L_i will be labeled using a housing price value $Pr(L_i)$. This value empowers us to examine the normalized

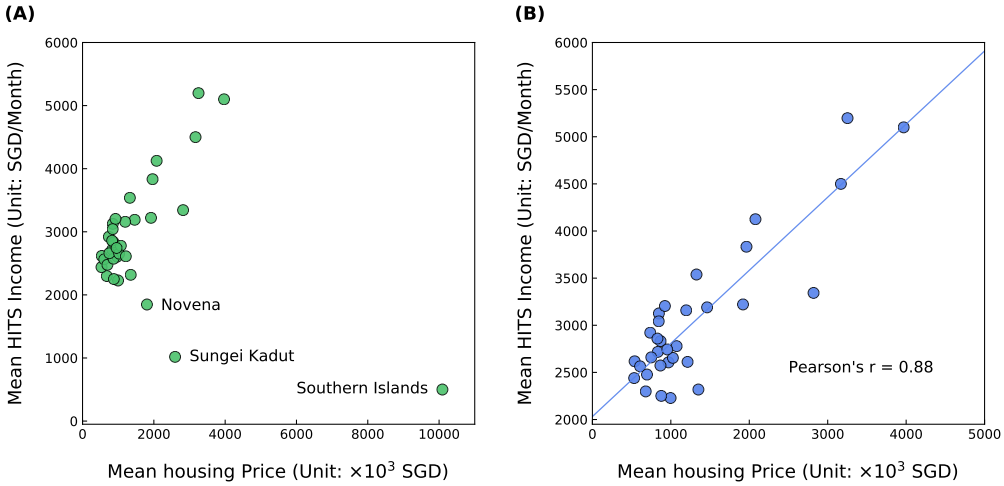

Figure 3: (A) The relationship between mean housing price and average monthly income at the level of Singapore's planning area; (B) The correlation between the two variables after removing the three outliers.

communication frequency and friendship probability between places with varying socioeconomic characteristics. Note that for each Voronoi cell, we have computed the standard deviation of the housing price values (within-cell std), and compared it with the overall standard deviation of the city. We find that the median ratio of with-cell std and overall std is 0.16. The result suggests that housing price values are relatively homogeneous within the Voronoi cells, and our approach could capture the heterogeneity of residential housing price in Singapore.

4 Results

4.1 Distance decay effect of cellphone communication network

In this section, we examine the distance decay effect of the cellphone communication network. Following the definitions of $I(L_i, L_j)$ and $F(L_i, L_j)$, we measure normalized communication intensity and friendship probability by aggregating pairs of cellphone towers separated at different distance values (d). The functions of $I(d)$ and $F(d)$ allow us to explore the impact of geographic distance on the network structure. As shown in Figure 4, both functions show a linear trend at the log-log scale, which indicates that each of the two variables ($I(d)$ and $F(d)$) and geographic distance (d) generally follow a power law. Fitting the two functions yield an exponent of -0.82 and -0.74, respectively. In particular, the likelihood of cellphone communications among people who live at a distance of d follows a distance decay $I(d) \sim d^{-0.82}$, while the probability of social ties at a given distance follows $F(d) \sim d^{-0.74}$. Note that the observed exponents (-0.82 and -0.74) are larger than the ones derived from some other studies using online social networks. For example, a study finds that friendship probability observed from the LiveJournal network follows a distance decay $p(d) \sim d^{-1.2}$ [9]. Another study based on Facebook data found that friendship probability is inversely proportional to geographic distance $p(d) \sim d^{-1.0}$ [8]. Compared to these online social networks, cellphone communications observed in this study decay more slowly with geographic distance.

The distance decay effect suggests that cellphone communications (I) or social connections (F) are more likely to occur among people whose residences are close to each other. An intuitive interpretation of $I(d) \sim d^{-0.82}$ is that neighborhoods separated at a distance of $10km$ tend to exhibit only 15% ($10^{-0.82} \approx 0.15$) of the communication strength compared to neighborhoods that are $1km$ apart. Similarly, friendship probability between neighborhoods decreases to 18% ($10^{-0.75} \approx 0.18$) when distance reaches $10km$. Note that $I(d)$ decays faster than $F(d)$, meaning that people tend to

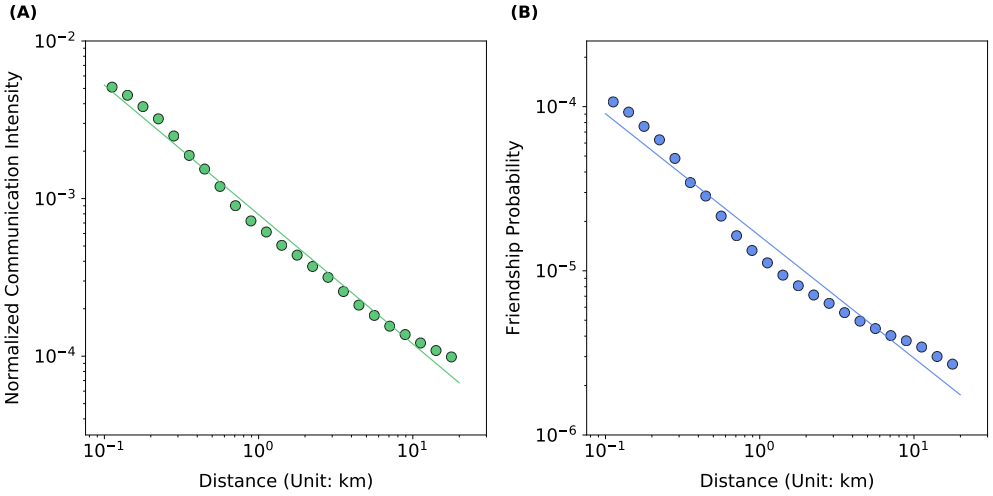

Figure 4: (A) Normalized communication intensity decays with geographic distance. Fitting the data with power law function yields $I(d) \sim d^{-0.82}$; (B) The probability of friendship also shows a distance decay effect $F(d) \sim d^{-0.74}$.

preserve more friendship with distance. However, the intensity of communication with these friends is weaker than with closer friends (i.e., "good friends are close to home").

The distance decay effects have notable implications for socioeconomic segregation in cities. If a city has a high level of residential segregation — namely, residents with similar socioeconomic characteristics tend to live close to each other — the city is likely to have a certain level of communication (or social-network) segregation. In other words, information tends to spread among similar others in a city with severe residential segregation.

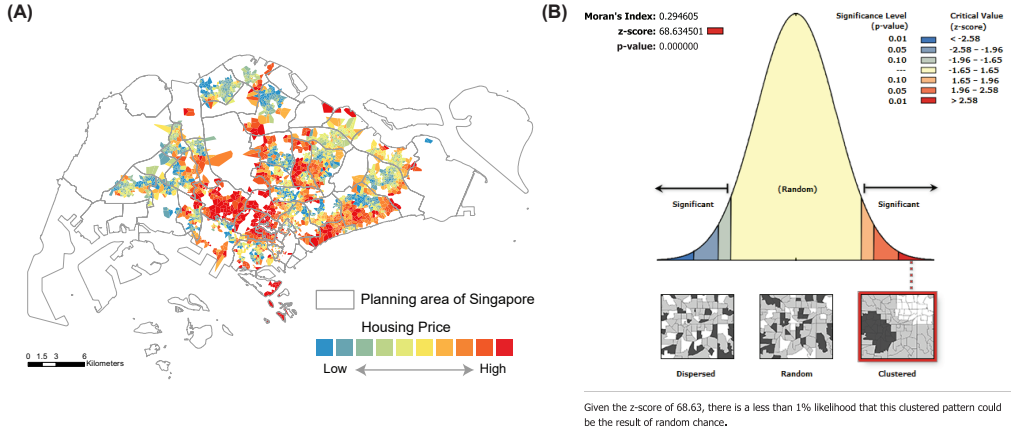

Figure 5: (A) The spatial patterns of average housing price at the level of Voronoi cells; (B) Global Moran's I is computed over the housing prices at these Voronoi cells. The analysis and report are derived from ESRI's ArcGIS product.

To better understand the spatial configuration of housing prices in Singapore, we measure the spatial autocorrelation based on the average housing price derived at the Voronoi cells (Figure 5A). In particular, we compute the Global Moran's I using the built-in function provided by ESRI's ArcGIS product. We use Euclidean distance to measure the distance between cells and the inverse distance method to conceptualize their spatial relationships. As shown in Figure 5B, the analysis yields a Moran's Index of 0.29 with a z-score of 68.63. This indicates that housing prices — which highly

reflect the income level of residents (Figure 3B) — are highly clustered. The clustering patterns suggest that neighborhoods with similar housing price values are generally closer than dissimilar ones. Given the distance decay effect of cellphone communications, the clustering patterns would imply more interactions between these similar neighborhoods, which contribute to a certain level of communication segregation.

4.2 Impact of socioeconomic characteristics on communication intensity

Previous studies have discussed the existence of homophily in social networks, namely, people's social connections tend to be homogeneous with regard to many sociodemographic characteristics [13]. Given the distance decay of cellphone communications observed from the mobile phone dataset, an intriguing question is whether people tend to connect with similar others when geographic distance is controlled. While this cannot be investigated at individual level due to lack of personal socioeconomic data, we examine whether places with similar housing prices would have relatively more interactions with each other that what would be predicted by distance decay effect.

To achieve this, we sort the Voronoi cells based on the average housing price derived in section 3.4. We divide them into five classes such that each class of cells covers 20% of the population (i.e., phone users). We label them from C_1 to C_5, with averaging housing price sorted in ascending order. In other words, C_1 denotes areas with the lowest average housing price while C_5 generally points to the rich neighborhoods. We adopt this classification method instead of other alternatives (e.g., separating top 1% of population from others) such that each class covers adequate amounts of phone users or Voronoi cells. It ensures that we have enough observations of cellphone communications among different classes at varying distance values.

For each pair of class C_i and C_j, we measure their normalized communication intensity at different distance values. Figure 6 shows the results. Each subplot demonstrates the communication intensity from each class to the five classes. Similar to the finding in Figure 4A, the communication intensity between classes decays with geographic distance. Interestingly though, when looking at the interactions among classes, we find that cells with similar housing prices tend to have higher communication intensities when distance is controlled. For instance, C_1 tends to interact more with their own or nearby classes at a variety of distance values (Figure 6A). This homophily effect is even more obvious when looking at C_5 (Figure 6E), which consistently maintain the highest level of interaction within their own class.

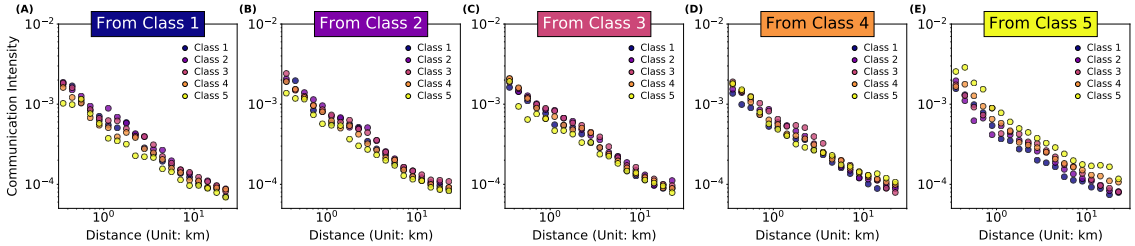

Figure 6: Normalized communication intensity among the five classes.

To further elaborate this, we visualize, for each class, the classes that have the most and second most interactions with them at different distance values. The results are shown in Figure 7. As can be seen, C_1 and C_2 tend to interact more with their own or nearby classes at most of the distance values. Again, class C_5 exhibits a relatively stronger homophily effect as compared to other classes. When distance is being controlled, these places will always have the strongest interactions with their peers. Class C_4 present a pattern that is different from the other classes. We find that C_4 does not exhibit the highest level of interaction with themselves. Rather, they tend to interact more with classes one or two step downward. It is possible that C_4, which points to many people in the upper middle class, tends to interact more with lower socioeconomic classes. For instance, many occupations that are well paid (e.g., doctors, lawyers) have frequent communications with different

socioeconomic tiers. While for class C_5, which points more to the extreme wealth, a clear "rich club" effect can be observed. Summarizing, we can conclude that, while homophily seems to have a significant effect on the intensity of communication, other factors such as social structures are likely to have a significant, and possibly heterogeneous across social classes, effect as well.

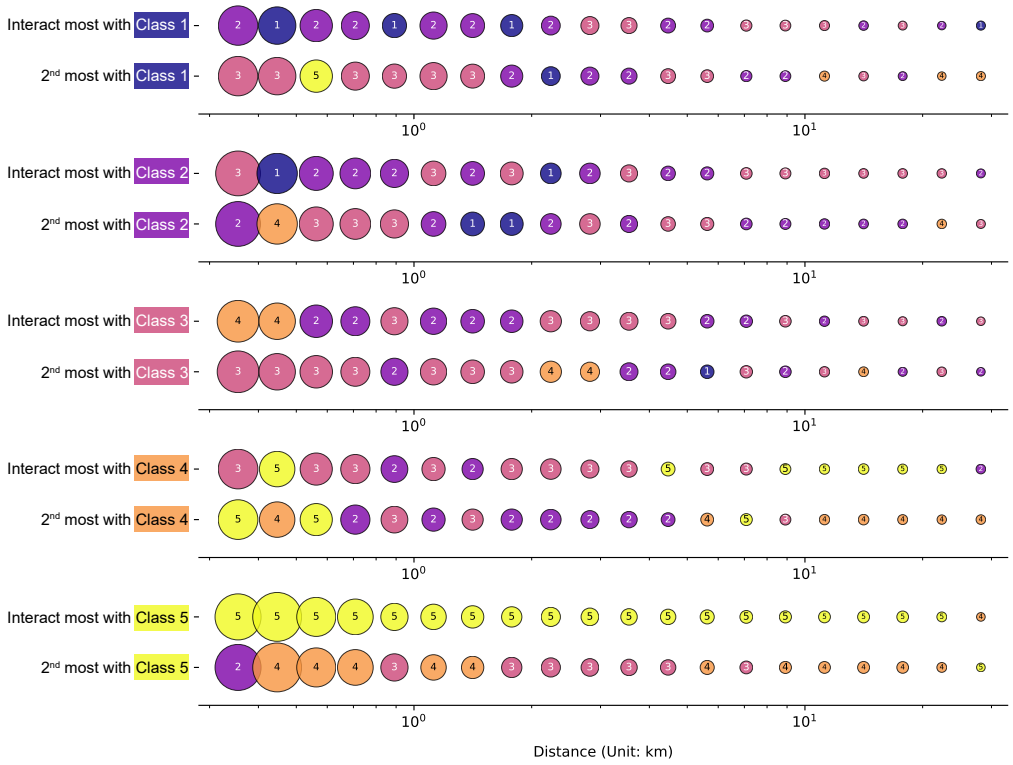

Figure 7: The classes that interact most and second most with each of the five classes at a given distance. Size of dot is proportional to the normalized communication frequency between the corresponding classes.

4.3 Formation of social ties

The communication intensities among the five classes have demonstrated the impact of homophily on the social network structure. Another interesting question is whether the formation of social ties would follow the homophily principle. To investigate this question, we perform another analysis by replacing communication intensity (I) with friendship probability (F), the measure that describes the formation of social ties. By deriving the friendship probability among the five classes at different distance values, we find that the effect of homophily still persists, but only when geographic distance is relatively short (Figure 8). Strikingly, when distance increases to a large value (e.g., over 10km), the classes that have the highest friendship probability with a given class tend to be C_4 or C_5. This observation is reaffirmed by Figure 9, which demonstrates the classes that have the highest and second highest friendship probability with each of the five classes at a given distance.

The results reveal an interesting aspect of social dynamics beyond the homophily principle. When people's residencies are far away from each other, the upper classes in the society are more likely to know or to be known by others. In other words, the formation of "long-range" ties tend to favor privileged people. Again, this might be related to the occupations of the upper classes as well as the overwhelming attentions they receive from other social groups. For example, it is difficult

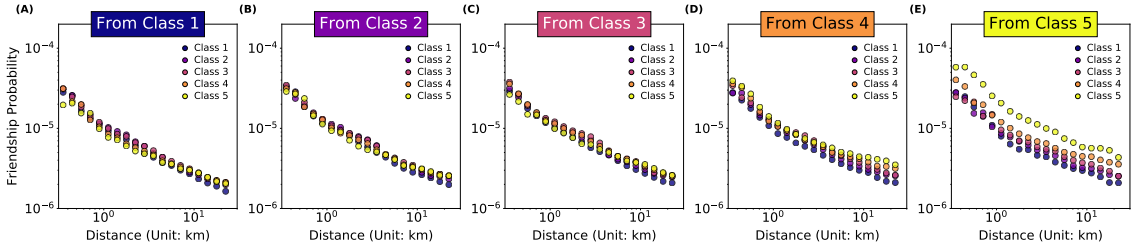

Figure 8: Friendship probability among the five classes.

for someone to connect with a regular person who is geographically distant. However, it is easier for a famous or rich person to be known by others regardless of where he or she resides in a city. This also connects to Stanley Milgram's small-world experiment and the theory of six degree of separation [33]. Namely, it is enough to have a "famous" friend reasonably close to be able to reach many other people in the world. Our results suggest that this "famous" and well-connected people is likely to be rich.

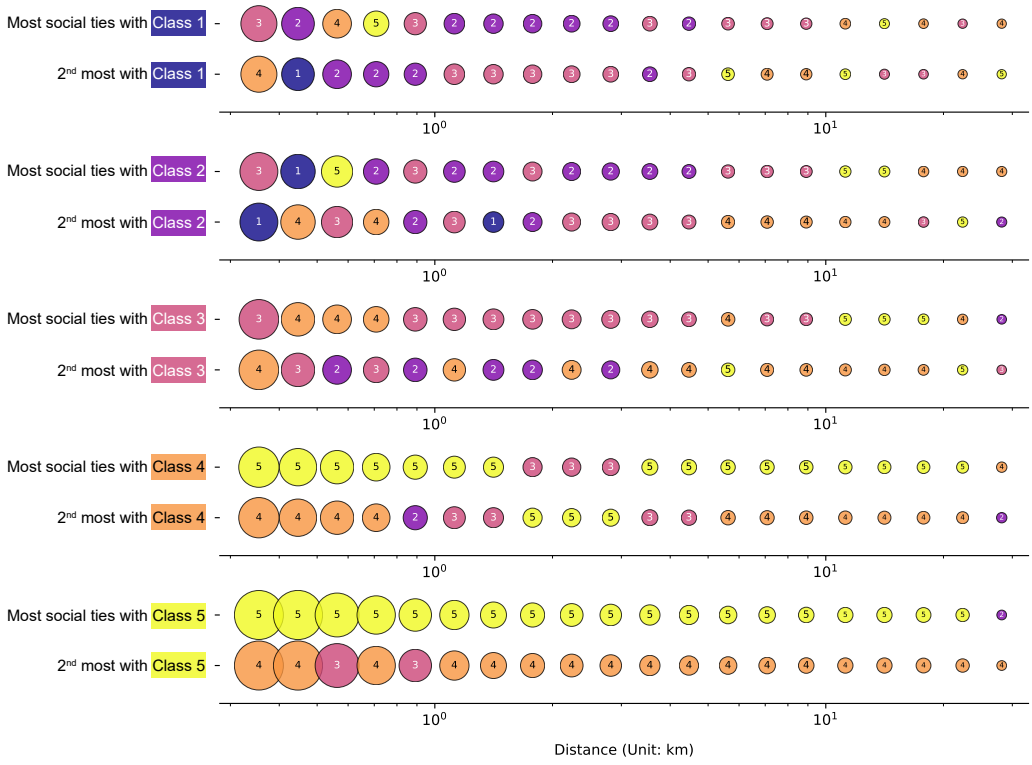

Figure 9: The classes that have the highest and second highest friendship probability with each of the five classes at a given distance. Size of dot is proportional to the friendship probability between the corresponding classes.

Note that although the formation of long-range ties tend to favor the upper classes, the intensities of cellphone communications at large distances are still higher among similar social classes (shown in Figure 6 and Figure 7). This indicates that for the wealthy populations, their connections with the middle and lower classes are more likely to be weak ties [34], while the majority of their interactions are still with similar others (e.g., other rich people).

4.4 Homophily distance

The results so far have demonstrated the joint impact of geographic distance and homophily on people's cellphone communications. Given this, it is possible that two neighborhoods with similar socioeconomic characteristics — but geographically distant — could have a higher communication intensity than two close neighborhoods with different characteristics. Figure 10A shows an example of this scenario. To explicitly quantify this effect, we adopt the concept of "homophily distance" [35] and compute this metric between two classes. Given two classes C_i and C_j, the homophily distance D_h is defined as the geographic distance such that the communication intensity between C_i and its peers (C_i) at D_h is roughly equal to the communication intensity between C_i and C_j at the minimal distance (i.e., when they are in close proximity). A large value of D_h suggests that C_i would overcome the friction of geographic distance to communicate with similar others. Here, we define the minimal distance as $1km$ to reflect the concept of "close proximity". To calculate D_h of C_i (from class) and C_j (to class), we fit a linear function of communication intensity with distance at the log-log scale (Figure 6). The point where the communication intensity at the minimal distance ($1km$) intersects with the fitted line gives the value of D_h.

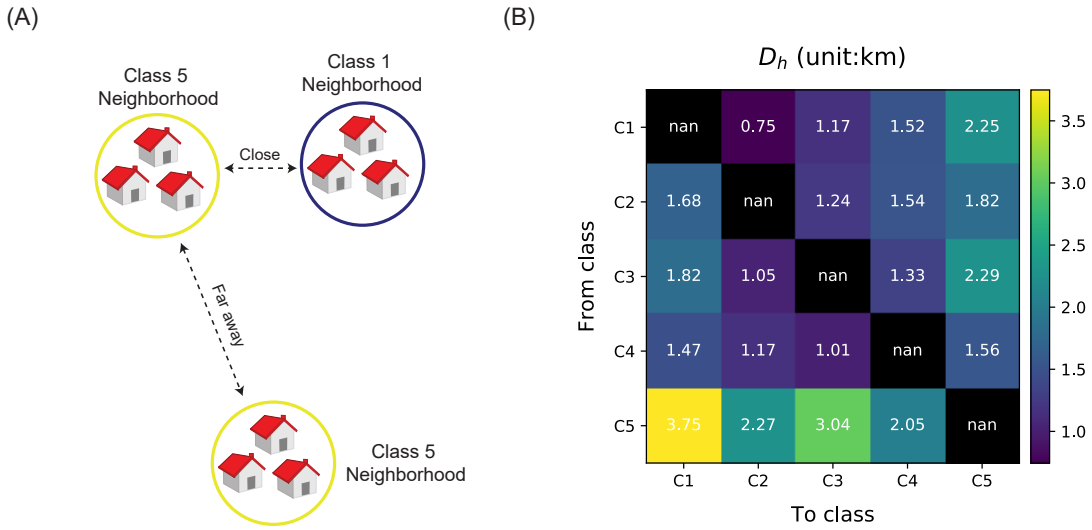

Figure 10: (A) Given the effect of distance decay and homophily, it is possible that two neighborhoods with similar socioeconomic characteristics but geographically distant could have a comparable of even higher communication intensity than two close neighborhoods with different characteristics; (B) The value of homophily distance for all combination of classes.

We compute D_h for all combinations of classes. The results are shown in Figure 10B. Almost all the values are above $1km$, indicating that places of the same class tend to maintain an adequate level of communication even when they are far apart. For instance, the homophily distance from C_1 to C_5 is $2.25km$, meaning that the communication intensity between C_1 and its peers that are $D_h = 2.25km$ away is comparable to the intensity between C_1 and C_5 at a distance of $1km$. This homophily effect is more pronounced when looking at the value from C_5 to C_1 ($D_h = 3.75km$). Note that the five classes are defined using the average housing price at the Voronoi cells and some of them could point to mixed-income neighborhoods. Thus, the homophily distances might be even higher if measured at the household or individual level. However, we are not able to measure this due to lack of individual level socioeconomic data. The results suggest that co-location or geographic proximity does not always ensure frequent communications. The socioeconomic similarity between places and the underlying populations play an equally important role in shaping the structure of the cellphone communication network.

5 Discussion and Conclusion

In this study, we establish a spatially embedded social network by coupling large-scale information on people's cellphone communications, residential locations, and socioeconomic characteristics. The findings suggest that cellphone interactions of people in Singapore are not only affected by geographic proximity, but also how social groups are distributed in the urban environment. We argue that the spatial organization of social groups in a city, which is often directed by urban planning policies (e.g., zoning strategies) [36], will have a direct impact on how people communicate in the virtual or online space.

By embedding phone users into urban space based on their estimated residence, we explore the geographic properties of the cellphone communication network. The results show that both communication intensity and friendship probability among places follow a power law decay with geographic distance. That means communications are more likely to occur among people who live close to each other. By further exploring the average housing price across places — which highly reflects the income level of residents in Singapore (Figure 3) — we find that the housing price values are highly clustered (Figure 5). Such an uneven distribution along with the distance decay effect tend to cause an imbalance of cellphone communications among social classes. The results indicate that physical separation of social groups in a city (e.g., residential segregation) will have a direct impact on shaping communication or social-network segregation.

By further examining cellphone connections among places with different housing price values, we find that it is not only geographic proximity, but also the principle of homophily that govern people's cellphone communications. In particular, places tend to have a higher communication intensity with ones that have similar housing price values when distance is controlled (Figure 6 and Figure 7). This homophily effect is reaffirmed by looking at the homophily distances among social classes (Figure 10). For instance, the communication intensity between the upper class (C_5) and its peers at a distance of $3.75km$ is comparable to the level between C_5 and the lowest class (C_1) that are immediately nearby. That means social classes would overcome the friction of distance or resist the "convenience of co-location" in order to connect with similar others. Note that we also observe this homophily effect when examining friendship probability, the indicator that describes the formation of social ties (Figure 8). Interestingly, though, the homophily effect is observed primarily at short distance values, or across all distance values for the upper class C_5. When distance is large, the middle-upper classes (C_4 and C_5) are always the ones that have the most connections with other socioeconomic tiers (Figure 9). The result reveals a "celebrity effect" on the formation of social ties.

The findings on the homophily effect suggest that neighborhoods with similar characteristics, even when they are far apart, could contribute to the emergence of social-network segregation. Therefore, traditional placed-based measures, such as creating mixed income neighborhoods, might not be enough per se to accomplish a high degree of social mixing in the online space. Novel online activities that can actively bridge different social classes can possibly promote social cohesion of cities given the prevalence of social media platforms and the recent public health emergency (COVID-19) [37] that further restricts physical human interactions.

Our study suggests that geographic proximity and homophily are two notable forces that jointly shape the structures of social networks. Although the current study focuses on Singapore, we believe these two forces have a far-reaching impact on human interactions in cities. Since classic spatial interaction models (e.g., gravity models) have a specific focus on modelling the distance effect, future work that aims to model or explain spatial-social networks could incorporate socioeconomic characteristics as a generic factor.

We want to point out a few limitations of this research. First, social ties in this study are defined as cellphone user pairs with at least one reciprocal contact during the study period. Although this measure, as suggested by a previous study in Portugal [4], was indicative of face-to-face interactions and therefore some sort of social relationships, this evidence does not immediately generalize to the city of Singapore which is the subject of this study. Hence, we acknowledge that the definition based on reciprocal phone contact provides a coarse indication of social ties in Singapore. Second, social interactions in this study are measured using phone calls and text messages. Although they

accounted for a notable fraction of human communications in 2011 (when the dataset was collected), people in Singapore were adopting social media (e.g., Facebook, Twitter) at that time as new channels of social interactions. In the future, it would be meaningful to examine the robustness of the findings by incorporating social media usage [38, 39, 40] into the analysis (e.g., Twitter mentions and follower-followee relationships). For instance, a recent work based on geo-located Tweets [41] has observed a certain degree of mobility homophily in the city of Stockholm, hinting to a possible robustness of our findings to social media datasets. Third, this study uses housing price as the variable to account for homophily. There exist other factors, such as ethnicity, that would affect people's communication patterns. Since our analysis is conducted at the cell tower level, we are not able to obtain information of ethnicity groups at such a fine spatial resolution. Fortunately, the Singapore government introduced the Ethnic Integration Policy (EIP) in 1989 for HDB (Housing Development Board) estates. The EIP sets limits on the total percentage of a block or neighborhood that can be occupied by a certainty ethnicity. Given that HDB estates host more than 80% of residential population in Singapore, the EIP policy tends to neutralize the impact of ethnic segregation on our results. But we do think it is important to consider these socioeconomic variables in future works (e.g., ethnicity and education background) to depict a more holistic picture of homophily in spatial-social networks.

This study demonstrates the importance of conceptualizing and modeling social networks in geographic space [23]. It also calls for more efforts on coupling physical and virtual (online) spaces for studying human dynamics [42]. Although exploratory in nature, the current study points to a few venues for future research. For instance, it would be meaningful to incorporate socioeconomic similarity of places into spatial interaction models [43, 44, 45] to better predict structures of spatial-social networks. It is also possible to detect changes of socioeconomic environments (e.g., gentrification) in cities by monitoring the interactions among neighborhoods and how they evolve through time and space [46].

Notes

[1]Planning areas, also known as DGPs, are the primary census divisions of Singapore created by the Urban Redevelopment Authority (URA). There are a total of 55 planning areas in Singapore at the time when the CDR data were collected (`https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore`).

[2]HDB, which is short for Housing Development Board, is a type of residential housing property that is publicly governed and developed in Singapore. The HDB flats were built primarily to provide affordable housing

References

[1] L. Strate, "The varieties of cyberspace: Problems in definition and delimitation," *Western Journal of Communication (includes Communication Reports)*, vol. 63, no. 3, pp. 382–412, 1999.

[2] J. Larsen, K. W. Axhausen, and J. Urry, "Geographies of social networks: meetings, travel and communications," *Mobilities*, vol. 1, no. 2, pp. 261–283, 2006.

[3] J. A. Carrasco, E. J. Miller, and B. Wellman, "How far and with whom do people socialize? empirical evidence about distance between social network members," *Transportation Research Record*, vol. 2076, no. 1, pp. 114–122, 2008.

[4] F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti, "Interplay between telecommunications and face-to-face interactions: A study using mobile phone data," *PloS one*, vol. 6, no. 7, p. e20814, 2011.

[5] F. Cairncross, "The death of distance: How the communications revolution will change our lives," 1997.

[6] S. Graham, "The end of geography or the explosion of place? conceptualizing space, place and information technology," *Progress in human geography*, vol. 22, no. 2, pp. 165–185, 1998.

[7] J. Goldenberg and M. Levy, "Distance is not dead: Social interaction and geographical distance in the internet era," *arXiv preprint arXiv:0906.3202*, 2009.

[8] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th international conference on World wide web*, pp. 61–70, ACM, 2010.

[9] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks," *Proceedings of the National Academy of Sciences*, vol. 102, no. 33, pp. 11623–11628, 2005.

[10] S. Gao, Y. Liu, Y. Wang, and X. Ma, "Discovering spatial interaction communities from mobile phone data," *Transactions in GIS*, vol. 17, no. 3, pp. 463–481, 2013.

[11] R. Lambiotte, V. D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, "Geographical dispersal of mobile communication networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, pp. 5317–5325, 2008.

[12] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, "Urban gravity: a model for inter-city telecommunication flows," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 07, p. L07003, 2009.

[13] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[14] D. S. Massey and N. A. Denton, "Trends in the residential segregation of blacks, hispanics, and asians: 1970-1980," *American sociological review*, pp. 802–825, 1987.

[15] Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, and M. Karsai, "Socioeconomic correlations and stratification in social-communication networks," *Journal of The Royal Society Interface*, vol. 13, no. 125, p. 20160598, 2016.

[16] G. Le Roux, J. Vallée, and H. Commenges, "Social segregation around the clock in the paris region (france)," *Journal of Transport Geography*, vol. 59, pp. 134–145, 2017.

[17] S. Musterd, S. Marcińczak, M. Van Ham, and T. Tammaru, "Socioeconomic segregation in european capital cities. increasing separation between poor and rich," *Urban Geography*, vol. 38, no. 7, pp. 1062–1083, 2017.

[18] Q. Wang, N. E. Phillips, M. L. Small, and R. J. Sampson, "Urban mobility and neighborhood isolation in america's 50 largest cities," *Proceedings of the National Academy of Sciences*, vol. 115, no. 30, pp. 7735–7740, 2018.

[19] D. S. Massey and N. A. Denton, "The dimensions of residential segregation," *Social forces*, vol. 67, no. 2, pp. 281–315, 1988.

[20] C. S. Fischer, *To dwell among friends: Personal networks in town and city.* University of chicago Press, 1982.

[21] J. Q. Stewart, "An inverse distance variation for certain social influences," *Science*, vol. 93, no. 2404, pp. 89–90, 1941.

[22] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 119–128, 2010.

[23] C. Andris, "Integrating social network data into gisystems," *International Journal of Geographical Information Science*, vol. 30, no. 10, pp. 2009–2031, 2016.

[24] F. P. Stutz, "Distance and network effects on urban social travel fields," *Economic Geography*, vol. 49, no. 2, pp. 134–144, 1973.

[25] E. Cho, S. A. Myers, and J. Leskovec, "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*, pp. 1082–1090, 2011.

[26] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1100–1108, 2011.

[27] Y. Xu, A. Belyi, I. Bojic, and C. Ratti, "How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data," *Transactions in GIS*, vol. 21, no. 3, pp. 468–487, 2017.

[28] S. Farber, M. O'Kelly, H. J. Miller, and T. Neutens, "Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure," *Journal of Transport Geography*, vol. 49, pp. 26–38, 2015.

[29] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *Fifth international AAAI conference on weblogs and social media*, 2011.

[30] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of urban technology*, vol. 17, no. 1, pp. 3–27, 2010.

[31] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *International Conference on Pervasive Computing*, pp. 133–151, Springer, 2011.

[32] Y. Xu, S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li, "Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach," *Transportation*, vol. 42, no. 4, pp. 625–646, 2015.

[33] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[34] M. S. Granovetter, "The strength of weak ties," in *Social networks*, pp. 347–367, Elsevier, 1977.

[35] Y. Xu, A. Belyi, P. Santi, and C. Ratti, "Quantifying segregation in an integrated urban physical-social space," *Journal of the Royal Society Interface*, vol. 16, no. 160, p. 20190536, 2019.

[36] J. Jacobs, *The death and life of great American cities*. Vintage, 2016.

[37] N. Oliver, B. Lepri, H. Sterly, R. Lambiotte, S. Deletaille, M. De Nadai, E. Letouzé, A. A. Salah, R. Benjamins, C. Cattuto, *et al.*, "Mobile phone data for informing public health actions across the covid-19 pandemic life cycle," 2020.

[38] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Link mining: models, algorithms, and applications*, pp. 337–357, Springer, 2010.

[39] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22436–22441, 2010.

[40] W. Xi, C. A. Calder, and C. R. Browning, "Beyond activity space: Detecting communities in ecological networks," *Annals of the American Association of Geographers*, vol. 110, no. 6, pp. 1787–1806, 2020.

[41] C. Heine, C. Marquez, P. Santi, M. Sundberg, M. Nordfors, and C. Ratti, "Analysis of mobility homophily in stockholm based on social network data," *PloS one*, vol. 16, no. 3, p. e0247996, 2021.

[42] S.-L. Shaw and D. Sui, "Understanding the new human dynamics in smart spaces and places: Toward a splatial framework," *Annals of the American Association of Geographers*, vol. 110, no. 2, pp. 339–348, 2020.

[43] G. K. Zipf, "The p1 p2/d hypothesis: on the intercity movement of persons," *American sociological review*, vol. 11, no. 6, pp. 677–686, 1946.

[44] W.-S. Jung, F. Wang, and H. E. Stanley, "Gravity model in the korean highway," *EPL (Europhysics Letters)*, vol. 81, no. 4, p. 48005, 2008.

[45] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.

[46] D. Zhu, F. Zhang, S. Wang, Y. Wang, X. Cheng, Z. Huang, and Y. Liu, "Understanding place characteristics in geographic contexts through graph convolutional neural networks," *Annals of the American Association of Geographers*, vol. 110, no. 2, pp. 408–420, 2020.