Quantifying segregation in an integrated urban physical-social space

Yang Xu[1,*,+], Alexander Belyi[2,3,+], Paolo Santi[4,5], and Carlo Ratti[4]

[1]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong
[2]Singapore-MIT Alliance for Research and Technology, 1 Create Way, Singapore
[3]Faculty of Applied Mathematics and Computer Science, Belarusian State University, Minsk, Belarus
[4]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Istituto di Informatica e Telematica del CNR, Pisa, Italy
[*]yang.ls.xu@polyu.edu.hk
[+]these authors contributed equally to this work

ABSTRACT

Our knowledge of how cities bring together different social classes is still limited. Much effort has been devoted to investigating residential segregation, mostly over well-defined social groups (e.g., race). Little is known of how mobility and human communications affect urban social integration. The dynamics of spatial and social-network segregation and individual variations along these two dimensions are largely untapped. In this article, we put forward a computational framework based on coupling large-scale information on human mobility, social network connections, and people's socioeconomic status (SES), to provide a breakthrough in our understanding of the dynamics of spatiotemporal and social-network segregation in cities. Building on top of a social similarity measure, the framework can be used to depict segregation dynamics down to the individual level, and meanwhile, provide aggregate measurements at the scale of places and cities, and their evolution over time. By applying the methodology in Singapore using large-scale mobile phone and socioeconomic datasets, we find a relatively higher level of segregation among relatively wealthier classes, a finding that holds for both social and physical space. We also highlight the interplay between the effect of distance decay and homophily as forces that determine communication intensity, defining a notion of characteristic "homophily distance" that can be used to measure social segregation across cities. The time-resolved analysis reveals the changing landscape of urban segregation and time-varying roles of places. Segregations in physical and social space are weakly correlated at the individual level, but highly correlated when grouped across at least hundreds of individuals. The methodology and analysis presented in this paper enable a deeper understanding of the dynamics of human segregation in social and physical space, which can assist social scientists, planners, and city authorities in the design of more integrated cities.

Introduction

The past half-century has witnessed an upsurge of population growth in the world's urban areas. Through this rapid urbanization, cities around the world — big or small — have accumulated unprecedented amounts of wealth and people. Such agglomerations are beneficial to the well-being of societies but raise issues such as income inequality[1,2] and social stratification[3]. The homophily mechanism in the formation of social relations[4], along with other political and economic forces, seem to fabricate an uneven distribution of social groups in cities[5]. These socioeconomic imbalances have led to varying degrees of urban segregation[6], causing issues of crime[7], inequalities in education attainment[8–10], disparities in health[11], and others. Besides these social problems, a more worrying fact is that segregations in urban areas have grown rapidly in the past few decades[12,13]. The widening gaps across socioeconomic classes call for more effective urban policies and mitigation strategies, among which many can benefit from improved abilities to follow segregation dynamics in time and space.

Unfortunately, our knowledge of how effectively cities bring together different social classes is quite limited. To date, the separation of social groups in cities has been observed mainly according to their place of residence[14], more rarely according to the places they visit[15]. The conceptualization and quantification of social segregation have largely been practiced in static spatial terms[16]. From the data perspective, population census — which dominates the data sources for classic segregation studies[17] — provides a static view of the uneven distribution of social groups in physical space. Subject to small sample sizes, high acquisition costs, and the inability to capture human movement patterns, such data is unable to deliver a timely and dynamic view of social mixing in cities. There is a need for segregation studies to expand from place-based measures to people-based measures, and to put more efforts to advancing temporally integrated analysis[16,18]. To fill these gaps, scholars

started to incorporate travel behaviour and activities of individuals into their analytical frameworks through the usage of travel surveys and spatial trajectories[19–27]. Leveraging the notion and measurements of human activity space, these studies substantially improved the understanding of segregation beyond residential spaces, and meanwhile, enabling individual-oriented and time-space views of social segregation. Despite the improvements made by these studies, their findings were all obtained through observations of human mobility in physical space. The impact of people's day-to-day communications, which represent another important dimension of their social behaviours, remain largely untapped. We are still in need of cost-effective ways to better understand — at the population scale — the impact of human mobility on physical segregation, the interactions among classes in the social space, and the interplay between them.

In recent years, with increasing availability of data resources and computational tools for social science research, the ways we understand socioeconomic systems have been radically transformed[28]. This transformation is also reflected in augmented capabilities to measure segregation in cities[29]. For instance, by coupling information of social network and mobility in mobile phone data with income data from bank records, researchers have offered a new way to describe segreation, revealing the "rich club" effect in social-communication networks[30]. A more recent work (https://inequality.media.mit.edu) aims to better understand places in cities by capturing the income inequality of visitors, enabling a dynamic and micro-level view of social segregation. These studies have contributed to the observation of segregation from a social-spatial perspective. However, a few important questions remain unaddressed. How or to what extent are individuals exposed to similar others in social network or physical urban space? Are people more socially isolated in one dimension more isolated in the other? Is there a unified framework to quantify individual variations along these two dimensions, and meanwhile, more aggregate forms of social segregation in a city?

We hereby introduce an analytical framework — through the usage of large-scale mobile phone and urban socioeconomic datasets — to address the above questions. Mobile phones, powered by various information and location-aware technologies, have enabled a 'digital census' that documents the whereabouts of people and their communication patterns. These digital breadcrumbs have been used to study large-scale population dynamics, revealing the intrinsic properties of human movements[31,32], structures of social networks[33], and the interplay between the two[34,35]. However, their potential in tackling segregation issues remains underexplored, largely because of the absence of socioeconomic characteristics in the datasets. To close this gap, we advocate a coupled usage of multiple urban datasets (e.g., mobile call detail records, high-resolution housing price, and income data), from which the socioeconomic status and interactions of large populations can be inferred at the same time to depict segregation and its evolution — in both physical and social space. We conduct a case study in Singapore, an island city-state in southeast Asia, to demonstrate the feasibility the proposed framework.

We start the problem formulation by presenting a collection X of individuals in a city, whose location footprints and communication patterns were documented by call detail records (CDRs). CDRs, usually collected by cellular operators for billing purposes, capture the whereabouts of individuals during their phone usage activities (e.g., call/text message). Given an individual $\alpha \in X$, her location trace can be represented by a sequence of tuples $(l_i, t_i, e_i, c_i, d_i)$, where l_i represents the cellphone tower location of the i^{th} record, t_i – the time stamp, e_i – the event type (i.e., outbound/inbound call, outbound/inbound message), c_i – the unique identifier of the phone user that α communicated with, d_i – the duration of the call ($d_i = 0$ when e_i being text message). Such information allows us to assess — at the city scale — not only interactions of large populations in physical space, but also their communications in social space.

To measure segregation, however, we need to distinguish individuals by their socioeconomic status (SES). Such information is usually not available in the CDR data. To overcome this issue while preserving individuals' privacy, we link phone users' residential locations (estimated from CDRs, see Methods section) with high-resolution socioeconomic datasets — in this context, the sale price of residential properties — to depict the populations' socioeconomic characteristics. The underlying assumption is that phone users who live in more affluent areas tend to have higher SES in general. We choose to use residential property price for two reasons. First, it was not possible to acquire income or other direct SES indicators from the cellphone dataset. Second, the sale price of individual housing property provides a fine-grained view of the socioeconomic configuration of the city, allowing us to approximate SES of the population down to individual level. (In Supplementary, we perform a correlation analysis between housing price and income data obtained from household interview travel survey. The result indicates that housing price can be used as a reasonable indicator of SES). Through this linkage, we approximate each phone user's SES using a discrete value inferred from the data fusion process (see Methods), from which individuals can be ranked along a city's socioeconomic spectrum (e.g, from poor to rich). Note that when estimating SES, the property price is used as the sole indicator regardless of tenure and housing type. Moreover, there could be multiple buildings (and the corresponding property price) that potentially match with an individual's estimated residence. To partially account for the data uncertainty, an SES assignment model is proposed and a robustness check based on multiple runs of the assignment is performed to evaluate its impact on the segregation measurements (Methods).

A new metric is then proposed to quantify the segregation dynamics. The metric takes the SES of phone users and their interactions as input. Such interactions — depending on whether they occur in physical or social space — can be quantified

through phone users' movements or cellphone communications. The segregation measure outputs an index, for each individual, that describes to what extent he or she is 'exposed' to similar others.

A Similarity Measure Based on Social Ranks

The first step in measuring social segregation is the definition of a robust notion of social similarity between phone users. Here, we introduce a similarity measure based on social ranks that will be used later to quantify segregation in physical and social space (Figure 1). Given a collection of phone users, we sort them based on the inferred SES values from low to high. A higher SES value indicates that the individual tends to belong to a higher socioeconomic class. This results in a finite sequence $(r_n)_{n=1}^{N} = (1,2,...,N)$, where r_n denotes the rank of the corresponding phone user, and N denotes the population size (Figure 1A). From now on, when we write about individual x we assume an individual with rank r_x. First, a social distance metric is defined to describe the distance from individual i to individual j. Given two individuals i and j, we define a set $A = \{x \mid |r_x - r_i| < |r_i - r_j|\}$ of individuals x that are closer to i than j is to i, then the social distance from i to j is defined as follows:

$$d_{i\to j} = \begin{cases} \frac{|A|+0.5}{N-1}, & \text{if there exists another } k(k \neq j) \text{ such that } |r_k - r_i| = |r_i - r_j| \\ \frac{|A|}{N-1}, & \text{otherwise} \end{cases} \quad (1)$$

where $|A|$ denotes the cardinality of A.

The social distance can be interpreted as the 'the total number of individuals that are closer to i than j is to i', normalized by the population size. Note that $d_{i\to j}$ might not be equal to $d_{j\to i}$, implying that the defined notion cannot be intended as a distance metric in mathematical terms. The rationale behind our definition of social distance is that people belonging to different social classes could have different views of how "far" other people are from them. For example, the poorest individual would consider a middle-class person as moderately far away. The middle-class person, however, would consider the poorest man — and probably the richest man as well — as the farthest person to him. We use the term 'distance' in an intuitive sense although the measure $d_{i\to j}$ does not satisfy symmetry.

Then, the social similarity between i and j can simply be calculated as $s_{i\to j} = 1 - d_{i\to j}$. The value of $s_{i\to j}$ ranges from zero to one, and a larger value indicates a higher social similarity. The complete equations for calculating the social distance/similarity measures are provided in the Methods section.

One important feature of the social similarity measure is that it allows us to easily compare the segregation level of an individual with the *null model*, which assumes that people interact with others randomly in physical or social space. Given any individual x, if we compute the weighted sum of social similarity between him/her to all individuals, with the weights being equal i.e., $\frac{1}{N}$, we can prove that the weighted similarity $\frac{1}{N} \cdot \sum_{i=1}^{N} s_{x\to j}$ is always 0.5, and this value is independent of the rank r_x (Figure 1E, see Supplementary for a formal proof). A value higher than 0.5 suggests that the individual tends to be relatively more segregated in physical or social space (i.e., interacts or communicates more with similar others), while a value lower than 0.5 suggests that the individual is exposed more to dissimilar others. To some extent, the social integration metric defined herein can be considered as an extension of assortativity metrics commonly used in network analysis, where emphasis is given to social attributes of the nodes in the network rather than to the mere structure of their connections.

With this social similarity measure, we can start answering the above-mentioned question, i.e., to what extent is an individual 'exposed' to similar others? This can be quantified as the social similarity of the individual to others, weighted by their interaction strength in either social (Fig. 1B) or physical space (Fig. 1C). In Fig. 1D, we illustrate the social similarity function of individuals with some selected ranks using $N = 1000$ as an example.

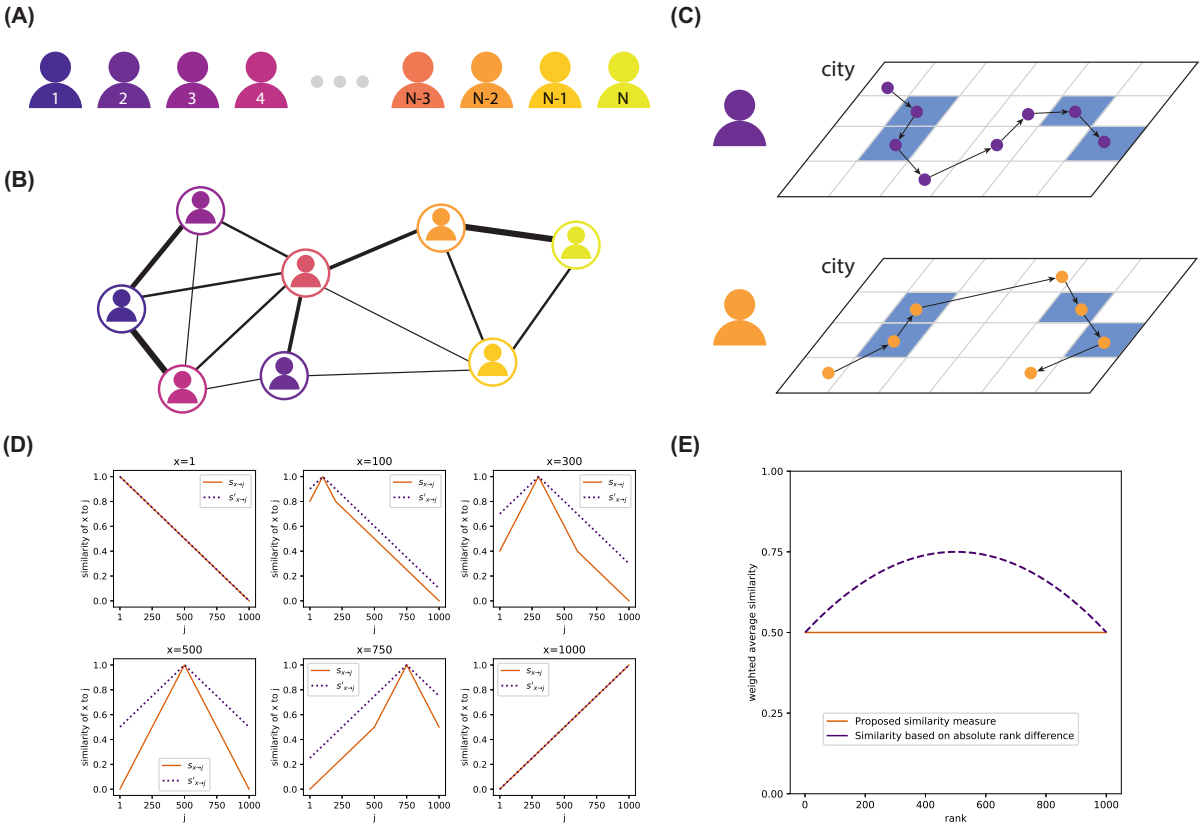

Figure 1. Framework of the segregation metric: **(A)** phone users are ranked by their socioeconomic status (SES) from low to high. Here the number denotes the rank of the corresponding phone user and N refers to the population size; **(B)** by observing cellphone communications among these individuals, a social network structure can be reconstructed to quantify segregation in the social space. In a segregated social network, interactions are more likely to be observed between individuals with similar SES; **(C)** by observing individual mobility in urban space, we can measure, for any pair of individuals, their *co-location* probability (i.e., how likely they tend to encounter each other in physical space). Similar to the tie strength between two individuals in the social network, such co-location probabilities can be used as the weights to quantify the presence or absence of segregation in physical space. In a segregated city, individuals tend to share space more often with similar others; **(D)** the orange lines illustrate the social similarity function $s_{x\to j}$ proposed in this study — using selected individuals with ranking $x = 1, 100, 300, 500, 750, 1000$ and $N = 1000$ as an example. The dotted purple lines shows the similarity between individuals if measured using absolute rank difference ($s'_{x\to j} = 1 - \frac{|r_x - r_j|}{N-1}$); **(E)** under the assumption of a *null model*, where individuals interact randomly with others in social or physical space, the segregation level of any individual can be quantified as the weighted sum of social similarity between him/her to all individuals, i.e., $\frac{1}{N}\sum_{j=1}^{N} s_{x\to j} = 0.5$. This value, which we name as 'baseline', is independent of the rank x (orange line). If the social similarity function is computed based on absolute rank difference, it will produce different 'baselines' for individuals with different ranks (dashed purple line, see Supplementary and Fig. S.7 for details on how our proposed social similarity measure enables better interpretations and comparisons of segregation measurements compared to ones derived from absolute rank difference).

Segregation in Social Space

Building on the defined social similarity function, we propose a *communication segregation index* (*CSI*) to quantify the segregation level of individuals in the social space. The *CSI* of an individual x is defined as the weighted sum of social similarity between x and his or her contacts $(y_1, y_2, ..., y_m)$, with weights being their communication strength $(f_1, f_2, ..., f_m)$:

$$CSI_x = \frac{\sum_{j=1}^{m} f_j \cdot s_{x \to y_j}}{\sum_{j=1}^{m} f_j} \tag{2}$$

Here f_j is computed as the total number of phone calls and text messages exchanged between x and y_j during the study period. While the duration of phone calls could also indicate social information, we used number of calls to account for text messages that do not have duration, but constitute significant part of communication.

The value of *CSI* ranges from 0 to 1. A *CSI* close to 1 indicates that the individual mainly communicates with similar others, while a value close to 0 means that the person mainly interacts with those who have quite different socioeconomic characteristics. If an individual communicates equally with any other person in the social network, his or her *CSI* would be equal to 0.5. Under the assumption of the ***null model***, where individuals connect randomly with others in the network, all the individuals will have an expected *CSI* value of 0.5. Here, we name this value a ***baseline*** of segregation in social space.

Through cellphone users' communications observed in the CDR data, we extract a city-scale social network, from which we compute the *CSI* of individuals. Out of about 1.8 million people with assigned SES (and hence having been ranked) about 1.4 million had mutual connections with someone else and got *CSI* value. As further explained in Methods, we calculate *CSI* values for 100 random assignments of house prices (hence, SES) to users. In each run we obtained very similar distributions of *CSI* values (see error bars shown in Fig. 2A), so here we present results for one of these assignments. This gives us a close to normal distribution with a mean and standard deviation of 0.546 and 0.200 (Fig. 2A). The two tails of the distribution suggest that there exist individuals who are quite segregated (e.g., $CSI > 0.8$) or socially 'integrated' (e.g., $CSI < 0.2$) in the network. Overall, cellphone communications in the network are slightly biased towards dyads that are similar to each other. Compared to a random network with a baseline *CSI* of 0.5, the social network produces a mean *CSI* of 0.546, meaning that people are 4.6 percent closer (i.e., $0.546 - 0.5 = 0.046$), than expected, to their own socioeconomic class — an evidence of moderate segregation in the social space. Note that when calculating individual *CSI*, we have filtered communications between users with the same home cell. This ensures that the segregation indices are less affected by communications between users who live together (e.g., family members and relatives). In Supplementary, we provide results when including individuals living in the same cell. The mean *CSI* increases from 0.546 to 0.576 and the standard deviation remains 0.200.

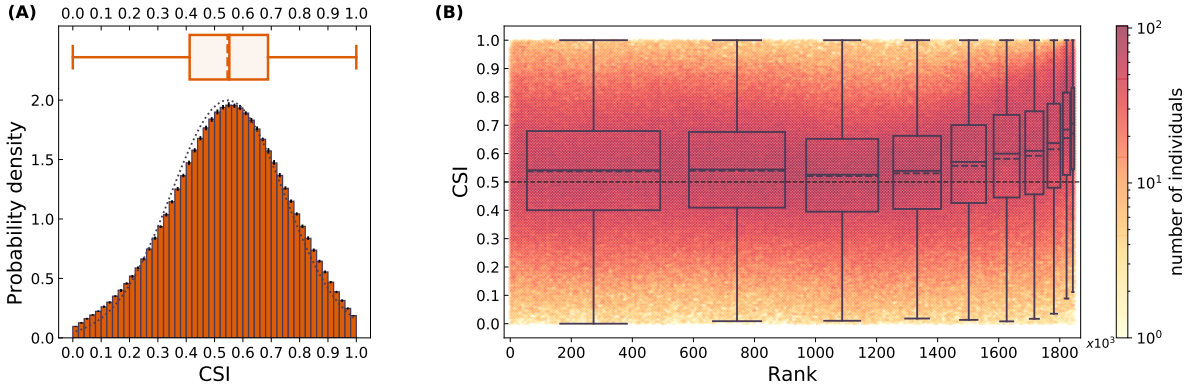

Figure 2. Distribution of *CSI* values: **(A)** A histogram and a box plot of *CSI* values obtained during 100 random assignments, error bars on top of each bar indicate variation (min and max values) observed over different assignments, dotted line shows fitted normal distribution; **(B)** Relationship between *CSI* and social ranks through a density plot for one of the random assignments. Darker colors indicate areas with more individuals, 10 box plots show distribution of *CSI* values within 10 classes with equal cumulative housing price, all means and medians are above baseline of 0.5. All box plots show median (center line), mean (dashed line), upper and lower quartiles (box limits), 1.5 interquartile range or min/max values (whiskers).

By comparing *CSI* of individuals with different social ranks, we find that communication segregation is highly dependent on SES (Fig. 2B). The aggregate segregation index remains relatively consistent across the poor and middle classes, but it consistently tends to increase among richer classes. Although all the socioeconomic tiers are observed to be more segregated than expected, the 'rich club' effect suggests that the top socioeconomic classes are the major contributors to the communication

segregation. This observation remains consistent over 100 assignments. For each of 10 groups maximum difference in values of mean, median, first and third quartiles between 100 assignments does not exceed 0.0015 (see Fig. S.5 for more details).

From many factors that could potentially affect communication strength in social networks we were interested in interplay between two major ones that were well studied in previous literature, namely: 'distance decay'[36–39] and 'homophily'[4]. The first suggests that interactions in social space are more likely to occur among people who are closer in physical space, while the second indicates that people are more likely to interact with similar others. Our analytical framework allows for the first time analyzing the interplay between these two forces affecting communication strength. To this end, we associate individual phone users to their home location, from which we measure the communication strength among people at different home distances. The 'distance decay' effect is very evident across the entire population (Fig. 3A). Deeper insights can be achieved by disaggregating the data for residents of neighborhoods with different housing prices. We observe that the 'distance decay' effect carries over all social classes. Furthermore, we do find evidence of 'homophily' when we look at the communication strength between individuals at a certain, given distance: in this case, individuals tend to prefer communication with individuals belonging to the same social class — (Fig. 3B to Fig. 3F).

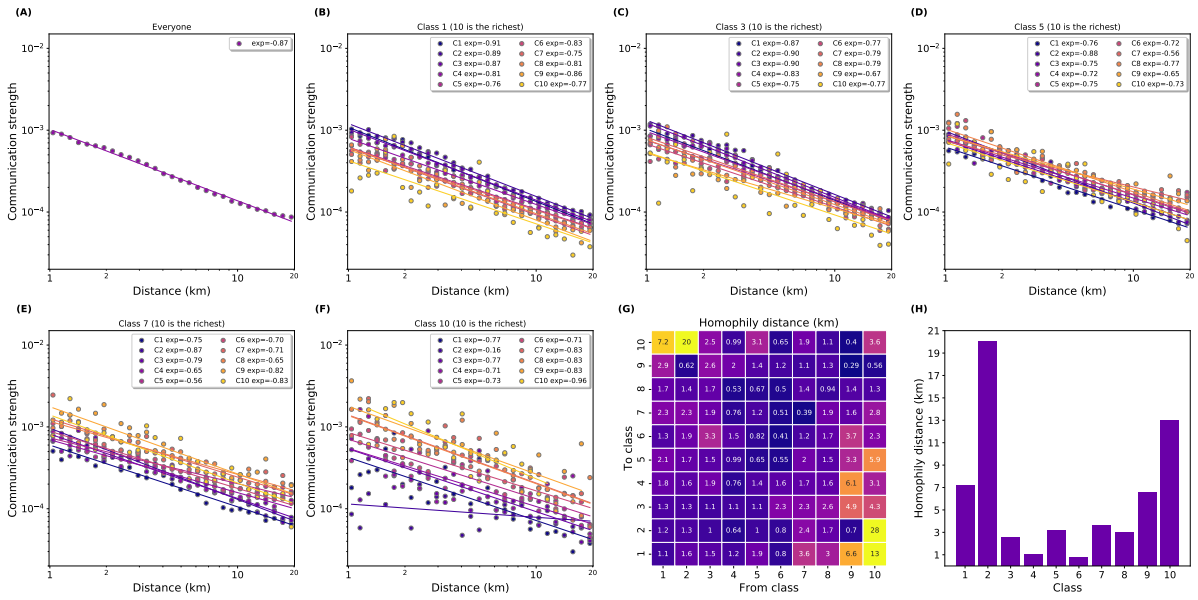

Figure 3. 'Distance decay' and 'homophily' effects on communication. **(A)** By assigning phone users to their home grid cell, we measure the communication strength among people separated at different home distances. A 'distance decay' effect is observed. Here, the *normalized communication strength* at a given distance is calculated as the total number of phone calls and messages exchanged between phone users who live at the corresponding locations, normalized by the total possible friend pairs (i.e., social ties) at this distance; **(B – F)** We further compute the average housing price of each grid cell, and categorize them into 10 classes based on equal cumulative housing price (i.e., each class would have the same 'buying power', which is measured as the sum of housing price value of all the corresponding phone users). By computing the normalized communication strength among selected classes, we find that locations with similar socioeconomic characteristics tend to have higher communication strength when distance being equal. Exponents indicate the slope of fitted lines that serve as a guide for the eye; **(G)** 'Homophily distance' between all class pairs. The off-diagonal values are generally larger, which indicates a stronger homophily effect for the generally lower and upper social classes; **(H)** For each class C_i, we plot its 'homophily distance' to its farthest social class. The values are generally larger at the two sides, which further suggest a stronger homophily effect for the poor and rich.

Given a decreasing trend of communication strength with distance in both physical and socio-economic space, an intriguing question is whether there exists a point at which the effect of 'distance decay' and 'homophily' would balance and, if so, at which point. In other words, we aim at characterizing a distance, which we name 'homophily distance', at which the effect of co-location (i.e., two persons who live in the same area but from very different social classes) is balanced by the effect of homophily (i.e., people interacting with similar others even if they live further). Formally, this characteristic 'homophily distance' is defined as follows. Given any two social classes C_i (from class) and C_j (to class), the homophily distance d is defined as the distance value d such that communication strength between C_i and C_j at the minimal distance, $f(C_i, C_j, 1)$, is

equal to the communication strength between C_i and its peers located at distance d, $f(C_i, C_i, d)$. The value of d — especially when the two classes C_i and C_j are far from each other — would express 'homophily' in spatial terms, with relatively larger values of d indicating a stronger tendency for C_i to connect with their own socioeconomic classes. Here again we note, that 'homophily distance' is not a proper mathematical distance, since it is not symmetrical (as could be seen in Fig. 3G). To calculate d we fit logarithm of communication strength values with a straight line and find the point where it intersects with horizontal line $y = f(C_i, C_j, 1)$.

By plotting the homophily distances between all classes (Fig. 3G), we find relatively higher values of d at the two off-diagonal corners. This suggests that when C_i refers to the generally lower or upper socioeconomic classes, it takes a long distance to let the 'homophily' effect to be outweighed by the co-location effect. To further illustrate this, for each class C_i, we pick its farthest social class (e.g., C_{10} is the farthest class to C_1) and plot their 'homophily distance' (Fig. 3H). Despite some fluctuations (e.g., values of C_1 and C_2), the values are generally higher at the two sides of the socioeconomic spectrum, meaning that the 'homophily' effect — the tendency for people to connect with similar others — are relatively stronger for the poor and rich.

Segregation in Physical Space

We propose a *physical segregation index* (*PSI*) to measure individual segregation in urban space. Different from studies of residential segregation, which only consider where people live, our metric provides a more dynamic view by estimating the encounter probability among people as they move around a city. To accomplish this, we perform a high resolution probability estimation separately on general weekdays and weekends. For each type of day, we divide the day into 24 one-hour time windows, i.e., $(T_1, T_2, ...T_{24})$ for general weekdays and $(T_{25}, T_{26}, ...T_{48})$ for general weekends. For each time window T_j, we estimate, for each individual x, the probability of stay $prob_x(L_i, T_j)$ at different urban locations $(L_1, L_2, ...L_q)$. From these estimations, we further measure the 'co-location' probability for all individual pairs in the city (see Methods). Similar to the definition of *CSI*, the derived co-location probabilities are used as the weights, along with the social similarity measure, to quantify individual segregation in physical space as $PSI_x(L_i, T_j)$. Aggregating these values for a window T_j outputs a time-dependent index for each individual x, i.e., $PSI_x(T_j)$. The value of $PSI_x(T_j)$ ranges from 0 to 1, with larger values indicating more exposure to similar others. Just as the *baseline* of segregation in social space, the *baseline* of $PSI_x(T_j)$ is also 0.5, which assumes that individuals have an equal probability of 'bumping' into others in the city.

By averaging the time-dependent indices, we obtain a single value, $PSI_x = \sum_{j=1}^{48} PSI_x(T_j)/48$, that describes the overall segregation of an individual x in physical space. Repeating this step for all individuals gives us a normal distribution with a mean and standard deviation of 0.571 and 0.074, respectively (Fig. 4A). This suggests that individuals tend to share urban spaces more often with similar others, although the city's overall deviation from the baseline is not large. Different from the segregation patterns in social space (Fig. 2A), the *PSIs* are more concentrated around the mean. Such a smaller interpersonal variation is likely to be an outcome of many physical constraints of human activities. These constraints — such as where one can choose to live and work, the locations available for other activities (e.g., dining, recreation) — could mediate the degree of social isolation for some individuals. While in social space, ones can choose to be more isolated or integrated by ignoring these physical constraints. By correlating *PSIs* with social ranks (Fig. 4B), we obtain a trend similar to that of *CSI*, meaning that the 'rich club' effect exists also in physical space. It is found that all social classes tend to be more segregated than expected. At the same time, however, all of them possess smaller interpersonal variations compared to that of *CSI*. This indicates that the mediation effect of physical constraints applies to all socioeconomic tiers in the city.

Besides the ability to quantify interpersonal variation, the metric can also depict segregation dynamics over time and space. By averaging $PSI_x(T_j)$ of all individuals for each time window T_j, we obtain a curve to describe the hourly variation of the city's overall segregation (Fig. 5A). The result reveals a notable contrast between day time and night time. The night-time values are generally higher, which indicates a higher degree of social segmentation at residential locations. The average *PSIs* drop notably during the day-time, suggesting that urban mobility has a positive effect on mixing social classes in the city. Similar conclusions were previously reached by couple of studies, although they did not consider SES and were focusing on segregation between a few discrete classes[22,29,40]. The degree of social mixing, however, is higher on weekdays than on weekends , with median *PSI* values ranging from 0.537 to 0.551 on weekdays between 10:00 and 20:00, and from 0.559 to 0.579 on weekends during the same hours (Fig. 5A). During day time on weekdays, many activities are employment-related and they primarily occur around where people work. These locations, especially ones that host large employment populations, could contribute to the mixing of people with different socioeconomic background. Weekends account for more activities related to socialization, recreation, and self maintenance. The dominance of such activities could bring people closer to their own socioeconomic classes. By further computing the *PSI* curve for different socioeconomic tiers (Fig. 5B), we find that social classes that are more isolated during night time also tend to be more segregated during the day time. That means the relationship between segregation and social ranks as we obtain in Fig. 4B remains relatively consistent over the 24 hours of a day.

From a spatial point of view, the movement patterns of large populations allow us to quantify the unique characteristics of

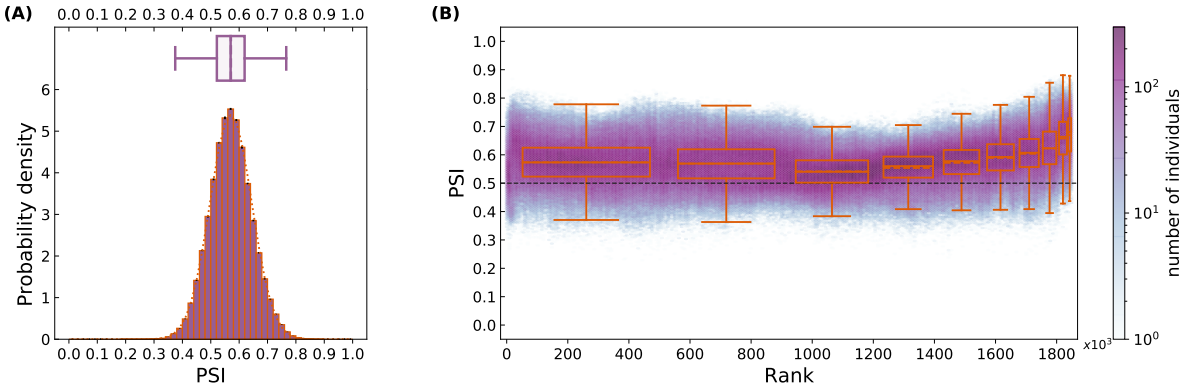

Figure 4. Distribution of *PSI* values: **(A)** Histogram and box plot of *PSI* values obtained during 100 random assignments averaged over all 48 time windows, error bars on top of each bar indicate variation (min and max values) observed over different assignments, dotted line shows fitted normal distribution; **(B)** Relationship between *PSI* and social ranks through a density plot of one of the random assignments averaged over all 48 time windows. Darker colors indicate areas with more individuals, 10 box plots show distribution of *PSI* values within 10 classes with equal cumulative housing price. All means and medians are above baseline value of 0.5. All box plots show median (center line), mean (dashed line), upper and lower quartiles (box limits), 1.5 interquartile range or min/max values (whiskers).

various urban locations and their evolution over time. By aggregating $PSI_x(L_i, T_j)$ values of all users x for each place L_i, we can compute what we call place-based $PSI(L_i, T_j)$ — a time-dependent index for each place L_i. The index quantifies the average level of exposure that visitors experience at location L_i during time window T_j (see Methods for more details). By plotting the place-based indices for a few time windows, we can clearly observe the changing landscape of physical segregation (Fig. 5C). Such information reveals not only the spatial patterns of residential segregation, but also the evolutionary roles of places. For instance, the Sentosa Island is a scenic spot and a tourist attraction, and meanwhile a place with luxury residential communities. The resort island is highly segregated at night, but becomes rather mixed in the day time and evening. The social segregation metric herein proposed is able to reveal such dynamics, that are shaped by the interplay of different types of people (dwellers vs. visitors) and their activities. Thus, the insights gained by applying our methodology and metrics allow going substantially beyond the traditional notion of residential segregation. For instance, depending on the type of places and time of day, the proposed metric can be used to better understand segregation at workplaces[15,40] or segmentation of social groups in leisure time activities[41,42].

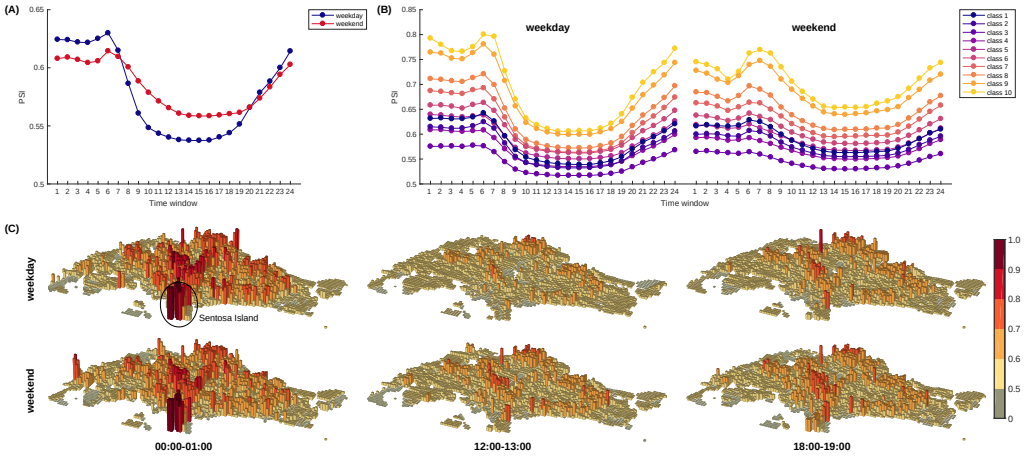

Figure 5. Variations of physical segregation with time. **(A)** Median *PSI* values for each one hour time window on weekdays and weekends. **(B)** Median *PSI* values of 10 socio-economic classes for each one hour time window on weekdays and weekends. **(C)** Average *PSI* of each grid cell for selected time windows, when we expect most people to be at home, at work or perform some outside activities on weekdays and weekends. Heights of the bars indicate values $PSI(L_i, T_j) - 0.5$ showing the difference of segregation from the baseline value.

Discussion

As we disclose the segregation dynamics in social and physical space, a critical question that remains unaddressed is to what extent these two dimensions align with each other. By correlating *CSI* and *PSI* at the individual level, we obtain a weak and positive correlation (Fig. 6A). Previous studies have suggested that the movement similarity between individuals is a strong indication of their proximity in the social network[34,35]. Such an interdependency potentially contributes to this correlation because individual exposure to others in urban space would mirror part of their communication patterns in the network space. On the other hand, the result indicates that, when measured at the individual level, the degrees of individual isolation in the communication network and urban space are not tightly bundled. Therefore, observing one's segregation in a given dimension cannot be leveraged to fully diagnose the degree of isolation in another.

However, the analysis brings very different results when correlation is measured at the aggregate level. By gradually grouping individuals with similar social ranks (Fig. 6B to Fig. 6F), we find a dramatic increase of correlation as individuals are aggregated (Fig. 6G) into larger groups. Interestingly, both the Pearson and Spearman correlation tends to level at or above 0.9 when numbers of social groups are smaller than 1000, corresponding to groups of cardinality larger than a few hundreds. The result reveals two interesting aspects of segregation dynamics. On the one hand, the social exposure an individual experienced can be very different in physical and social-network space. Such differences can be affected by their occupations, living styles, as well as variations in their online and offline behaviour. This echoes back to the concept of "choice homophily" that describes individual and psychological preferences of social interactions[43]. On the other hand, the communication and physical segregation become highly correlated when social groups are large in size. This suggests that there might exist structural causes, such as the homogeneity of neighborhoods, housing policies, access to public transit, and digital divide, that shape segregation patterns in cities. This corresponds to the view of "induced homophily" that aims to reveal structural opportunities for social interactions[43,44]. It seems that the collective dynamics of segregation in physical and social-network space reinforce each other. While public policies could have a more direct impact on the social mixing in physical space, it would be very interesting to monitor the co-varying dynamics of these two types of segregation, especially after the introduction of particular intervention strategies (e.g., housing policies that require a certain level of social mixing). This is one of the possible directions for future research.

In this study, we introduce a framework to measure segregation in an integrated urban physical-social space. The metrics enable us to depict segregation not only for individuals, but also for places in a city, and their evolution over time. As technology progresses, new datasets that are suitable for segregation studies are increasingly being generated. They are also offering rich information of human dynamics at finer granularities. Classical segregation measures, which were mainly designed to tackle segregation using static data (e.g., census) and over well-defined social groups (e.g., race and ethnicity), are not adequate to support segregation analysis in highly dynamic settings. The current framework contributes to the segregation literature from the following perspectives. First, it enables the measurement of segregation at the individual level, and the notion of baseline segregation makes it easy to compare the degree of isolation across individuals and social groups. Second, the similarity measure based on social ranks makes it possible to quantify segregation over continuous variables. With such a measure, some important segregation types (e.g., income segregation) can be well quantified without the need to group individuals into socioeconomic tiers in arbitrary ways. For example, the Theil's entropy index[45] was often used in existing studies to quantify residential segregation. Computing the Theil's entropy requires the underlying population to be categorized into discrete social groups (e.g., race or ethnicity). When measuring income segregation using Theil's index, the social classes need to be predefined. However, the ways social classes are defined can be somewhat arbitrary. In Supplementary, by dividing phone users into predefined numbers of social classes (based on their inferred SES), we apply the Theil's entropy index to measure the segregation of the city during different time periods of a day. By comparing the results with ones derived from our proposed measure (Fig. 5), we find that the Theil's index and the proposed physical segregation index (*PSI*) achieve compatible results when describing the overall segregation of the city as well as the spatial heterogeneity of social mixing (see Fig. S.9 and Fig. S.10 for more details). The *PSI* measure, however, does not require population to be divided into discrete social classes. Meanwhile, it is able to portray the degree of physical segregation down to the individual level (Fig. 4).

This study also reveals the linkage between segregation in the social and physical space, and highlights the importance of data fusion for studying the coupling dynamics. The proposed segregation indices can be applied or modified over data that are already available in many countries and cities. Examples include cellular operator's data[46,47], social media data[6,48], geolocated transactions[49,50], and multi-modal datasets that couple the social-spatial and economic activities (e.g., spending behaviour) of the populations[30,51]. As measurement of segregation concerns not only the behaviors of people, but also their socioeconomic characteristics, there will be privacy concerns regarding accessing and especially merging these kinds of datasets. Due to these reasons, in this work we approximate individuals' socioeconomic status through the estimated housing price at their home location. The approximation, although not perfect, offers a new perspective on privacy issues while capturing a multidimensional view of urban segregation. By simulating random assignment of housing price, we show that our conclusions about level of segregation of different classes remain consistent and robust even when uncertainties exist in the measurements.

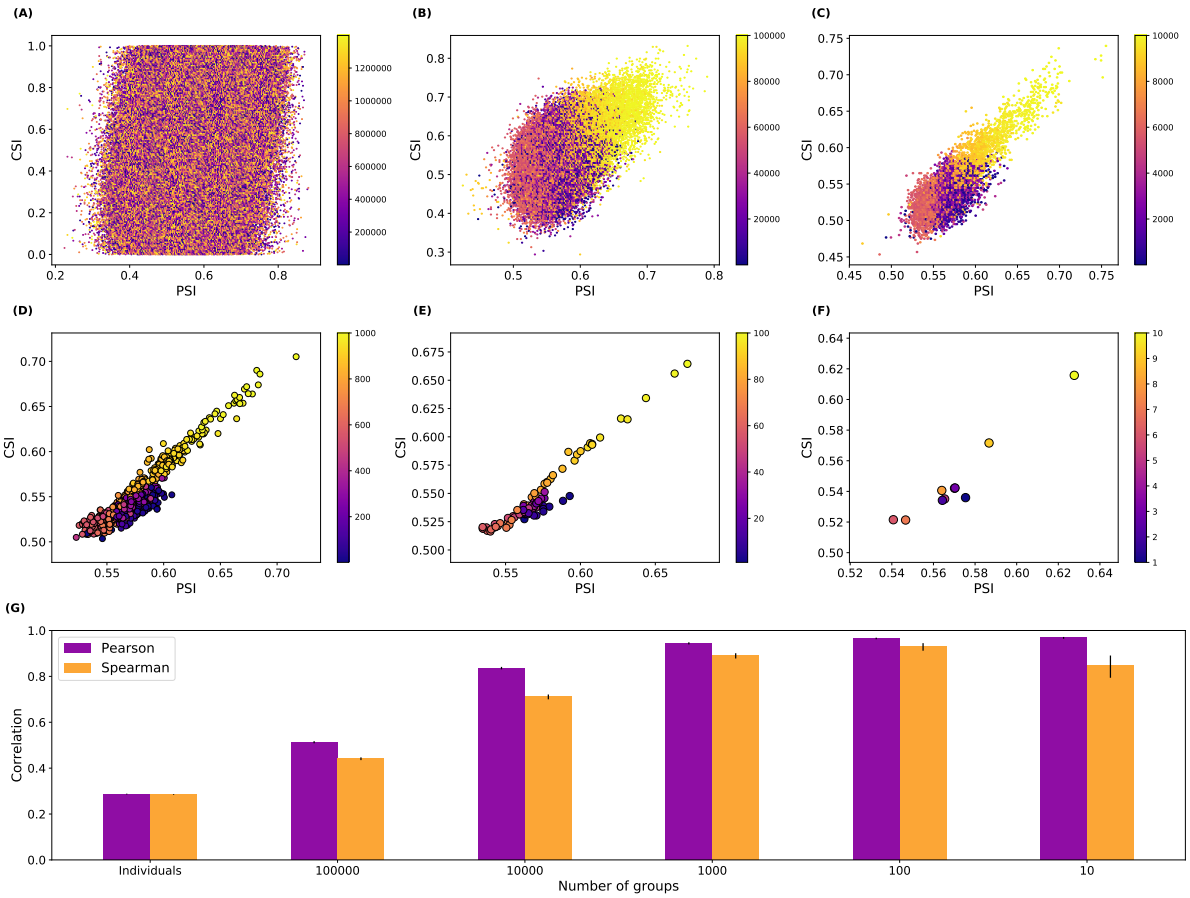

Figure 6. Correlations between *CSI* and *PSI* values. Scatter plots are produced based on one random assignment to illustrate the relationship between *CSI* and *PSI* at **(A)** individual level and when individuals are further aggregated into **(B)** 100,000, **(C)** 10,000, **(D)** 1000, **(E)** 100, and **(F)** 10 groups. Darker colors represent individuals or groups in lower socioeconomic classes. **(G)** Pearson's and Spearman's correlations between *CSI* and *PSI* at the above six levels of aggregation. Error bars indicate variation (min and max values) over the 100 random assignments.

Our results reveal that the degree of urban segregation is not only affected by the socioeconomic configuration of the city (e.g., where rich and poor reside), but also the presence of 'homophily' in the society[4]. The analysis of 'homophily' distance suggests a relatively stronger homophily effect for the poor and rich, and reveals their tendencies to connect with similar others even when the 'distance decay' effect is controlled. It would be interesting to apply our metrics in other cities in the future, and examine whether the coexistence of this two-level effect (distance decay and homophily) can be rediscovered.

In this research, we apply the proposed metrics in one single city, Singapore. The level of segregation, as observed in both physical and social space, tends to be relatively stable across the lower and middle classes, but notably higher across the wealthier classes. Such relationship, which is tied to the uniqueness of the city, can be affected by many things such as housing policy, transportation networks, and urban spatial structure. Through the years, Singapore has deployed efficient public transportation systems to facilitate the everyday mobility of the city populations[52]. It also implemented planning strategies that fostered a more polycentric urban form[53–55]. These can all have a positive effect on the social mixing of the city. On the other hand, Singapore still possesses a high level of income inequality. Although the city tried to foster the mixing of different ethnic groups through public housing policies, less has been done to explicitly foster the mixing between social classes. Disentangling these effects is beyond the scope of this research. However, it would be meaningful to relate them to our measurements in the future to better understand the influential factors of social segregation. We believe the relationship between SES and the level of segregation (Fig. 2 and Fig. 4) is not universal across cities. For instance, some cities might possess a 'U' shape and some others might show more distinct patterns. What are the major types of relationships that distinguish cities? Do cities with similar segregation dynamics share similar institutional and policy contexts? These questions, which are worth investigating, call for cross-city comparisons in the future. And we hope that the framework presented in this study could be really helpful with providing answers to such questions and the methods described here will be applied to other cities once the appropriate

data become available to researchers.

We want to point out a few limitations of this research. First, although we have evaluated the robustness of the results based on multiple runs of SES assignment, we are not able to fully capture the difference of individuals, especially among those who are identified with the same home cell. In many places of Singapore, due to housing in building with large number of floors, it could happen that many people are assigned to the same home cell and with the same SES indicator. The precision of SES inference is limited by the model's ability to differentiate people who live in the same building. Such a limitation leads to an underestimate of individual variations, which will influence the segregation measurements. In this regard, the results need to be interpreted with caution. Second, since we limit our analysis to active phone users, certain demographic tiers (e.g., the elders who tend to use mobile phones less frequently) will be underrepresented. This will to some extent distort the distribution of SES indicator (housing price). The effect of this bias, which is not investigated in this research, can be better understood by applying the methodologies over datasets that cover a longer time period (e.g., 1 year). Also, since the phone calls and text messages are observations of the study, mobile phone users who rely more on mobile application-based communication could be underrepresented in the analysis. However, we believe that in 2011, the time at which the dataset was collected, online messengers were still not that widespread and people used to text. This limitation could be addressed by applying described methods to new datasets that contain information about mobile data usage. Third, the segregation measurements used in this study are mainly approached from the perspective of income inequality based on housing data. As suggested by previous literature[6,15,40], racial segregation is a significant factor that impacts social segregation in cities. As the datasets do not capture the racial makeup of the population, we are unable to investigate or control the impact of racial segregation on our measurements. It would be meaningful to examine this issue in the future when appropriate datasets become available. Nevertheless, the proposed framework contributes to the broad literature by providing a new perspective of segregation in a coupled physical-social space. The insights could help better predict and explain segregation under different urban settings and social contexts, which could inspire other studies that focus on inequality and segregation-health associations, and inform new policies for social integration.

To sum up, our analysis has clearly shown that traditional, "place-based" practices to improve social integration such as stringent housing policies and easy access to transportation infrastructure, as those put in place in the city of Singapore, are not sufficient per-se to achieve a high degree of mixing across social classes. A possible strategy to further reduce social segregation would be to complement "place-based" practices with "activity-based" practices, where groups of people from different social classes are actively engaged in social bonding activities. For instance, organizing gatherings, discussion workshops, and other events directed towards suitably selected groups could positively contribute to build bonds across social classes.

Methods

Data

We analyze a large scale call detail record (CDR) dataset collected from a major mobile phone operator in Singapore, with a market share of more than 45% as of 2011. The dataset covers a period of 50 days in 2011. In this study, we use a subset which includes all individuals with *active days of phone usage* ≥ 25 *days*. On the one hand, it allows us to focus more on local residents by filtering short-term subscribers such as tourists. On the other hand, it would mitigate the data sparsity issue and ensure that the users kept for the segregation analysis have sufficient observation days, which increase the reliability of home location estimation and mobility analysis (e.g., estimation of co-location probability). At the end of this filtering stage, the dataset contained data observations referring to about 2.1 million users, while total population of Singapore in 2011 was just below 5.2 million.

Two socioeconomic datasets — residential property price and the Household Interview Travel Survey (HITS) — are used to facilitate the estimation of phone users' socioeconomic status (SES). The housing price dataset is acquired from a private company in Singapore (99.co). The dataset includes information of thousands of residential properties collected between 2011 and 2012. Each record maps a unique housing property and the information of the property type (condo, landed, or HDB), the geographic coordinates (latitude and longitude) as well as the total sale price of one housing unit.

The HITS dataset was collected by the Singapore Land Transport Authority (LTA) in 2012. It includes a one-day travel diary of 35,715 individuals (sampling rate of roughly 1%) along with other socioeconomic attributes that were self-reported by the respondents (e.g., monthly income).The HITS data is not used in this study to directly infer phone users' SES. We use information of individual income in the data and correlate it with housing price, which proves that housing price is a reasonable indicator of the SES of the underlying populations (see Supplementary for detailed information).

Home detection and SES assignment

In the dataset we use, all events are associated with a cell tower which a phone was connected to at the moment when event occurred. We use a Voronoi polygon as an approximation of a cell tower service area. But in urban areas with high population towers are much denser than in areas with low population. This leads to a very uneven distribution of Voronoi polygon sizes.

To account for this and to make our approach applicable for comparison of different cities, we adopted 500×500 m grid, i.e. we divide the whole study area into a network of square cells of size 500 meters by 500 meters. We try to identify one of these cells as a phone user's home location. We also calculate user's probability of staying at each cell to calculate *PSI*, as described further. To determine user's home location, for each cell we count number of nights when at least one event occurred at this cell between 7pm and 9am. We count an event as occurring at a cell whenever it occurs at the base tower whose Voronoi polygon intersects with this cell. A cell with the events on the highest number of nights is assigned as home location. In Supplementary in Figure S.3 we show that obtained distribution is well correlated with official census data.

For every grid cell where we have at least one record of apartment/house purchase and for every user whose home location was assigned to this cell, we assign a user to the housing price drawn at random from the uniform distribution over the linear interpolation of the prices. More precisely, for every user living at a cell with a sorted list of prices $(p_1, p_2, \dots p_m)$, we draw a number u from the uniform distribution $u \sim U(1,m)$. Then the assigned price p is calculated as $p = p_{\lfloor u \rfloor} + (p_{\lceil u \rceil} - p_{\lfloor u \rfloor})(u - \lfloor u \rfloor)$. This price is used as an approximation of individuals' SES and based on it we can rank all users and calculate *CSI* and *PSI*. In total, we were able to assign SES to about 1.8 million users. To verify robustness of our method and obtained results, we repeat this assignment 100 times, and every time we calculate new values of *CSI* and *PSI* for each user. While each individual's rank, *CSI* and *PSI* can change significantly from one assignment to another, overall distributions of *CSI* and *PSI* values are shown to remain very stable. We discuss this fact further in Supplementary. In figures of the main body, we present values for one assignment and show error bars indicating minimum and maximum value, where appropriate.

The social distance/similarity measure

Given individual x with rank r_x in the sequence $(r_n)_{n=1}^N = (1,2,...,N)$, if $r_x \leq \frac{N}{2}$, then for any person j with rank $r_j < r_x$, the social distance from x to j is:

$$d_{x\to j} = \frac{(2r_x - 2r_j - 1) + 0.5}{N-1} = \frac{4r_x - 4r_j - 1}{2(N-1)} \tag{3}$$

When $r_x < r_j \leq 2r_x - 1$, the social distance measure can be calculated similarly. When $r_j \geq 2r_x$, since there is no other individual with rank r_k that satisfies $|r_k - r_x| = |r_x - r_j|$, the distance $d_{x\to j}$ will be:

$$d_{x\to j} = \frac{r_j - 1}{N-1} \tag{4}$$

Similarly, we can derive the social distance measure for $r_x > \frac{N}{2}$.

Thus, the complete equations for calculating the social distance metric are summarized as follows:

$$\begin{aligned}
&\text{when } r_x \leq \frac{N}{2} \qquad d_{x\to j} = \begin{cases} 0, & \text{if } r_j = r_x \\ \frac{|4r_x - 4r_j| - 1}{2(N-1)}, & \text{if } r_j \leq 2r_x - 1 \text{ and } r_j \neq r_x \\ \frac{r_j - 1}{N-1}, & \text{if } r_j > 2r_x - 1 \end{cases} \\
&\text{when } r_x > \frac{N}{2} \qquad d_{x\to j} = \begin{cases} 0, & \text{if } r_j = r_x \\ \frac{|4r_x - 4r_j| - 1}{2(N-1)}, & \text{if } r_j \geq 2r_x - N \text{ and } r_j \neq r_x \\ \frac{N - r_j}{N-1}, & \text{if } r_j < 2r_x - N \end{cases}
\end{aligned} \tag{5}$$

The social similarity from x to j can simply be calculated as:

$$s_{x\to j} = 1 - d_{x\to j} \tag{6}$$

Note that when individuals are ranked based on the housing price in the random assignment model, the sequence $(r_n)_{n=1}^N = (1,2,...,N)$ produces an incomplete order of individuals because some phone users might be assigned with the same SES value (i.e., the associated housing price). To address this issue, we adopt *fractional ranking*. Basically, for individuals that have a tie in ranking, their fractional ranking is computed as the mean of what they would have under ordinal rankings. For example, if individual A has the lowest housing price value, while individual B and C share the same price which is lower than that of individual D, then the rank of the four individuals would be 1, 2.5, 2.5, and 4, respectively.

Physical Segregation Index

Given the cellphone trace of an individual x as a list of tuples $\{(l_1,t_1),(l_2,t_2),...,(l_n,t_n)\}$, where l_i denotes the user's location (i.e., cellphone tower) at time point t_i, the probability that user x stays at location L during a defined time period T is:

$$prob_x(L,T) = m_x(L,T)/n_x(T) \tag{7}$$

where $m_x(L,T)$ denotes the total number of times x is observed at location L during time period T, and $n_x(T)$ denotes the total number of times x is observed during time period T. Note that:

$$\sum_{L \in Loc_x} m_x(L,T) = n_x(T) \tag{8}$$

where Loc_x is the set of locations visited by x during T.

Note that the individual's cellphone communications do not take place regularly over time. People could make several phone calls in a short period of time and then none for hours. Hence, $prob_x(L,T)$ could be biased due to the "bursty" nature of CDRs. To control this effect, when we measure $m_x(L,T)$, if an individual x is observed multiple times at location L during a one-hour time window (e.g., 07:00 – 08:00), we only consider them as one entry. To interpolate this probability from cellphone towers on the grid cells, we calculate $int(L_{ct},L_{gc})$ – the area of intersection between a Voronoi polygon L_{ct} and a grid cell L_{gc}. Then the probability that user x stays at grid cell L_{gc} is calculated as $prob_x(L_{gc},T) = \frac{prob_x(L_{ct},T)\cdot int(L_{ct},L_{gc})}{area(L_{ct})}$, where $area(L)$ is the area of L. By doing so, we could establish a matrix A, which records each mobile phone user's probability of stay at different locations:

$$A = \begin{pmatrix} prob_{x_1}(L_1,T) & prob_{x_1}(L_2,T) & \cdots & prob_{x_1}(L_k,T) \\ prob_{x_2}(L_1,T) & prob_{x_2}(L_2,T) & \cdots & prob_{x_2}(L_k,T) \\ \vdots & \vdots & \ddots & \vdots \\ prob_{x_N}(L_1,T) & prob_{x_N}(L_2,T) & \cdots & prob_{x_N}(L_k,T) \end{pmatrix} \tag{9}$$

Here x_i denotes the i^{th} mobile phone user, N denotes the total number of mobile phone users in the dataset; k denotes the total number of spatial units (i.e., grid cells). Note that each row in the matrix sums to 1.

We next introduce how the physical segregation index (PSI) can be calculated for an individual x. At each location $L \in Loc_x$ that has been visited by x (i.e., $p_x(L,T) > 0$), we retrieve all other individuals y for whom $p_y(L,T) > 0$, i.e., $y \in U(L,T)$ where $U(L,T) = \{y | p_y(L,T) > 0\}$. The physical segregation index of x at location L is defined as:

$$\begin{aligned} PSI_x(L,T) &= \sum_{y \in U(L,T)} prob_x(L,T)\cdot prob_y(L,T)\cdot s_{x\to y} \quad / \sum_{y \in U(L,T)} prob_x(L,T)\cdot prob_y(L,T) \\ &= \sum_{y \in U(L,T)} prob_y(L,T)\cdot s_{x\to y} \quad / \sum_{y \in U(L,T)} prob_y(L,T) \end{aligned} \tag{10}$$

where $s_{x\to y}$ refers to the social similarity measure we defined in the previous section. Thus, $PSI_x(L,T)$ can be viewed as the weighted average of social similarity between person x and other visitors at location L. The weight of $s_{x\to y}$ is $prob_x(L,T)\cdot prob_y(L,T)$, which is the spatial co-location rate of the two individuals. A high value of $PSI_x(L,T)$ indicates that x is more likely to meet with similar others at this location L.

We can then calculate the physical segregation index for x during time window T as the sum of segregation index at all locations, weighted by the corresponding stay probability:

$$\begin{aligned} PSI_x(T) &= \sum_{L \in Loc_x} prob_x(L,T)\cdot PSI_x(L,T) \quad / \sum_{L \in Loc_x} prob_x(L,T) \\ &= \sum_{L \in Loc_x} prob_x(L,T)\cdot PSI_x(L,T) \end{aligned} \tag{11}$$

According to equation 11, the location where x has a higher probability of stay will have a higher impact on $PSI_x(T)$.

Similarly, taking the perspective of each spatial unit, we can calculate place-based physical segregation index of place L during time window T by aggregating $PSI_x(L,T)$ of all users who visited L during T:

$$PSI(L,T) = \sum_{\{x | L \in Loc_x\}} prob_x(L,T)\cdot PSI_x(L,T) \quad / \sum_{\{x | L \in Loc_x\}} prob_x(L,T) \tag{12}$$

$PSI(L,T)$ quantifies the average level of exposure that visitors of L experience. Since all *PSI* measures are defined as weighted similarity, all of them scale from 0 to 1, taking higher values when similar users have high co-location rate or, in other words, when people are surrounded by similar others. For the same reason, all *baseline PSI*s, e.i. when all people have the same probability of being at some place, take the same value as the average similarity, i.e. equal to 0.5 (see Supplementary for more details).

References

1. Ravallion, M. Income inequality in the developing world. *Science* **344**, 851–855 (2014).
2. Saez, E. & Zucman, G. Wealth inequality in the united states since 1913: Evidence from capitalized income tax data. *The Q. J. Econ.* **131**, 519–578 (2016).
3. Chodrow, P. S. Structure and information in spatial segregation. *Proc. Natl. Acad. Sci.* 201708201 (2017).
4. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annu. review sociology* **27**, 415–444 (2001).
5. Schelling, T. C. *Micromotives and macrobehavior* (WW Norton & Company, 2006).
6. Wang, Q., Phillips, N. E., Small, M. L. & Sampson, R. J. Urban mobility and neighborhood isolation in america's 50 largest cities. *Proc. Natl. Acad. Sci.* **115**, 7735–7740 (2018).
7. Peterson, R. D. & Krivo, L. J. *Divergent social worlds: Neighborhood crime and the racial-spatial divide* (Russell Sage Foundation, 2010).
8. Wodtke, G. T., Harding, D. J. & Elwert, F. Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *Am. Sociol. Rev.* **76**, 713–736 (2011).
9. Quillian, L. Does segregation create winners and losers? residential segregation and inequality in educational attainment. *Soc. Probl.* **61**, 402–426 (2014).
10. Chetty, R., Hendren, N. & Katz, L. F. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *Am. Econ. Rev.* **106**, 855–902 (2016).
11. Williams, D. R. & Collins, C. Racial residential segregation: a fundamental cause of racial disparities in health. *Public health reports* **116**, 404 (2001).
12. Bischoff, K. & Reardon, S. F. Residential segregation by income, 1970-2009. *Divers. disparities: Am. enters a new century* **43** (2014).
13. Musterd, S., Marcińczak, S., Van Ham, M. & Tammaru, T. Socioeconomic segregation in european capital cities. increasing separation between poor and rich. *Urban Geogr.* **38**, 1062–1083 (2017).
14. Massey, D. S. & Denton, N. A. The dimensions of residential segregation. *Soc. forces* **67**, 281–315 (1988).
15. Hellerstein, J. K. & Neumark, D. Workplace segregation in the united states: Race, ethnicity, and skill. *The review economics statistics* **90**, 459–477 (2008).
16. Kwan, M.-P. Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and giscience. *Annals Assoc. Am. Geogr.* **103**, 1078–1086 (2013).
17. Massey, D. S. & Denton, N. A. Trends in the residential segregation of blacks, hispanics, and asians: 1970-1980. *Am. sociological review* 802–825 (1987).
18. Kwan, M.-P. From place-based to people-based exposure measures. *Soc. science & medicine* **69**, 1311–1313 (2009).
19. Wong, D. W. & Shaw, S.-L. Measuring segregation: An activity space approach. *J. geographical systems* **13**, 127–145 (2011).
20. Farber, S., O'Kelly, M., Miller, H. J. & Neutens, T. Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *J. Transp. Geogr.* **49**, 26–38 (2015).
21. Järv, O., Müürisepp, K., Ahas, R., Derudder, B. & Witlox, F. Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Stud.* **52**, 2680–2698 (2015).
22. Le Roux, G., Vallée, J. & Commenges, H. Social segregation around the clock in the Paris region (France). *J. Transp. Geogr.* **59**, 134–145 (2017).
23. Tan, Y., Kwan, M.-P. & Chai, Y. Examining the impacts of ethnicity on space-time behavior: Evidence from the city of Xining, China. *Cities* **64**, 26–36 (2017).

24. Silm, S., Ahas, R. & Mooses, V. Are younger age groups less segregated? Measuring ethnic segregation in activity spaces using mobile phone data. *J. Ethn. Migr. Stud.* **44**, 1797–1817 (2018).

25. Dannemann, T., Sotomayor-Gómez, B. & Samaniego, H. The time geography of segregation during working hours. *Royal Soc. open science* **5**, 180749 (2018).

26. Östh, J., Shuttleworth, I. & Niedomysl, T. Spatial and temporal patterns of economic segregation in Sweden's metropolitan areas: A mobility approach. *Environ. Plan. A: Econ. Space* **50**, 809–825 (2018).

27. Zhang, X., Wang, J., Kwan, M.-P. & Chai, Y. Reside nearby, behave apart? Activity-space-based segregation among residents of various types of housing in Beijing, China. *Cities* **88**, 166–180 (2019).

28. Gao, J., Zhang, Y.-C. & Zhou, T. Computational socioeconomics. *Phys. Reports* **817**, 1 – 104, DOI: https://doi.org/10.1016/j.physrep.2019.05.002 (2019). Computational Socioeconomics.

29. Silm, S. & Ahas, R. The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset. *Soc. Sci. Res.* **47**, 30 – 43, DOI: https://doi.org/10.1016/j.ssresearch.2014.03.011 (2014).

30. Leo, Y., Fleury, E., Alvarez-Hamelin, J. I., Sarraute, C. & Karsai, M. Socioeconomic correlations and stratification in social-communication networks. *J. The Royal Soc. Interface* **13**, 20160598 (2016).

31. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779 (2008).

32. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).

33. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).

34. Cho, E., Myers, S. A. & Leskovec, J. 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 (ACM, 2011).

35. Wang, D., Pedreschi, D., Song, C., Giannotti, F. & Barabasi, A.-L. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1100–1108 (Acm, 2011).

36. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins, A. Geographic routing in social networks. *Proc. Natl. Acad. Sci.* **102**, 11623–11628 (2005).

37. Lambiotte, R. *et al.* Geographical dispersal of mobile communication networks. *Phys. A: Stat. Mech. its Appl.* **387**, 5317–5325 (2008).

38. Backstrom, L., Sun, E. & Marlow, C. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, 61–70 (ACM, 2010).

39. Krings, G., Calabrese, F., Ratti, C. & Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech. Theory Exp.* **2009**, L07003, DOI: 10.1088/1742-5468/2009/07/l07003 (2009).

40. Ellis, M., Wright, R. & Parks, V. Work together, live apart? Geographies of racial and ethnic segregation at home and at work. *Annals Assoc. Am. Geogr.* **94**, 620–637 (2004).

41. Kamenik, K., Tammaru, T. & Toomet, O. Ethnic segmentation in leisure time activities in Estonia. *Leis. Stud.* **34**, 566–587 (2015).

42. Kukk, K., Van Ham, M. & Tammaru, T. Ethnicity of leisure: A domains approach to ethnic integration during free time activities. *Tijdschrift voor economische en sociale geografie* **110**, 289–302 (2019).

43. McPherson, J. M. & Smith-Lovin, L. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *Am. sociological review* 370–379 (1987).

44. Kossinets, G. & Watts, D. J. Origins of homophily in an evolving social network. *Am. journal sociology* **115**, 405–450 (2009).

45. Theil, H. *Statistical decomposition analysis* (North-Holland Publishing Company Amsterdam, 1972).

46. Deville, P. *et al.* Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci.* **111**, 15888–15893 (2014).

47. Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).

48. Crandall, D. J. *et al.* Inferring social ties from geographic coincidences. *Proc. Natl. Acad. Sci.* **107**, 22436–22441 (2010).

49. Singh, V. K., Bozkaya, B. & Pentland, A. Money walks: implicit mobility behavior and financial well-being. *PloS one* **10**, e0136628 (2015).

50. Sobolevsky, S. *et al.* Cities through the prism of people's spending behavior. *PloS one* **11**, e0146291 (2016).

51. Di Clemente, R. *et al.* Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nat. communications* **9** (2018).

52. Xu, Y., Belyi, A., Bojic, I. & Ratti, C. Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Comput. Environ. Urban Syst.* **72**, 51–67 (2018).

53. Zhong, C., Arisona, S. M., Huang, X., Batty, M. & Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *Int. J. Geogr. Inf. Sci.* **28**, 2178–2199 (2014).

54. Xu, Y., Belyi, A., Bojic, I. & Ratti, C. How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data. *Transactions GIS* **21**, 468–487, DOI: 10.1111/tgis.12285 (2017). https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12285.

55. Zhang, X., Xu, Y., Tu, W. & Ratti, C. Do different datasets tell the same story about urban mobility—a comparative study of public transit and taxi usage. *J. Transp. Geogr.* **70**, 78–90 (2018).

Author contributions statement

Y.X., A.B., P.S. and C.R. defined the problems and conceived the experiments. Y.X. and A.B. designed and performed the analysis. Y.X, A.B. and P.S. analyzed the results. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Information

Distribution of housing units and prices

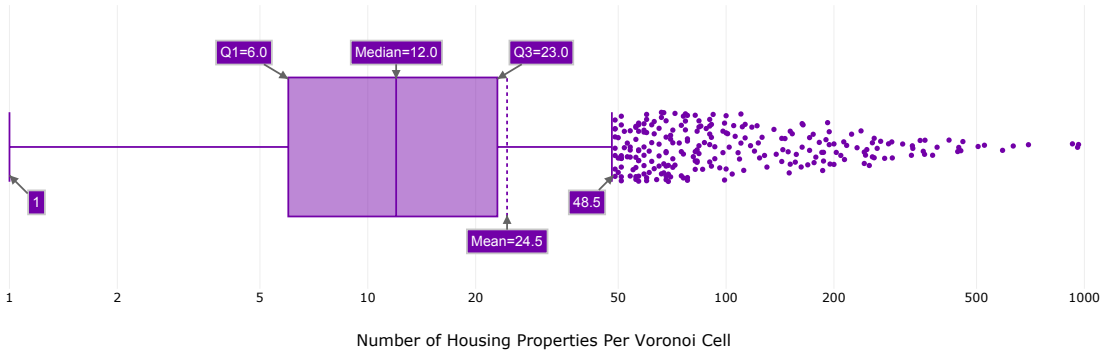

Figure S.1. Distribution of number of price values per tower service area.

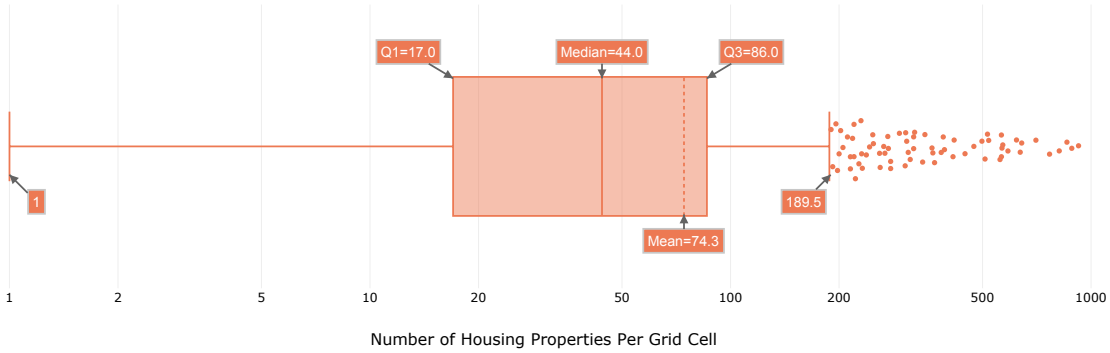

Figure S.2. Distribution of number of price values per 500m grid cell.

In Fig. S.1 and Fig. S.2 we show the distribution of numbers of housing prices per Voronoi cell and per 500m grid cell. Since most of the Voronoi cells in dense residential areas are much smaller than 500m, by changing spatial resolution to the 500m grid cells we increase the average number of prices per cell. This improves the quality of our estimate of the distribution of prices per cell and helps to assign more appropriate ranks to users.

Correlation between census and estimated populations

To evaluate whether our estimated home locations of each individual reflect the population distribution in Singapore we first calculate the total number of cellphone users with home location in each planning area and then compare these values with census data available from Department of Statistics Singapore, for the year 2010. As shown in Figure S.3, we find that the total number of cellphone users sampled in each planning area is strongly correlated with the population distribution recorded by the census data, with a Pearson's correlation of 97%.

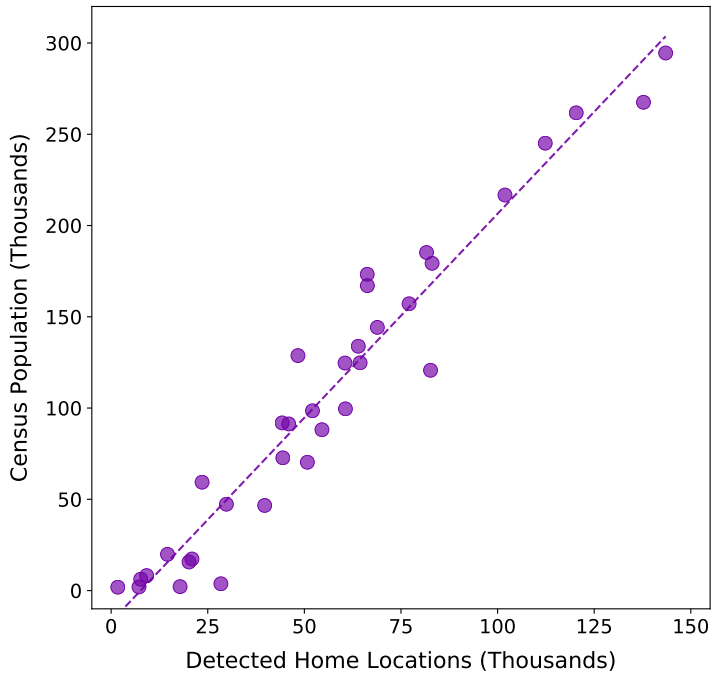

Figure S.3. Correlation between the number of detected home locations and census population (by planning area in Singapore) (Pearson's $R = 0.97$).

Correlation between housing price and income

To assess the feasibility of using housing price as an indicator of individuals' socioeconomic status (SES), we perform a correlation analysis at the level of Singapore's planning area by relating the housing price data and income data from the Household Interview Travel Survey (HITS). We first extract all the individuals from the HITS data who reported their monthly income (12,111 in total). We then aggregate individuals — based on the postal code of their reported residencies — by planning area, and calculate the average monthly income at each planning area. We then compute the average sale price of housing units in each planning area and correlate the two variables. We find that the average monthly income matches relatively well with the mean housing price (Fig. S.4A) except for three outliers (Southern Islands, Sungei Kadut, and Novena). By further exploring the HITS dataset, we think this is partially caused by the sampling bias when individuals were selected for the travel survey. For example, only two individuals in Southern Islands were sampled from the 2011 HITS survey, and both them reported a monthly income of 500 SGD. However, Southern islands is well known as a planning area with many luxury housing communities. Housing price, in this sense, could be a more reasonable indicator of individual SES given the sparsity of HITS income data. Fig. S.4B shows the relationship between the two variables after filtering these three outliers. The Pearson's R is 0.88, which suggests that it is reasonable to use housing price to approximate the SES of the underlying populations. Note, that in the main study we did not exclude residents of these tree outlying areas, as we believe that the bias comes mostly from the HITS data.

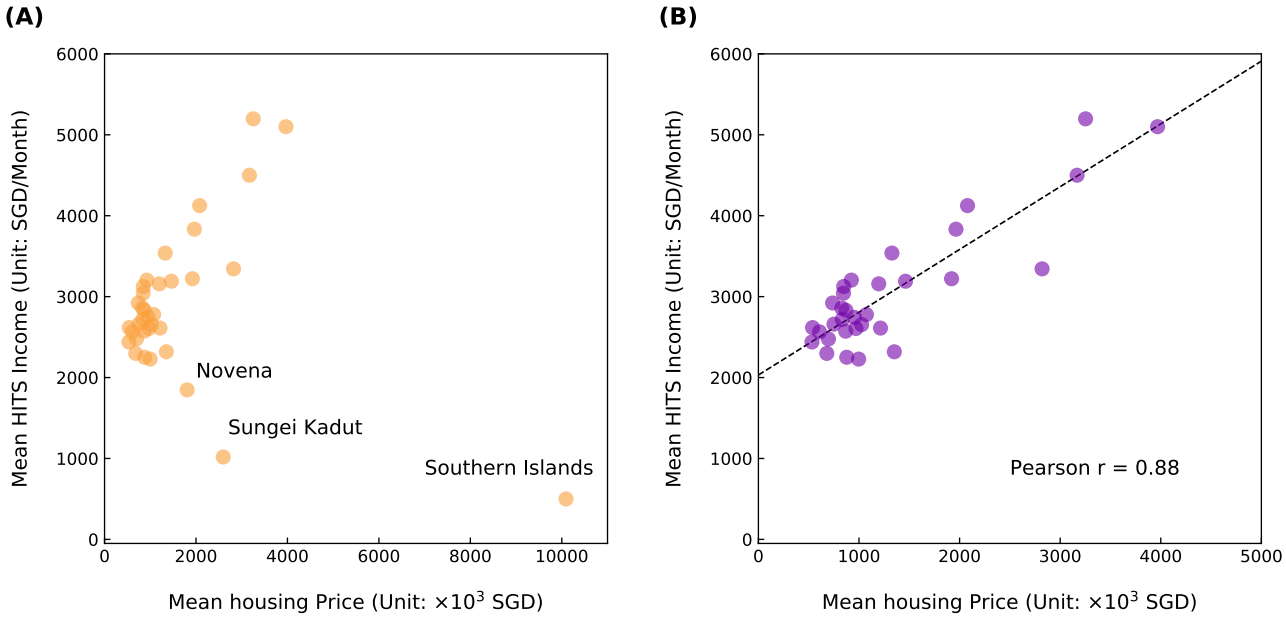

Figure S.4. The relationship between mean housing price and average monthly income **(A)** at the level of Singapore's planning area; **(B)** The correlation between the two variables after filtering the three outliers (Pearson's $R = 0.88$).

Random assignment of housing price

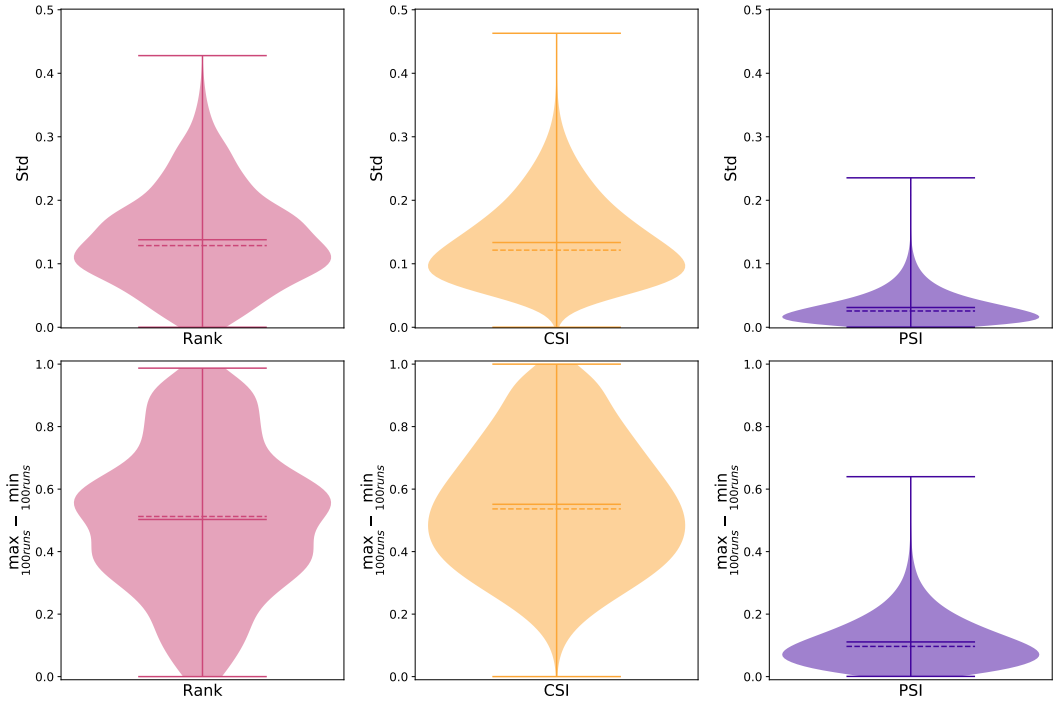

Figure S.5. Distribution of standard deviation and spread between min and max value of ranks, *CSI* and *PSI* of each user between 100 random assignments of housing prices.

In Fig. S.5 we show how ranks, *CSI* and *PSI* change between assignments. One can see that for some users these values could change quite a lot. However, the overall standard deviations are relatively low, indicating that individuals tend to be assigned with similar SES values across the 100 assignments. Moreover, the overall distributions of *CSI* and *PSI* remain very similar from one assignment to another. This confirms that our conclusions about differences in segregation between classes are robust and do not depend on a particular way of assigning housing prices.

Segregation level of an individual under the assumption of a null model

Under the assumption of a *null model*, where each individual interacts equally with other individuals in a city (in the social or physical space), the segregation level of an individual can be quantified as the weighted sum of social similarity between him/her to all individuals, i.e., $\frac{1}{N} \cdot \sum_{j=1}^{N} s_{x \to j}$. We can prove that, in case of unique ranks $r_x = x$, this value is independent of the social rank x and is always 0.5. Given an individual's rank $x \leq \frac{N}{2}$, the weighted sum of social similarity can be calculated as:

$$\begin{aligned}
\frac{1}{N} \cdot \sum_{j=1}^{N} s_{x \to j} &= \frac{1}{N} \cdot \left(\sum_{j=1}^{x-1} s_{x \to j} + s_{x \to x} + \sum_{j=x+1}^{N} s_{x \to j} \right) \\
&= \frac{1}{N} \cdot \left(\sum_{j=1}^{x-1} s_{x \to j} + s_{x \to x} + \sum_{j=x+1}^{2x-1} s_{x \to j} + \sum_{j=2x}^{N} s_{x \to j} \right) \\
&= \frac{1}{N} \cdot \left(2 * \sum_{j=1}^{x-1} s_{x \to j} + s_{x \to x} + \sum_{j=2x}^{N} s_{x \to j} \right) \\
&= \frac{1}{N} \cdot \left(\sum_{j=1}^{x-1} \frac{2N - 4x - 1}{N-1} + \sum_{j=1}^{x-1} \frac{4j}{N-1} + 1 + \sum_{j=2x}^{N} \frac{N}{N-1} - \sum_{j=2x}^{N} \frac{j}{N-1} \right) \\
&= \frac{1}{N} \cdot \frac{N^2 - N}{2(N-1)} \\
&= \frac{1}{2}
\end{aligned}$$

We can prove the case similarly for $x > \frac{N}{2}$. The proof shows that this value is independent of social rank x.

So, if we assume that every person has equal chances to communicate with any other person, then:

$$\begin{aligned}
E(CSI_x) &= E \frac{\sum_{j=1}^{N} f_j \cdot s_{x \to j}}{\sum_{j=1}^{N} f_j} = \sum_{j=1}^{N} s_{x \to j} \cdot E \frac{f_j}{\sum_{i=1}^{N} f_i} \\
&= \sum_{j=1}^{N} s_{x \to j} \cdot \sum_{k=1}^{\sum_{i=1}^{N} f_i} \frac{1/N}{\sum_{i=1}^{N} f_i} = \frac{1}{N} \sum s_{x \to j} = \frac{1}{2}
\end{aligned}$$

And if at any place L and time period T all people have equal chances to be there, i.e., $prob_y(L,T) = prob(L,T)$, then:

$$PSI_x(L,T) = \frac{\sum_{y \in U(L,T)} prob_y(L,T) \cdot s_{x \to y}}{\sum_{y \in U(L,T)} prob_y(L,T)} = \frac{\sum_{y=1}^{N} prob(L,T) \cdot s_{x \to y}}{\sum_{y=1}^{N} prob(L,T)} = \frac{1}{N} \sum s_{x \to j} = \frac{1}{2}$$

and then

$$PSI_x(T) = \sum_{L \in Loc_x} prob_x(L,T) \cdot PSI_x(L,T) = \frac{1}{2} \sum_{L \in Loc_x} prob_x(L,T) = \frac{1}{2}$$

and

$$PSI(L,T) = \frac{\sum_{\{x | L \in Loc_x\}} prob_x(L,T) \cdot PSI_x(L,T)}{\sum_{\{x | L \in Loc_x\}} prob_x(L,T)} = \frac{1}{2}$$

CSI calculated including individuals living in the same cell

In Fig. S.6 we present a distribution of *CSI* values calculated based on all users' communications without filtering calls between users living in the same cell. One can see that distributions presented in Fig. 2 and Fig. S.6 are very similar. A slight and expected difference is that the histogram is shifted to the right with mean equal to 0.576 (vs. 0.546 with filtering) while having the same standard deviation of 0.200. This confirms the expectation that a substantial proportion of all users communications consists of calls to the relatives and other people living together.

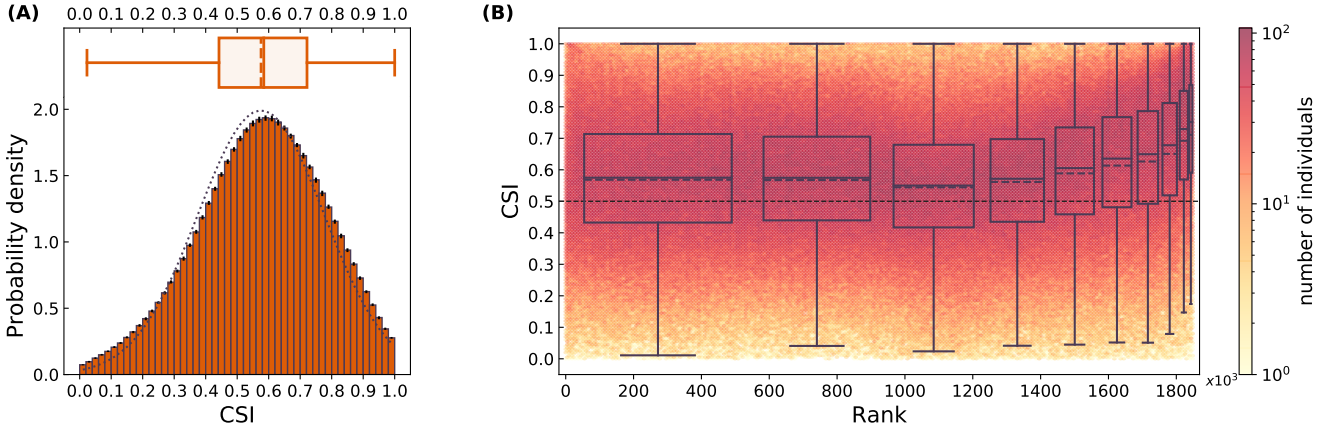

Figure S.6. Distribution of *CSI* values calculated without filtering calls between people living together.

CSI based on absolute rank difference

We understand that many readers might find our measure of social distance confusing at first. A more intuitive way of defining it could be just to take the absolute difference between people's ranks. In this case *CSI* could be defined by the same formula 2 from the main text, but using social distance defined as $d_{x\to j} = \frac{|r_x - r_j|}{N-1}$. This measure would totally make sense, but would lack a nice property of having the same base-line value for people from different classes. To see this, consider a person right in the middle of hierarchy, with rank $N/2$. For this person, her social distance to any other person range from 0 to $1/2$, i.e. social similarity range from $1/2$ to 1, and then average of these values will be $3/4$. While for the person with rank 1, her base-line *CSI* will stay 0.5.

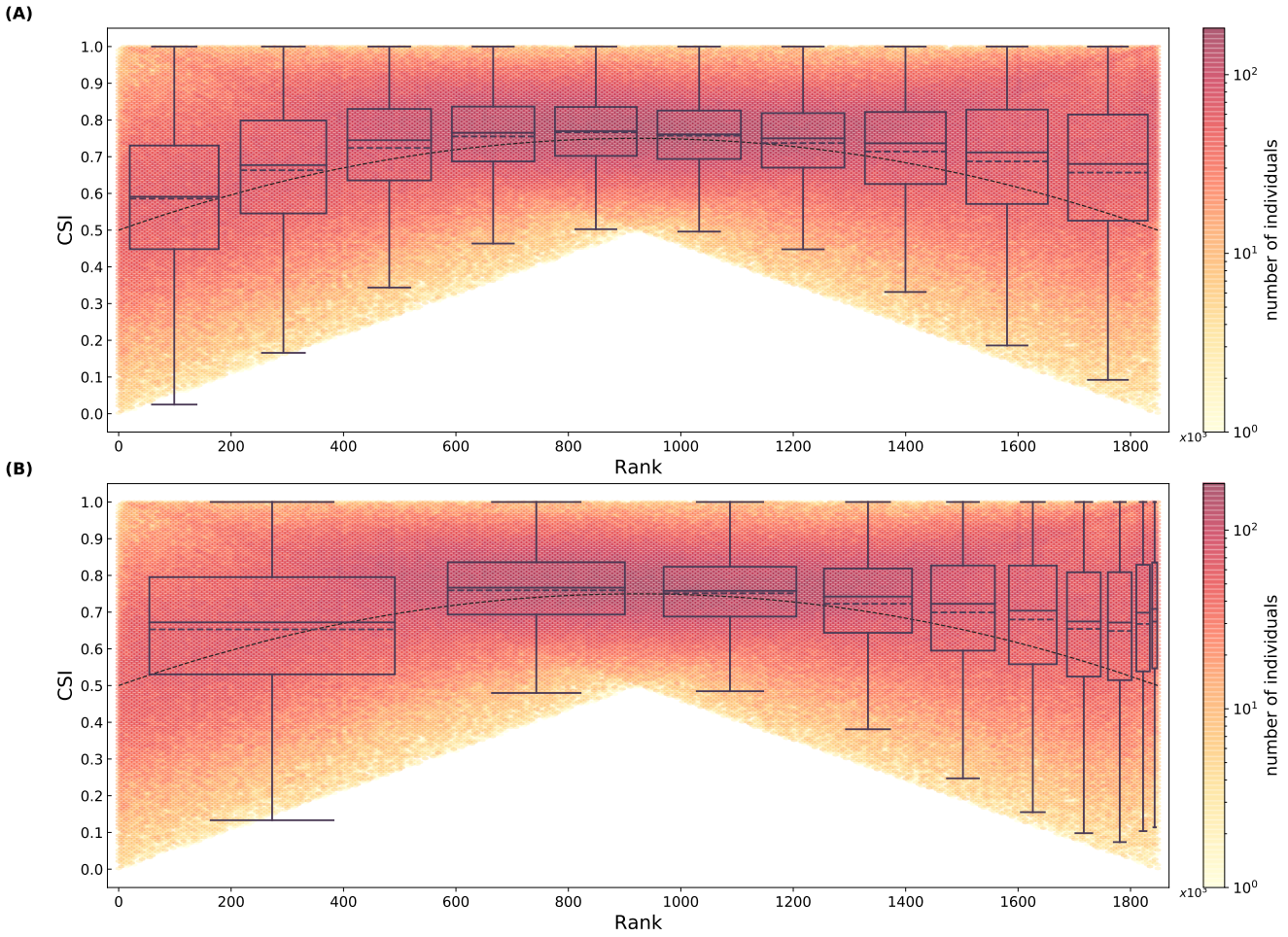

Figure S.7. Density of *CSI* based on absolute rank difference. We plot a line showing base-line *CSI* as well as box plots representing distribution of *CSI* values for **(A)** 10 equal-sized groups; **(B)** 10 groups with equal total housing price.

In Fig. S.7, we show density plot similar to one from Fig. 2B in the main text, but for this version of *CSI*. As it could be seen from the figure, just by raw *CSI* values it is hard to make any conclusions about the general level of social segregation and regarding different classes. To be able to make such conclusions one needs to compare these values with expected base-line values, outlining the importance of the social segregation metrics introduced in this paper.

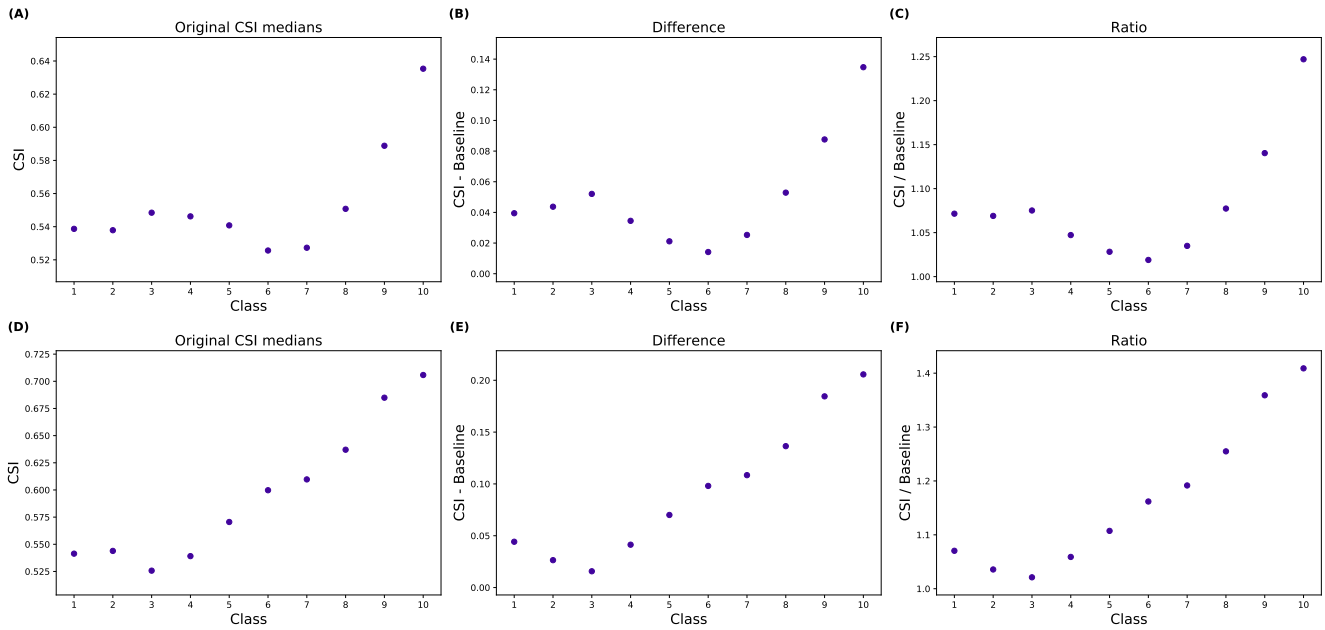

Figure S.8. Median *CSI* values of 10 classes. In the first row people split into equal-sized classes, in the second row – by equal cumulative housing price. **(A, D)** *CSI* measure proposed in this work; **(B, E)** Difference between median *CSI* based on absolute rank difference and median baseline; **(C, F)** Ratio between median *CSI* based on absolute rank difference and median baseline.

In Fig. S.8, we compare median values of original *CSI*, defined in the main text, with normalized *CSI* defined in this section. To normalize the values, we can either subtract base-line values from *CSI* values, or divide *CSI* values by base-line. We present both version in Fig. S.8. From this figure, one could see that distribution of values for both versions of *CSI* are fairly similar. This indicates that these values represent qualitatively the same thing – level of segregation of a group of people. But *CSI* measure proposed in the main text does not require any normalization, and its values could be easily interpreted.

Measure segregation of the city using Theil's entropy index

The Theil's entropy index is frequently used in previous studies to quantify the residential unevenness of a city[45]. It measures the departure of the entropy of each spatial unit (e.g., census tract) – which is determined, for example, by the ethic/racial composition at that place – from the racial or ethnic entropy of the whole city. In this research, we apply the Theil's entropy index to quantify the unevenness of interactions across different social classes (e.g., income groups). The index is calculated as follows:

$$H = \sum_{i=1}^{n} \frac{S_i(E - E_i)}{EN}$$

where N denotes the total number of phone users in the city; n refers to the total number of spatial units; S_i and E_i stand for the expected number of phone users and the corresponding Shannon entropy of spatial unit i, respectively. E is the overall Shannon entropy of the city:

$$E = \sum_{m=1}^{M} p_m \cdot log \frac{1}{p_m}$$

Here M denotes the total number of social classes that is predefined in the study, for example, income groups or classes derived from the housing price of phone users' residential locations. p_m stands for the proportion of class m users in the city. Similarly, E_i is calculated as:

$$E_i = \sum_{m=1}^{M} p_{i,m} \cdot log \frac{1}{p_{i,m}}$$

where $p_{i,m}$ denotes the proportion of class m users in spatial unit i.

Unlike traditional measure of residential segregation which associates individuals to fixed locations (home), in this research, we aim to quantify the interactions of phone users across different classes, and examine how the level of segregation in a city changes over time. To take human movements into account, this research starts by first dividing a day into several time windows (e.g., 24 one-hour time windows). For a given time window T, we can estimate, for each phone user, the probability of stay at different spatial units (i.e., grid cells). Thus, the value of $p_{i,m}$, for a specific time window T, can be calculated as the proportion of class m (i.e., C_m) users at location i:

$$p_{i,m} = \frac{1}{S_i} \sum_{x \in C_m} prob_x(i)$$

where $prob_x(i)$ denotes the stay probability of phone user x' at location i. Note that:

$$S_i = \sum_{m=1}^{M} \sum_{x \in C_m} prob_x(i)$$

The Theil entropy index ranges from zero to one. A value of zero indicates that all the spatial units have the same entropy that is equal to the value of the whole city. A value of one indicates that each spatial unit only hosts one particular class, which results into an entropy value of zero. There are several important considerations when the phone user pool is divided into different social classes. First, we need to determine the number of classes (M). Second, we need to specify the criterion for phone user classification. Here, we use two ways to divide phone users into M classes:

- *Quantile*: each class includes the same number of mobile phone users.
- *Total Buying Power*: each class has the same amount of buying power (e.g., total housing price). In this research, the total buying power for each class is calculated as the sum of housing price value of all the phone users in that class.

Regarding the number of classes M, we test different values and compare the results.

Here we present results of analysis similar to the one for *PSI* presented in the main text, again just for one random assignment of housing prices. By distinguishing weekdays and weekends, we divide each type of day into 24 one-hour time windows. We then calculate Theil's entropy index for each time window. Fig. S.9 shows the Theil's entropy indices of Singapore based on various combination of M and classification schemes (i.e., *Quantile* or *Total Buying Power*). We can see that different parameter sets produce very similar results on the overall level of segregation of the city as well as its diurnal patterns. During the day time, the city is more segregated on weekends than on weekdays.

Fig. S.10 shows values of 1.0 minus the normalized entropy value of grid cells for time windows *12AM – 1AM, 12PM – 1PM, 6PM – 7PM* on weekdays and weekends. Places with a high value of entropy (and correspondingly low value of 1 minus entropy) correspond to socially-mixed areas, while those with a low entropy (high bars) refer to those that are more segregated. Hence, the method can be used to quantify: (1) the overall level of segregation in a city, and (2) the spatial heterogeneity.

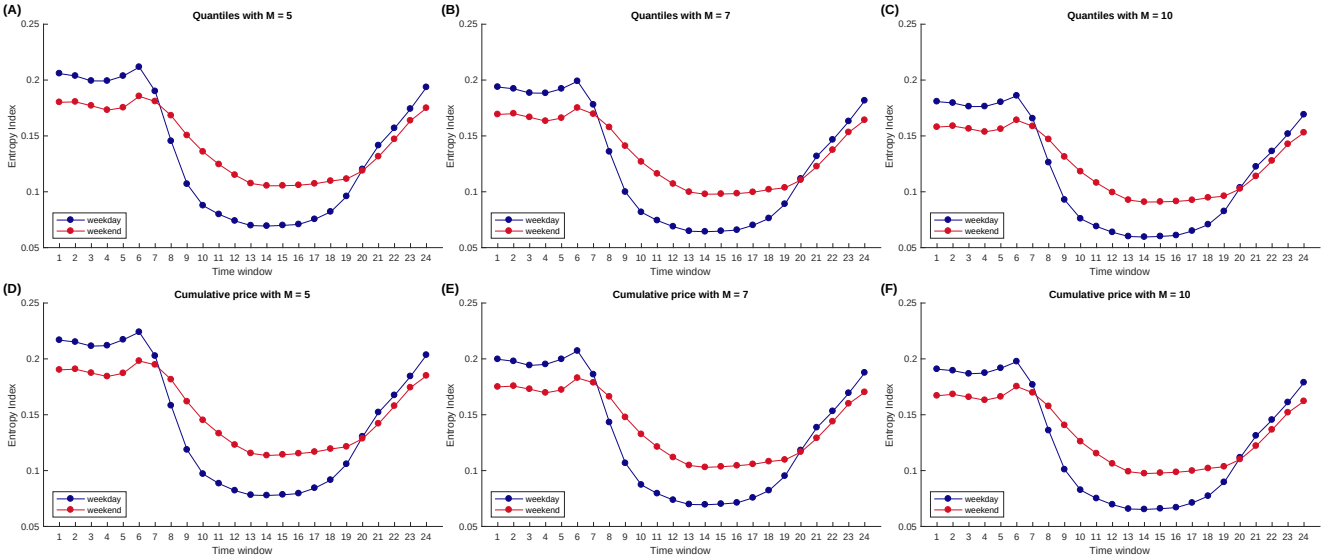

Figure S.9. Theil's entropy index and its temporal evolution over time. Results are generated based on: **(A)** Quantile classification with $M = 5$; **(B)** Quantile classification with $M = 7$; **(C)** Quantile classification with $M = 10$; **(D)** Buying Power classification with $M = 5$; **(E)** Buying Power classification with $M = 7$; **(F)** Buying Power classification with $M = 10$. The x-axis denote time windows, and y-axis denotes value of Theil's entropy index.

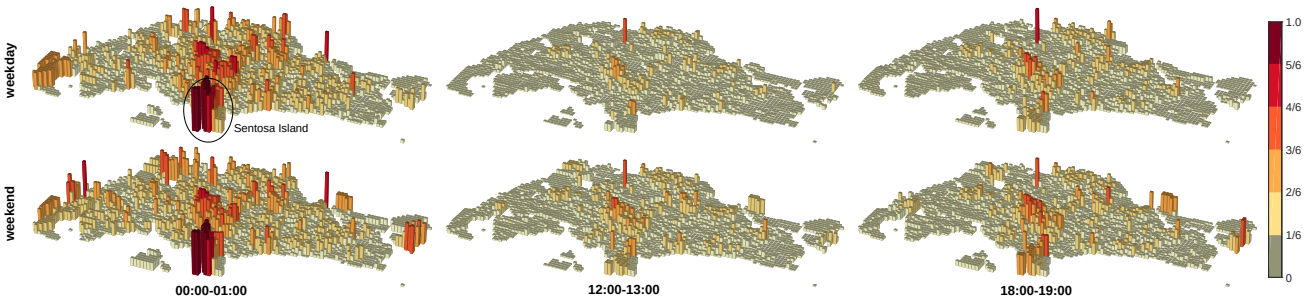

Figure S.10. $1 - \frac{E_i}{\log M}$ representing segregation of each grid cell at selected time windows (result based on quantile classification with $M = 10$)