JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 21 (2020), pp. 177–202
doi:10.5311/JOSIS.2020.21.587

RESEARCH ARTICLE

How does socio-economic and demographic dissimilarity determine physical and virtual segregation?

Michael Dorman[1], Tal Svoray[1,2], and Itai Kloog[1]

[1]Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2]Department of Psychology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Received: October 30, 2019; returned: February 2, 2020; revised: March 23, 2020; accepted: May 4, 2020.

Abstract: It is established that socio-economic and demographic dissimilarities between populations are determinants of spatial segregation. However, the understanding of how such dissimilarities translate into actual segregation is limited. We propose a novel network-analysis approach to comprehensively study the determinants of communicative and mobility-related spatial segregation, using geo-tagged Twitter data. We constructed weighted spatial networks representing tie strength between geographical areas, then modeled tie formation as a function of socio-economic and demographic dissimilarity between areas. Physical and virtual tie formation were affected by income, age, and race differences, although these effects were smaller by an order of magnitude than the geographical distance effect. Tie formation was more frequent when "destination" area had higher median income and lower median age. We hypothesise that physical tie formation is more "costly" than a virtual one, resulting in stronger segregation in the physical world. Economic and cultural motives may result in stronger segregation of relatively rich and young populations from their surroundings. Our methodology can help identify types of states that lead to spatial segregation, and thus guide planning decisions for reducing its adverse effects.

Keywords: Boston, edge weights, followers, income, mobility, network, spatial segregation, Twitter, US, virtual segregation

1 Introduction

There is, at present, an explosion of interest in social network analysis, primarily thanks to the advent of large online data sources on large social groups [5, 10, 47]. In particular, location-based social networks (LBSN) such as Twitter provide opportunities to study spatial dimensions of human behavior in greater detail than previously possible [32, 54, 77].

One of the notable domains in which network analysis of LBSN can bring substantial contribution is the study of spatial segregation in human society. Segregation can be thought of as the extent to which individuals of different groups occupy or experience different social-environments. A measure of segregation consists of a definition of the social environment of each individual, quantifying the extent to which these social-environments differ across individuals [73, 88]. Studies of segregation have three main analytical aims [72]: to investigate the patterns of segregation, to investigate the causes of segregation, and to investigate the consequences of segregation. The first aim leads towards the other two aims, as patterns suggest processes.

Spatial segregation is inherently geographical. Groups generally form distinct patterns of over- and under-representation across residential regions. The resulting urban mosaic is often described with terms that have figurative associations as well as spatial expression, such as ghetto, ethnic enclave, gated community, suburb, exurb, inner city, and edge city [11]. There is no single geographic scale of segregation and thus spatial segregation can exist at several levels simultaneously, ranging from specific households to neighborhoods to nation-states to the world [37]. Scales can affect how segregation is measured and represented [72]. For example, some features can be observed with a geographic scale of 1,000 or more kilometers (the concentration of the black population in the southeastern United States) and, in contrast, some features of racial residential patterns are observed at the smaller scales of states, metropolitan areas, municipalities, neighborhoods, city blocks, and even households.

Many methods of measuring segregation, and particularly spatial segregation, have been formulated and proposed [25, 31, 88]. The most commonly used measure is the Index of Dissimilarity (ID) [18]. The ID quantifies the evenness with which two demographic groups are distributed among areal units comprising a geographical area. It can be interpreted as a measure of displacement, quantifying the percentage of one of the two groups that would have to move in order to produce an even distribution. Massey & Denton [58] further identified five dimensions of segregation: unevenness, exposure, clustering, concentration, and centralization. Among them, unevenness and clustering are regarded as the most important [66, 73]. Unevenness is a measure of spatial heterogeneity, the variation displayed across the map. Clustering measures the scale of spatial similarity, the extent to which closely located neighborhoods are alike. For instance, in a "checkerboard", the standard pattern of black white alternation, can be compared with a hypothetical board for which the top half is wholly black and the bottom wholly white.

Studies of spatial segregation are mostly focused on measuring population distribution patterns in residential space, based on census data [19, 58, 81], ignoring the fact that social isolation likely extends from residential place to other locations and to other dimensions of activity. Recently, it has been recognized that segregation studies should go beyond residential place to daily activity space [2, 22, 50, 70], and shift from location-based to people-based analysis [46]. A key paper by Krivo et al. [45] argues that studies which consider residential neighborhoods as the only context of social isolation ignore the fact that social

isolation likely extends from residential place to other geographical locations (e.g., workplace) and to other dimensions of activity (e.g., face-to-face encounters during daily travel). For example, although residential place remains an important hub in individuals' daily life, the importance of other places (employment, recreation, etc.) has increased with the growth of human mobility in urban areas [48, 87]. Thus, a fuller understanding of urban segregation requires critical analyses of not only the socio-demographic compositions of residential neighborhoods, but also the types of social-environments that individuals are exposed to in daily life.

The degree of mobility and communication between areas of contrasting socio-economic background are important aspects in the formation and maintenance of spatial segregation [68]. In reference to LBSN, these two complementary components [15, 62, 79] are thought to be the strength of *physical / mobility* ties and of *friendship / virtual* ties, respectively. Virtual ties refer to the social network structure of the LBSN. For example, Twitter users declare the people they are interested in "following", in which case they get notified when that person has posted a new message. A user who is being followed by another user does not necessarily have to reciprocate by following them back, which makes the ties on the Twitter social network directed [23]. The structure of these virtual ties shapes information flow within the network [12, 27], whereas lack of information flow can be considered an expression of segregation. Virtual ties can be contrasted with physical ties, the latter referring to physical movement of individuals between two regions in geographical space [9, 26], rather than the flow of information in virtual space. Namely, the occasion in which a person who lives, or spends most of his day, in one place was present at least one time in another site is informative regarding segregation and urban interaction [76, 82]. When referring to individuals, virtual and physical ties can be either present or not (e.g., the user does or does not follow another user). When referring to geographical areas, tie strength can naturally be defined as the proportion of users maintaining the given relation, relatively to maximal possible connectivity (e.g., the proportion of follower ties between users in area A and users in area B, relatively to the maximal number of ties obtained if all users followed everyone else) [49].

Virtual and physical ties comprise two complementary descriptors of daily activity patterns. For example, the emergence of "rich" enclaves may lead to fragmentation of public space in terms of physical mobility, where the underprivileged are increasingly cut-off. Limited communication between contrasting neighborhoods, likewise, may lead to enhanced prejudice and reduced solidarity, thus standing in the way of political solutions to common problems. Understanding the formation of these intangible mobility and communication barriers [73] between sub-populations of contrasting socio-economic background is a first step towards reducing their negative outcomes [85].

Most studies on physical or virtual ties on Twitter aimed at inferring spatial community structure and interactively exploring it [4, 20, 36, 78, 90]. This approach is particularly useful for detecting "natural" geographical boundaries [32] reflecting human interaction patterns, as opposed to predefined boundaries based on administrative division. It is less useful, however, for studying the effects of socio-economic factors on the formation of such ties, as population characteristics contributing to community cohesion are not quantitatively considered.

Fewer LBSN-based studies attempted to quantify the determinants of tie formation itself, whether on the level of individual Twitter users [17], populations, or geographical areas, and whether concerning mobility [26, 30], communication [12, 15], or friendship [79, 80]

ties. Most studies, however, only considered elementary predictors of tie formation, such as geographical distance alone [49,51,52,79] (but see [3,62]).

Studies of population-level socio-economic and demographic (SD) predictors of tie formation have either leaned towards an exploratory analysis of spatial patterns in specific case-studies [30,76], or have over-simplified the representation of physical space [17]. For example, Shelton et al. [76] used Twitter data to examine spatial segregation in terms of mobility between two contrasting socio-economic regions: West-End and East-End Louisville, Kentucky. Based on inferred movement patterns between the two regions, the authors argued against the commonly held view of a rigid bi-directional segregation (the "9th Street Divide") between these two areas. Later on, Huang and Wong [30] studied the movement activity of Twitter users among four distinct socio-economic regions in Washington, D.C., rather than a single pair of regions. Yet, the considered variables were group characteristics (e.g., travel distance), rather than ties forming a complete network (e.g., mobility flow between all region pairs).

Recently, Ma et al. [54] spatially analyzed a large-scale friendship tie network between users in the Brightkite and Gowalla LBSNs. Aggregating social ties between users in each network, the authors created both location-location and city-city spatial networks reflecting tie counts across the entire area of the continental US. The authors put emphasis on spatial characteristics such as the relative abundance of network ties between large cities, thus demonstrating that "the number of social connections does not correlate well with geographic proximity, but depends on the characteristics of a place" [54]. Wang et al. [82] have analyzed small-scale segregation patterns, evaluating mobility characteristics as function of racial and socio-economic composition at the neighborhood level, based on Twitter data. Again, the study revealed otherwise unapparent patterns of racial segregation through mobility: where the different groups travel and whom they are exposed to, extending the work of Shelton et al. [76] from a sample of two regions in Louisville to a sample of $\sim$36,000 block groups in the 50 largest cities in the US [82].

The purpose of the present study is to take the next step and evaluate the "characteristics of place" [54] in terms of their effect on follower and mobility ties—in a systematic, comprehensive and quantitative way. Accounting for geographical distance and population size, we focus on the less obvious effects of SD characteristics, in light of spatial segregation as reflected through the recorded behavior of Twitter users.

A more comprehensive understanding of network tie formation determinants can shed new light on patterns of spatial segregation in the various activity dimensions of human society—which is commonly recognized as the next frontier in studying segregation [87]. Generalizing the above studies [30, 54, 76, 82], we propose a comprehensive hypothesis-testing oriented methodology, that operates on a large sample of regions (rather than few individual regions) covering a wide and heterogeneous spatial extent, using two tie metrics that represent both mobility and friendship (rather than just one), while bearing in mind SD "characteristics of place" (income, age, and race). Using this methodology, our aim here is to study how dissimilarity in population characteristics translates into segregation—in terms of mobility and friendship—between the geographical areas these populations occupy.

The study aim is achieved using three operational *objectives*:

1. Constructing four weighted networks that represent physical and virtual tie strength on two distinct spatial scales, based on geo-referenced Twitter data, and a fifth survey-based commute network for validation.

2. Fitting models where tie strength between given areas in the latter five networks is explained with their distance, SD dissimilarity, and the SD $\times$ distance interactions.
3. Using a model selection procedure to determine which factors substantially affect tie formation, their effect size and effect direction, concerning each tie type (physical and virtual) on each spatial scale.

Our specific *hypotheses* are:

1. Spatial segregation exists in both physical and virtual dimensions—although it may be weaker in the virtual dimension, due to lower tie formation costs.
2. Racial dissimilarity enhances spatial segregation, due to the homophily principle [59].
3. Income and age differences induce asymmetric segregation due to the unbalanced motivation that dissimilar populations, such as the rich and poor, have for maintaining contact with each other [86].

2 Methods

2.1 Twitter Data

Access to Twitter data is provided through several Application Program Interfaces (APIs). The Streaming API was used to continuously collect text contents and metadata of all geo-referenced tweets falling within the study areas. The REST API was subsequently used to collect the list of users each user follows (otherwise known as their "friends"). The analyzed dataset was thus comprised of point locations (lon-lat) each unique user posted from, combined with an indication on whether a social tie exists between each pair of users.

Though geo-tagged tweets comprise only 1-2% of the total volume [53], thanks to the relatively prolonged collection period we could accumulate a large sample size of over 20M tweets sent by $\sim$900K unique users (Table 1), that can serve as a good approximation of human activities for study purposes [54]. It should be noted that Twitter data are more strongly associated with sampling bias (Twitter users vs. general population) and location bias (messages sent from particular types of locations vs. continuous monitoring), compared to mobile phone records. On the other hand, geolocated Twitter data are spatially accurate at GPS precision, while phone record data resolution is dictated by distances between cellular towers, which are typically in the order of several kilometers [21]. Overall, Twitter data have been shown to be representative of population-level behavioral patterns [35] and have been successfully used in numerous studies on formation of virtual [36, 79] and physical [30, 76, 82, 90] ties.

	GBA	US
Period start	2016-03-29	2016-05-26
Period end	2017-02-12	2016-10-05
Period length	320 days	132 days
Bounding box area	49,805 km^2	13,094,663 km^2
Geo-located tweets	1,855,513	21,896,420
Unique users	73,563	876,764

Table 1: Description of Twitter data for the Greater Boston Area (GBA) and the US.

The data collection process was repeated on two spatial scales:

1. County scale, the contiguous USA (US) (∼8,000,000 km^2)
2. Census tract scale, a rectangular area of ∼50,000 km^2 in the Greater Boston Area (GBA)

The specific study areas (US and GBA) were selected for two reasons. First, Twitter usage is particularly high in the US [26, 80], and particularly in the GBA which is one of the major educational and Information Technology hubs in the US. This assures a large sample size for estimating and modeling follower and mobility ties between geographical units. Second, the GBA region was extensively used in many of our group's previous works [38–42].

Although the GBA is contained within the US, a separate collection process was conducted to achieve a more detailed sampling of tweets, given API rate limit considerations. The sample size and time frame of Twitter data used in this study are specified in Table 1.

2.2 Socio-economic and demographic (SD) data

The US Census is arguably the most important data set for social science research in the United States [1]. The census geography maintains a strict hierarchy where states contain counties, counties contain census tracts, census tracts contain block groups, and block groups contain blocks. We worked on two spatial scales: county and census tract. To this end, we used the American Community Survey (ACS) 5-Year Estimates (2010-2014) data for obtaining SD data for the studied areas at each spatial scale. Three key characteristics [65] were extracted: median household income (ACS code: `B19013e1`), median age (`B01002e1`) and total population in each racial/ethnic group (`B02001e2`, `B02001e3`, `B02001e4`, `B02001e5`, `B02001e6`, `B02001e7`) (see Table S1 and Figures S1–S6 in the Supplemental Materials).

Dissimilarity between each pairs of areas was expressed as Euclidean distance, either one-dimensional (median income, median age) or multi-dimensional (racial composition). For example, in case the destination area has median income of $50,000 and the origin area has median income of $60,000, the one-dimensional dissimilarity was defined as -$10,000. Note that the one-dimensional dissimilarity measure is directional, thus expressing not just the absolute difference but also whether the destination has higher or lower value of the metric. For racial composition, multi-dimensional dissimilarity was calculated using function `dist` in R, using `method="euclidean"`, following transformation of racial/ethnic population from counts to proportions to remove the effect of total population size. Note that multi-dimensional dissimilarity is not directional, as it only expresses the absolute degree of compositional difference between the origin and destination areas. Also note that our dissimilarity measure, simply reflecting multivariate Euclidean distance, is not to be confused with the more specific Index of Dissimilarity (see Section 1) which refers to compositional difference between areas and frequently used in segregation studies [18].

2.3 Commute data

To validate the results obtained using Twitter, we reproduced the analysis of socio-economic effects on mobility realization using an external data source—the 2009-2013 5-Year ACS Commuting Flows dataset[1]. This dataset contains counts of workers in com-

[1]https://www.census.gov/data/tables/time-series/demo/commuting/commuting-flows.html

muting flow between each pair of counties in the USA [64]. To standardize the number of commuters by potential commuters count, we also obtained the county-level labor force estimates (`B23025e4`) from the ACS dataset (Table S1).

2.4 Network construction

Twitter network data were aggregated from individual-user scale to areal scale, as estimates of SD characteristics are only available on the latter. Aggregation involved the following steps:

1. Assigning each Twitter user to the areal unit where he/she was most active, which we marked as his/her "center of activity". The center of activity was defined as the areal unit with the greatest number of tweets for a given user [54,82]. One of the reasons for using this broad definition, rather than attempting to detect place of residence [29,30], is that Twitter accounts for organizations, agencies, services, etc., rather than authentic individuals, are increasingly more common [75]. The term "place of residence" is naturally irrelevant for such users. Nonetheless, these users are still relevant in terms of virtual and physical ties within the online community on Twitter, as they reflect the flow of information and levels of mutual interest between different spatial areas. For example, an organization account may establish virtual follower relations with other organizations and individuals, or post tweets from different physical locations, thus contributing to the segregation or lack thereof between geographical areas in the same way that authentic users do.
2. Calculating virtual and physical tie strength metrics between all possible pairs of areal units A and B in the study area, according to the following algorithms for each tie type:
   (a) Follower = The number of follower ties between a user from area A and a user in area B, divided by the number of potential ties—i.e., the number of unique users from area A multiplied by the number of unique users from area B (Figure 1).
   (b) Mobility = The number of users from area A who have sent at least one tweet when physically located in area B, divided by the number of potential ties—i.e., the number of unique users in area A.
   (c) Commute = The commuting flow count from area A to area B, divided by total labor force in area A.
3. Assigning each pair of areas A and B with the geographical distance between their centroids, as well as the corresponding set of SD dissimilarity metrics (see Section 2.2).
4. Repeating steps 1-3 for the two spatial scales—namely, for the census tract scale in the GBA (Figure 2) and for the county scale in the US (Figure S7).

It is important to note that the tie strength indices are directional—i.e., tie strength for $A \rightarrow B$ is not necessarily the same as tie strength for $B \rightarrow A$. Self-ties, where the origin and destination are the same (i.e., $A \rightarrow A$), were excluded from the analysis. Also note that the tie strength indices are inherently standardized by total network activity for removing group size bias [49]. For example, areas that are more densely populated or characterized by younger population [60] are expected to have more Twitter users and thus more LBSN

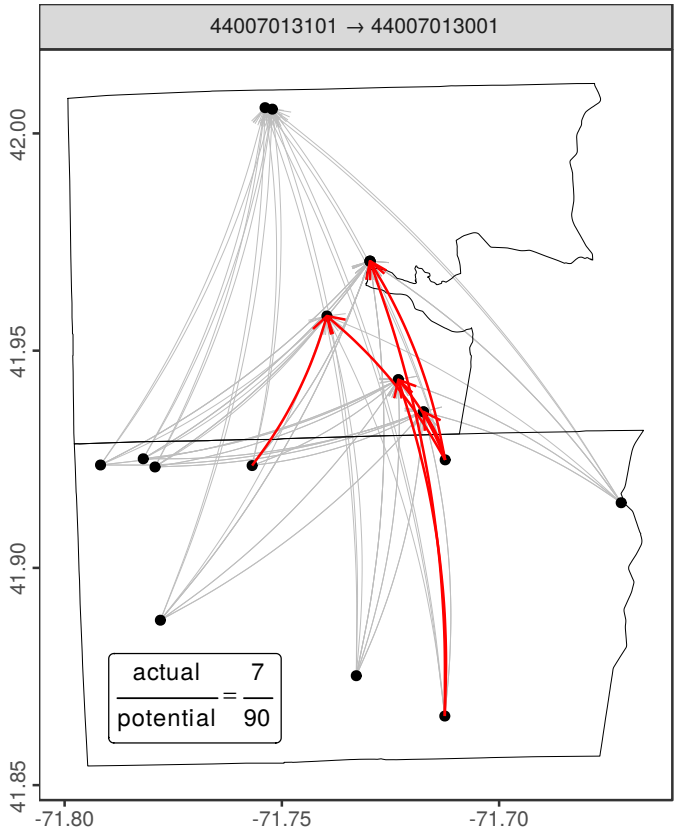

Figure 1: Calculation of the follower tie ratio between two census tracts in the Greater Boston Area (GBA) (44007013101 and 44007013001). Grey segments represent all 90 possible follower ties extending from Twitter users whose estimated "center of activity" is located in tract 44007013101 towards users in whose "center of activity" is in tract 44007013001. Red segments represent the 7 ties that are actually realized. Follower tie ratio for the 44007013101 $\rightarrow$ 44007013001 edge is therefore equal to $7/90$.

activity, whether physical or virtual. Standardizing by total activity removes this effect and allows us to compare areas with varying group sizes.

Descriptive statistics of the five analyzed networks are given in Table 2. An illustration of calculating a single follower tie strength estimate ($A \rightarrow B$) between two census tracts in the GBA is given in Figure 1. Illustrations of the complete networks, built once estimates are available for all area pairs, are given in Figure 2 and Figure S7. Note that to maintain visual clarity, the latter figures do not display the entire network, but only sub-networks including adjacent areas—i.e., ties between all areas A and B which share a common border. The actual analyzed networks (Tables 2–3) consist of all possible ties between any pair of areas.

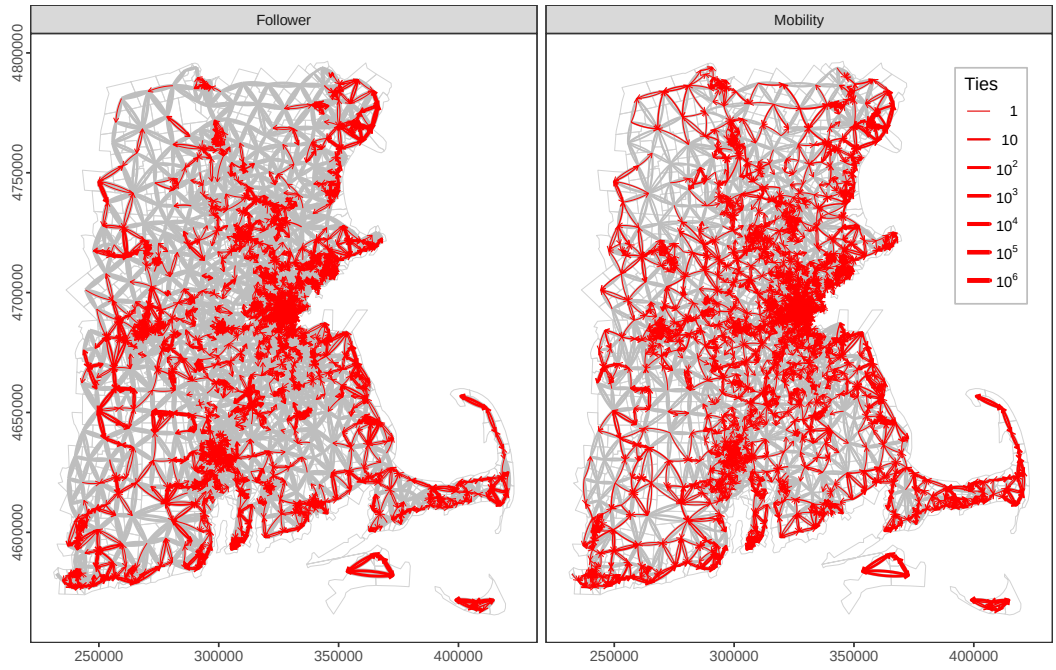

Figure 2: Observed follower and mobility ties between adjacent census tracts in the Greater Boston Area (GBA). Grey lines represent potential ties count, red lines represent actual count. Note that for visual clarity, these figures do not display the entire networks, but only sub-networks of adjacent areas—i.e., ties between all areas A and B which share a common border. Models (Table 3, Figure 3) were fitted to data on all tie pairs, not just the adjacent ones. Also note that line width is on a logarithmic scale.

	GBA		US		
	Follower	Mobility	Follower	Mobility	Commute
Vertices	1,760			3,108	
Edges	3,097,600			9,659,664	
Non-zero ties	7.84%	2.85%	10.40%	2.40%	1.37%

Table 2: Description of weighted directed networks representing virtual (i.e., follower) and physical (i.e., mobility) ties between predefined geographical areas (census tracts and counties, respectively) in the Greater Boston Area (GBA) and the US.

2.5 Statistical analysis

A preliminary visual evaluation of geographical distance effect on network tie strength was conducted by fitting a Generalized Additive Model (GAM) to each of the four Twitter-based networks (two spatial scales $\times$ two tie types) (Figure 3). The dependent variable was the tie strength estimate (Figure 1), while the independent variable was geographical

distance between the respective areas. The observations comprised all network edges—i.e., all ties between pairs of areas.

Displaying the relation between tie realization and distance on a log-log scale, followed by fitting a power law function, has been shown to facilitate comparison of distance decay intensity among different scales [21]. To fit the power law function, network ties in the range of 0-200 km and 0-700 km in the GBA and US scales, respectively, were aggregated into 20 equal breaks. A linear regression model was then fitted to the average tie realization per distance bin as function of its midpoint distance (Figure 4). The coefficient β of the linear fit on the log-log scale:

$$\log_{10} Y = \alpha - \beta \times \log_{10} d$$

where Y is tie realization and d is geographical distance, thus reflects the power law coefficient β on the original scale:

$$Y = 10^{\alpha} \times d^{-\beta}$$

The main statistical approach follows section 9.2 "Modeling Network Flows: Gravity Models" in Kolaczyk and Csárdi [43], pp. 162-170. The input data for the statistical analysis comprised the five weighed networks: Twitter-derived networks for two spatial scales $\times$ two tie types, and the US commute network. In the main analysis, we considered not only distance, but also SD dissimilarity, and the interactions of SD dissimilarity with geographical distance. In each case, we statistically tested whether realization of follower or mobility ties was associated with the latter variables, and if so—how. For example, tie formation probability may be higher between more proximate areas (negative geographical distance effect), between areas characterized by higher racial composition similarity (negative racial composition dissimilarity effect) and either effect may vary when the other does (distance $\times$ racial composition dissimilarity interaction).

Generalized Linear Models (GLMs) with binomial response (i.e., "logistic regression") were used since the dependent variables consisted of proportional data. Thus, the dependent variables were "success" vs. "failure" counts—i.e., the ratio between the number of actual and potential network ties (Figure 1). The independent variables were: geographical distance, median income arithmetic difference, median age arithmetic difference, and racial composition multivariate euclidean distance, as well as the interaction of geographical distance with these three SD variables.

We found no multicollinearity among the four examined variables on the network edges—i.e., geographical distance and dissimilarity of income, age, and racial composition. For example, the strongest Pearson's correlation in the GBA dataset was 0.37 (between age and income dissimilarities), which is considered "Low" [63]. Although spatial autocorrelation between network edges is not clearly defined, and pairwise neighbor weighting is computationally unfeasible for sample sizes of 3M or 9M edges (Table 2), we ran preliminary evaluations on a sample of randomly chosen 10,000 edges in the GBA. We used Moran's I global test for autocorrelation [8], with the 8-nearest-neighbors edge centroid criterion for defining neighbor weights. There was no significant autocorrelation in the follower network (p-value = 0.53) or the mobility network (p-value = 0.21).

A model selection procedure—based on the Akaike Information Criterion (AIC)—was conducted to evaluate the relative support for the full model and all simplified models lacking one or more of the predictors, in each of the five networks (Table 3). In each case, models were ordered by the AIC score—from lowest AIC (i.e., highest relative support) to highest

AIC (i.e., lowest relative support). The hypotheses underlying the inclusion of variables present in the most parsimonious models (i.e., having the lowest AIC) were considered supported by the data [34]. The five most parsimonious models were eventually used to generate and visualize predicted tie strength in the studied parameter space (Figures 5–7), to characterize effect sizes and directions. We also calculated Akaike weights (AIC_w), which express relative weight of evidence for each model, summing to 1 across all models (Table 3). An AIC_w value for model i can be interpreted as the probability that model i is the best model for the observed data, given the candidate set of models [34]. Finally, we calculated explained deviance (pseudo-R^2) per model. Explained deviance in a GLM is analogous to R^2 in a Linear Model, expressing the proportion of variation explained by the model [92].

Model predictions were calculated using non-standardized GLM coefficients, thus allowing for interpretation in the original units for each variables (e.g., kilometers for distance, or $ for income difference) (Figures 5–7). In addition, Table 3 reports the standardized model coefficients. Standardized regression coefficients basically refer to how many standard deviations a dependent variable will change per standard deviation increase in the predictor variable. Standardized coefficients are thus useful when numerically comparing effect sizes among variables in the same model, or among different models based on the same data. For example, we can compare the effects of independent variables (e.g., distance) among the three models (mobility, follower, and commute) within each given study region. The comparison is valid due to identity of the underlying data: an effect of +1 for the standardized distance coefficient reflects the same effect size for all three models (in terms of standard deviations of the response variable), since the distributions of edge distances—and therefore their standard deviations—are the same in the three datasets.

Following [82], we re-fitted all Twitter-based GLM models weighting each observation (i.e., network tie) based on the ratio of Twitter users to the total population size. The weighting mechanism thus (partially) addresses the fact that Twitter users are not fully representative of the local population. The weighted model results are given in Table S2. Since the weighted and unweighted models were highly consistent (Tables 3 and S2), predictions are only reported for the unweighted models (Figures 5–6).

2.6 Software

Accessing the Twitter APIs for data collection was done using Python [74] package `twarc`[2]. All other analyses were done in R [71]. Spatial processing of the Twitter and census data were executed using R packages `sp` [8] and `rgeos` [7]. Network construction and statistical calculations were done using package `igraph` [16]. GAMs were fitted using package `mgcv` [89]. Moran's I test was done using package `spdep` [8]. Model selection procedure of GLMs was done using package `MuMIn` [6]. Figures were produced with package `ggplot2` [84].

3 Results

The characteristics of the five networks representing follower, mobility and commute tie strength between areal units on two spatial scales (see Section 2) are provided in Table 2.

[2]https://github.com/docnow/twarc

Network density—i.e., the proportion of non-zero ties—was higher in the follower networks (7.84% and 11.0%, in the GBA and the US, respectively) than in the mobility networks (2.85% and 3.12%). In other words, a higher proportion of area pairs were characterized by at least one follower tie, than by at least one physical observation of a user who is a resident of one area "visiting" the other area. The commute network density was lower still (1.37%), indicating that regular work-related commute takes place between a small subset of (adjacent) county pairs out of all possible county pairs in the US.

In the preliminary analysis, the relations between geographical distance and tie realization proportion were visually examined by fitting GAMs to tie (i.e., edge) properties of the four Twitter-based networks. The most obvious observation (Figure 3) was that follower ties and mobility ties markedly differ in their form of distance decay. Follower ties (1) exhibited a relatively shallow decline with increasing distance and (2) never reached "zero" realization. Conversely, mobility ties (1) exhibited a relatively steep decline and (2) quickly reached "zero" realization when distances get large (i.e., several hundred kilometers).

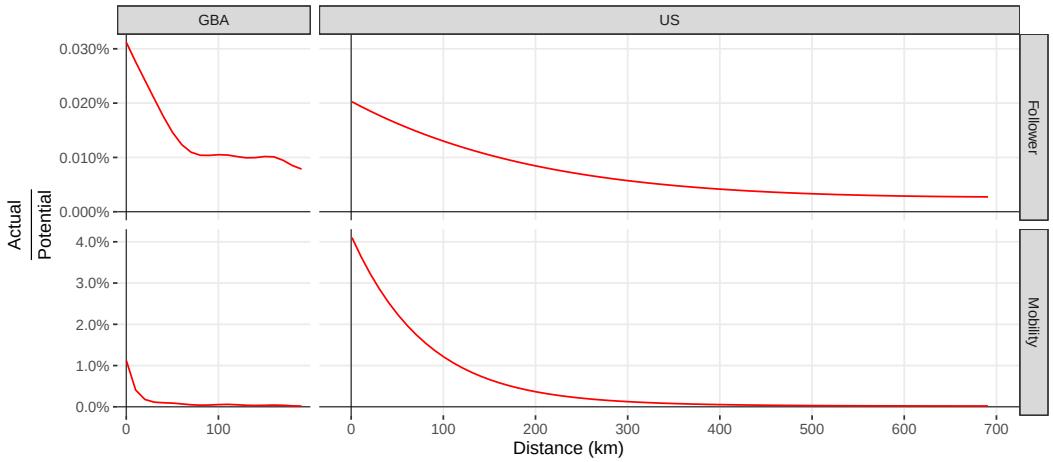

Figure 3: Follower and mobility tie proportions as function of geographical distance in the Greater Boston Area (GBA) and the US. Lines show the average trend based on a Generalized Additive Model (GAM). The x-axis range covers 99.7% and 21.7% of observed data in the GBA and US areas, respectively.

Steepness of distance decay was also quantitatively assessed by fitting a linear model on a log-log scale and extracting the slope β (Section 2.5). Distance decay was by far steepest when considering commute ($\beta = 2.06$), compared to mobility ($\beta = 1.17$) and follower relations ($\beta = 0.90$) in the US (Figure 4). Comparing the Twitter-based metrics only, mobility distance decay ($\beta = 1.00$ and $\beta = 1.17$ in the GBA and US, respectively) was more steep than that of follower ties ($\beta = 0.79$ and $\beta = 0.90$).

In the main analysis, according to the AIC-based model selection procedure, the full models (i.e., those including all examined factors) had overwhelmingly highest relative support in 4 out of 5 cases (Table 3). Only in the case of the mobility tie models in the GBA did the most parsimonious model lack the "income $\times$ distance" effect—although in that case the full model came second, not far behind in terms of relative support (AIC_w of 0.31

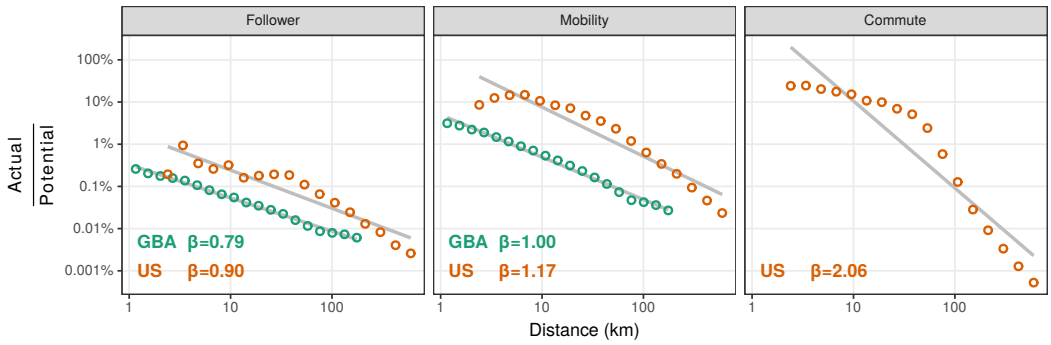

Figure 4: Follower, mobility and commute tie proportions as function of geographical distance in the Greater Boston Area (GBA) and the US, on a log-log scale. Linear regression fitted lines are shown in grey. The values of β refer to the linear slope, reflecting the decay parameter. The x-axis range covers 99.7% and 21.7% of observed data in the GBA and US areas, respectively.

vs. 0.69). In other words, the hypothesis that geographical distance, SD dissimilarity, and their interactions (except for the "income $\times$ distance" interaction for mobility ties in the GBA) affect follower and mobility tie formation was supported by the data.

In agreement with the preliminary visual examination (Figure 3), the effect of geographical distance on tie formation was (1) consistently negative and (2) larger for mobility ties than for follower ties. In other words, follower and mobility tie formation probability was reduced with increasing geographical distance between given areas, more steeply when considering mobility. In terms of effect size, observing best models' predicted values in the relevant parameter space (Figures 5–7) as well as standardized coefficients (Table 3) revealed that distance effect on mobility tie formation was stronger by an order of magnitude compared with follower tie formation. For example, predicted decline of follower ties realization between short distance of $\sim$300 m (5% quantile) and long distances $\sim$3000 m (95% quantile) in the US was 2-fold, from 0.00455% to 0.00243% (Figure 5). Under the same scenario, the predicted decline of mobility ties realization was 61-fold, from 0.194% realized ties on $\sim$300 m distances to 0.003% realized ties on $\sim$3000 m (Figure 6). The effect of distance on commute frequency was higher still—the x-axis for commute predictions (Figure 7) does not show the full range of distances in the US but only distances up to 200 km, as commute realization above that distance was practically zero.

The effects of median income and median age were also largely consistent among examined tie types and scales. Income effect was positive in all cases (Table 3, Figures 5–7), with no substantial "income $\times$ distance" interaction effect size in the studied parameter space. Namely, both follower and mobility tie realization constantly increased when the destination area had a relatively higher income compared with the origin area. For example, predicted tie realization at short distances of $\sim$300 km (5% quantile) in the US increases from 0.0046% to 0.0049% (1.1-fold) and from 0.194% to 0.366% (1.9-fold) considering follower and mobility, respectively, when average median income difference increases from

Area	Type	$Inter.$	$Dist.$	$Inc.$	Age	$Race$	$I. \times D.$	$A. \times D.$	$R. \times D.$	AIC_w	pR^2
GBA	Follower	**-8.863**	**-0.409**	**0.027**	**-0.021**	**-0.076**	**0.005**	**0.018**	**-0.082**	**0.996**	0.09
		-8.863	-0.409	0.025	-0.021	-0.076		0.019	-0.082	0.004	0.09
		-8.863	-0.410	0.027	-0.023	-0.076	0.013		-0.082	<0.001	0.09
		-8.863	-0.409		-0.015	-0.076		0.015	-0.082	<0.001	0.09
		-8.863	-0.410	0.019	-0.024	-0.076			-0.082	<0.001	0.09
		-8.863	-0.409		-0.019	-0.076			-0.082	<0.001	0.09
	Mobility	**-7.259**	**-1.183**	**0.104**	**-0.257**	**-0.308**		**0.056**	**0.032**	**0.690**	0.17
		-7.259	-1.183	0.102	-0.257	-0.309	-0.003	0.057	0.032	0.310	0.17
		-7.267	-1.191	0.099	-0.257	-0.339	-0.006	0.058		<0.001	0.17
		-7.267	-1.191	0.104	-0.257	-0.339		0.056		<0.001	0.17
		-7.269	-1.191	0.111	-0.298	-0.307	0.016		0.036	<0.001	0.17
		-7.269	-1.191	0.096	-0.299	-0.309			0.033	<0.001	0.17
US	Follower	**-10.243**	**-0.188**	**0.056**	**-0.049**	**-0.033**	**0.010**	**-0.013**	**0.100**	**1**	0.07
		-10.244	-0.188	0.056	-0.053	-0.033	0.010		0.100	<0.001	0.07
		-10.244	-0.187	0.060	-0.049	-0.033		-0.011	0.100	<0.001	0.07
		-10.244	-0.187	0.060	-0.053	-0.033			0.100	<0.001	0.07
		-10.243	-0.188	0.054		-0.033	0.009		0.099	<0.001	0.07
		-10.243	-0.187	0.057		-0.033			0.100	<0.001	0.07
	Mobility	**-7.857**	**-1.235**	**0.454**	**-0.396**	**0.068**	**0.072**	**0.005**	**0.386**	**0.998**	0.23
		-7.857	-1.234	0.454	-0.399	0.068	0.072		0.386	0.002	0.23
		-7.861	-1.237	0.407	-0.395	0.059		0.003	0.377	<0.001	0.23
		-7.861	-1.237	0.407	-0.397	0.059			0.377	<0.001	0.23
		-7.772	-1.165	0.436	-0.391	-0.124	0.057	0.037		<0.001	0.22
		-7.770	-1.160	0.435	-0.414	-0.125	0.054			<0.001	0.22
	Commute	**-36.912**	**-20.083**	**3.004**	**0.288**	**4.053**	**1.769**	**0.288**	**2.513**	**1**	0.74
		-36.982	-20.127	3.022		4.066	1.782		2.520	<0.001	0.74
		-36.897	-20.081	3.242	0.445	0.058	1.917	0.445		<0.001	0.74
		-37.003	-20.147	3.242		0.058	1.919			<0.001	0.74
		-36.925	-20.092	3.276	0.461		1.937	0.461		<0.001	0.74
		-37.034	-20.16	3.274			1.938			<0.001	0.74

Table 3: Model selection results for Generalized Linear Models (GLMs) of follower, mobility and commute tie probability in the Greater Boston Area (GBA) and the US, as function of geographical distance, socio-economic and demographic (SD) dissimilarity, and interactions. The six most highly supported models are shown per model selection procedure. Models are ordered by decreasing AIC, starting from the most supported model (in bold). The AIC_w column shows Akaike weights, which express relative support for each model. An AIC_w value for model i can be interpreted as the probability that model i is the best model for the observed data, given the candidate set of models. The pR^2 column, pseudo-R^2, shows the explained deviance, which is analogous to R^2 in a GLM. The remaining columns show standardized coefficients of each independent variable in each model, when present. (Inter. = Intercept, Dist./D. = Distance, Inc./I. = Income, A. = Age, R. = Race)

0$ (i.e., both areas have similar income) to 30,000$ (i.e., destination area has higher median income by 30,000$).

The effect of age was consistently negative (Table 3), again with a negligible effect size for the "age $\times$ distance" interaction (Figures 5–7). In other words, both follower and mobility tie realization consistently increased when destination area had lower median age. The effects of income and age dissimilarity thus maintained their direction (positive and negative, respectively) irrespective of distance, in both the Boston and US areas (Figures 5–7).

The effect of racial composition dissimilarity on tie formation was consistent among four out of five models, with the exception of the follower ties in the GBA (Table 3, Figure 5). In the four models, the highest tie realization rates were associated with low dissimilarity (i.e., high similarity) in racial composition—as might be expected. Additionally, a strong

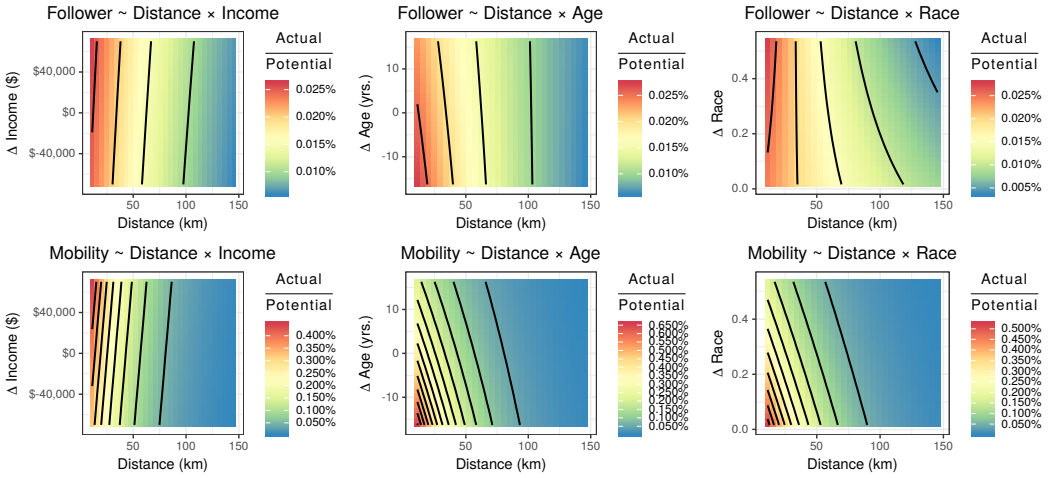

Figure 5: Predicted follower and mobility tie proportions, as function of geographical distance and socio-economic dissimilarity in the Greater Boston Area (GBA), based on models described in Table 3. Explained deviance was 8.6% and 17.1% (follower and mobility, respectively).

"race × distance" interaction effect was observed on the US scale—suggesting that race dissimilarity becomes irrelevant when long-distance ties are concerned, compared with short-distance ties which were more frequent when racial composition is similar (Figure 6). Predicted follower tie formation in Boston, however, was highest at short distances and high dissimilarity in racial composition (Figure 5), contrary to our expectation.

Explained deviance—the closest analogous metric to R^2 in a GLM—was 0.09 and 0.07 in follower tie models and 0.17 and 0.23 in the mobility tie models, for the GBA and the US, respectively, and 0.74 in the US commute model (i.e., follower<mobility<commute). Effect sizes of examined variables were also larger when predicting mobility compared with follower ties—most notably for the geographical distance effect (Figure 3), but for the SD variables as well (Table 3). Overall, the range of predicted tie realization within the 5-95% inter-quantile parameter space (in all independent variables) was 0.004-0.031% (7.5-fold) and 0.002-0.006% (3-fold) for follower models, compared with 0.003-1.071% (357-fold) and <0.001-0.989% (>989-fold) for mobility models, in the GBA and US areas, respectively. In other words, spatial variation in mobility ties, at least considering the portion explained by our examined variables, was much higher than that of follower ties.

4 Discussion

Our analysis bolsters the negative effect of geographical distance on Twitter virtual (i.e., follower) [14,79,80] and physical (i.e., mobility) [52] tie formation probabilities (*hypothesis 1*). We suggest that the relative weakness and low-cost of virtual tie formation [83] makes them less sensitive to distance, compared with physical (mobility) ties. Nevertheless, even for virtual ties distance is not "dead" [61] and proximity still makes a difference (Figures

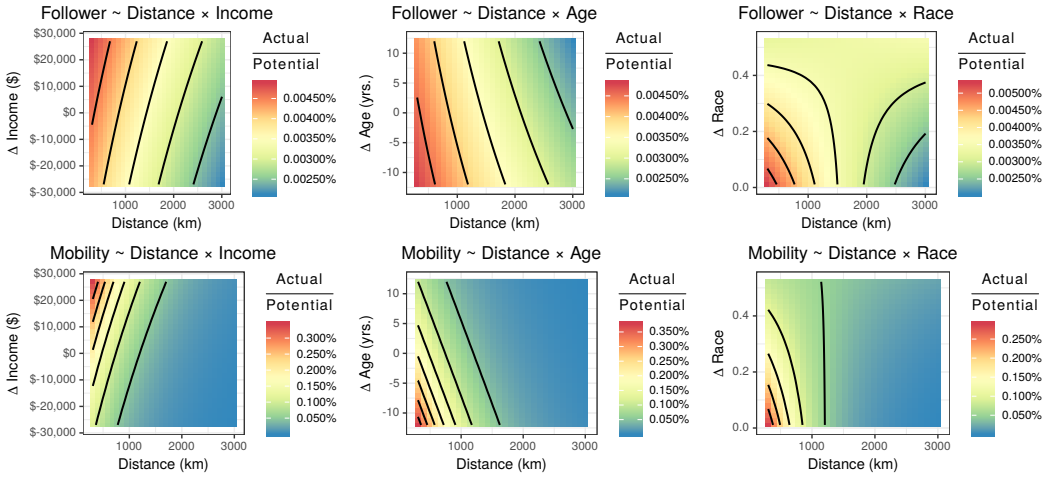

Figure 6: Predicted follower and mobility tie proportions, as function of geographical distance and socio-economic dissimilarity in the US, based on models described in Table 3. Explained deviance was 7.5% and 23.5% (follower and mobility, respectively).

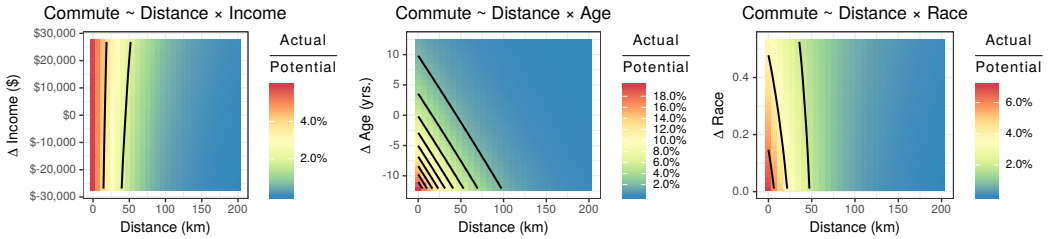

Figure 7: Predicted commute tie proportions, as function of geographical distance and socio-economic dissimilarity in the US, based on the model described in Table 3. Explained deviance was 73.8%.

3–4). We hypothesize that at least some of the follower ties are still complementary to face-to-face interaction, or directed towards similar interests relevant for a particular location, thus maintaining the association with distance [14, 80].

Conversely, mobility (i.e., physical) ties were strongly governed by distance—shrinking towards an average zero realization rate above a distance of several hundred kilometers. The cost of physical travel is higher than the cost of creating a virtual follower tie on Twitter [49]. We hypothesize that the additional cost is responsible for (1) sharper decline (2) towards a zero average rate, in case of mobility, as opposed to follower, tie formation (Figure 3). Naturally, commute travel was even more strongly affected by distance. The cost of maintaining regular commute is higher than that of conducting any given one-time travel (Figure 4). Indeed, regular commute travel to distances >160 km accounts for just 2.6% of commuter flows in the USA [64].

The distance decay weakening observed in Twitter virtual follower ties ($\beta = 0.79$ and $\beta = 0.90$, the the GBA and US, respectively) compared to physical mobility ($\beta = 1.00$ and $\beta = 1.17$) was in agreement with the difference observed between calls ($\beta = 1.45$) and mobility ($\beta = 1.60$) inferred from mobile phone data [21]. Overall, however, the present Twitter-based distance decay estimates (β between 0.79 and 1.17) were lower than estimates obtained from mobile-phone interactions (e.g., $\beta = 1.45$ [21], $\beta = 1.58$ [67], and $\beta = 2$ [44] for calls; $\beta = 1.60$ [21] and $\beta = 1.98$ [91] for movement). This is in line with previous studies which found relatively lower distance decay coefficients in online social networks, such as $\beta = 0.6$ in a former major Hungarian online social network [49]. The generally steeper decay observed in mobile phone call data compared to Twitter-based follower data may be explained by the fact that virtual interaction through mobile phones is mainly used for closer relationships that are fostered by other means too [49], thus more costly and more tightly associated with proximity. The steeper decay observed in mobile phone-based compared to Twitter-based movement data may be due to the high frequency of phone data collection, reflecting daily routine and thus emphasizing local activity, while Twitter usage may reflect more spatially diffuse leisure activity.

In addition to the effect of geographical distance, previous studies demonstrated that both types of tie formation—physical and virtual—are affected by dissimilarity in populations-level characteristics, such as spoken language [80], cultural barriers [36], and political or other interests [27, 28]. For example, residents of more similar socio-economic background potentially share more common interests and opportunities to socialize [59]. The novel aspects addressed here concern the innovative application of spatial network-analysis methodology to comprehensively model physical and virtual tie formation probability in different situations, with respect to SD settings as well as geographical distance.

Evaluating both distance and SD characteristics effects in the same model, our work is the first to show that the effect of geographical distance was stronger by an order of magnitude compared with the effects of SD characteristics dissimilarity—namely income, age, and race differences. Furthermore, the "distance $\times$ SD" interactions had smaller and less consistent effect size than the main effects. Finally, much of the variation in tie formation probability remained unexplained. We hypothesize that other population characteristics and common interests (such as political views) [24] may explain some of the remaining variation.

The two study areas—US and GBA—were generally characterized by similar patterns (Figures 5–7). The only substantial difference in tie formation determinants was observed in relation to the effect of racial composition. Namely, follower tie formation in the GBA was most frequent among nearby tracts of low racial composition similarity. We hypothesize that this unexpected result is caused by the relatively low variation in racial composition (84.1% white population) and concentration of other races in a few specific locations (Figure S3) which may be characterized by relatively high follower tie rates, due to unaccounted factors (such as economic activity). Conversely, follower, mobility, and commute tie formation in the US as a whole showed a consistent pattern, whereby ties are formed more frequently between counties of higher racial composition similarity (*hypothesis 2*). This pattern is consistent with individual-based social network studies. For example, in a national probability sample, only 8% of adults with networks of size two or more mention having a person of another race with whom they "discuss important matters", less than one seventh the heterogeneity that we would observe if people chose randomly from the population [56].

The fact that our Twitter-based findings were in agreement across the two analyzed scales and with the results of applying the same procedure to an independent data source—the commute dataset—strengthens their validity (Figures 5–7). This suggests that our results do indeed reflect real-world human behavior, rather than being an artifact of LBSN data [69]. Nevertheless we acknowledge the fact that assignment of individual-based data into areal units and subsequently to SD characteristics of those units—which is common to all three analyses—is associated with ecological fallacy [30]. We also acknowledge that geotagged Twitter data are not a representative sample of the population [33, 55, 60], generally biased towards younger users of higher income from urban areas. Furthermore, the data obtained for each given user is not necessarily a random representation of their mobility and social ties, for instance due to the fact that tweeting habits may be different depending on where the user travels [82]. The potential bias due to these factors cannot be completely ruled out, though it is unlikely due to the above-mentioned agreement across scales and methods, including the comparison to the census-based dataset on commute patterns. Moreover, weighted model results (Table S2) were highly consistent with the unweighted model results (Table 3), suggesting that Twitter data adequately represent population behavioral patterns with respect to the present study objectives. Similarly, Wang et al. [82] found highly consistent results when comparing regression models weighted by the ratio of Twitter users to the true population, in each geographic unit, with unweighted models.

Using a bi-directional network approach and directional predictors (income and age difference) our results highlight and quantify the asymmetric nature of spatial segregation in society (*hypothesis 3*). Follower, mobility, and commute ties were more frequently formed when directed towards areas of relatively higher median income and lower median age. The observed directional income effect is in line with previous small-scale studies on directional segregation in populations of contrasting socio-economic background. For example, daily movement from poor areas into rich ones was more frequent than the other way around in Bangkok's highly unequal economy [86]. While "the poor work for affluent residents in low-paid jobs as maids, cleaners, gardeners, drivers, or guards" thus regularly traveling to high-income areas, the affluent citizens do not have similar reasons to visit enclaves where poor people live [86]. A similar asymmetric mobility pattern was observed between two contrasting socio-economic regions of Louisville, based on Twitter data [76]. The present study confirms the generality of the phenomenon over the entire US, and demonstrates its validity not only in physical but also in virtual space.

We hypothesize that areas characterized by lower (i.e., younger) median age may be more economically influential and thus attracting more network attention, whether virtual (more influential persons or companies to "follow") or physical (more reasons to travel towards the area, e.g., for work or for recreation). Additionally, it has been shown that older people are disproportionately more likely to connect with younger ones, especially their children, compared to the general age homophily in social relations [57], which could further contribute to the asymmetry. Finally, areas with higher median income and lower median age may be characterized by a higher proportion of relevant "experts" that provide specialized knowledge, advice and services, thereby attracting further physical and virtual attention in the network [13].

It should be noted that segregation patterns are not merely a direct outcome of SD population characteristics and physical distance, but they are also shaped by the pre-existing spatial structure of cities. For example, Huang and Wong [30] analyzed travel patterns in Washington, D.C., using Twitter data on census tract resolution, concluding that "the urban

spatial structure, particularly where jobs are mainly found and the geographical layout of the region, plays a critical role in affecting the variation in activity patterns between users from different communities". Although urban structure, rather than SD differences, may also partially account for our results, we expect their role in our case to be minor, for several reasons. First, we analyzed virtual as well as mobility ties and found similar patterns in both. Virtual ties are clearly unconstrained by urban structure: any Twitter user can follow any other user at the same "cost", regardless of their spatial connectivity in the real world. Second, our large-scale analysis (US counties)—where urban structure is largely masked due to the aggregation of whole cities into the same areal unit—revealed similar results when compared with the local-scale (GBA) analysis. Third, we expect that our large sample of census tracts and counties covering a wide area (Table 2) to reflect a variety of different urban structures, thereby avoiding bias towards any specific structure, such as the one revealed in Washington, D.C. [30].

The present study comprises a first step towards a quantitative understanding of spatial segregation based on virtual and physical activity in a large human population. Understanding social factors that shape spatial community formation may initiate progress beyond exploratory community delineation [36, 64], towards modelling and prediction of spatial segregation. Specifically, identifying the types of situations which result in strong segregation may lead to better planning decisions for reducing its adverse effects. For example, the presented methodology can be used by urban planners to calculate the expected degree of physical and virtual segregation (Figures 5–7) between any given areal units, and to evaluate the expected degree of segregation under different scenarios. Conversely, actual metrics of segregation (Figure 2) can be contrasted with expected ones (Figures 5–7) (e.g., by calculating model residuals) to detect areas of unexpectedly high or low segregation levels, for reasons other than income, age, and race. For example, a follow-up study could investigate the potential causes of extremely low mobility or communication between adjacent census tracts (Figure 2) which are otherwise similar in SD characteristics.

To conclude, in the present study we examined spatial segregation in physical and virtual activity spaces, by applying a novel network-analysis approach to Twitter data. We showed that spatial segregation is more enhanced in physical space than in virtual space. The contribution of social characteristics to segregation was found to be smaller by an order of magnitude compared with geographical distance. Nonetheless, SD effects were ubiquitous and consistent at both region- and country-scale, and in virtual and physical ties alike. Specifically, tie formation was more frequent between pairs of areas characterized by more similar racial composition, and between pairs of areas where the "destination" has higher median income and lower median age. The presented methodology can help identify and map the intangible barriers for populations movement in physical space and their communication in the virtual one. Understanding the formation of such barriers is the first step towards reducing the negative effects of spatial segregation in human society.

References

[1] ALMQUIST, Z. W., ET AL. US Census spatial and demographic data in R: the `UScensus2000` suite of packages. *Journal of Statistical Software 37*, 6 (2010), 1–31. doi:10.18637/jss.v037.i06.

[2] ATKINSON, R., AND FLINT, J. Fortress UK? Gated communities, the spatial revolt

of the elites and time–space trajectories of segregation. *Housing studies 19*, 6 (2004), 875–892. doi:10.1080/0267303042000293982.

[3] BAILEY, M., CAO, R., KUCHLER, T., STROEBEL, J., AND WONG, A. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives 32*, 3 (2018), 259–80. doi:10.1257/jep.32.3.259.

[4] BAKILLAH, M., LI, R.-Y., AND LIANG, S. H. Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan. *International Journal of Geographical Information Science 29*, 2 (2015), 258–279. doi:10.1080/13658816.2014.964247.

[5] BARABÁSI, A.-L. The network takeover. *Nature Physics 8*, 1 (2011), 14. doi:10.1038/nphys2188.

[6] BARTON, K. *MuMIn: Multi-Model Inference*, 2018. R package version 1.42.1.

[7] BIVAND, R., AND RUNDEL, C. `rgeos`*: Interface to Geometry Engine - Open Source ('GEOS')*, 2018. R package version 0.3-28.

[8] BIVAND, R. S., PEBESMA, E. J., GOMEZ-RUBIO, V., AND PEBESMA, E. J. *Applied spatial data analysis with R*, 2nd ed. Springer, New York, 2013. doi:10.1007/978-1-4614-7618-4.

[9] BLANFORD, J. I., HUANG, Z., SAVELYEV, A., AND MACEACHREN, A. M. Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PloS one 10*, 6 (2015), e0129202. doi:10.1371/journal.pone.0129202.

[10] BORGATTI, S. P., MEHRA, A., BRASS, D. J., AND LABIANCA, G. Network analysis in the social sciences. *Science 323*, 5916 (2009), 892–895. doi:10.1126/science.1165821.

[11] BROWN, L. A., AND CHUNG, S.-Y. Spatial segregation, segregation indices and the geographical perspective. *Population, space and place 12*, 2 (2006), 125–143. doi:10.1002/psp.403.

[12] CONOVER, M. D., DAVIS, C., FERRARA, E., MCKELVEY, K., MENCZER, F., AND FLAMMINI, A. The geospatial characteristics of a social movement communication network. *PloS one 8*, 3 (2013), e55957. doi:10.1371/journal.pone.0055957.

[13] CORNWELL, E. Y., AND CORNWELL, B. Access to expertise as a form of social capital: An examination of race-and class-based disparities in network ties to experts. *Sociological Perspectives 51*, 4 (2008), 853–876. doi:10.1525/sop.2008.51.4.853.

[14] CRANDALL, D. J., BACKSTROM, L., COSLEY, D., SURI, S., HUTTENLOCHER, D., AND KLEINBERG, J. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences 107*, 52 (2010), 22436–22441. doi:10.1073/pnas.1006155107.

[15] CROITORU, A., WAYANT, N., CROOKS, A., RADZIKOWSKI, J., AND STEFANIDIS, A. Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems 53* (2015), 47–64. doi:10.1016/j.compenvurbsys.2014.11.002.

[16] CSARDI, G., AND NEPUSZ, T. The `igraph` software package for complex network research. *InterJournal Complex Systems* (2006), 1695.

[17] DE CHOUDHURY, M. Tie formation on Twitter: Homophily and structure of egocentric networks. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (2011), IEEE, pp. 465–470. doi:10.1109/PASSAT/SocialCom.2011.177.

[18] DUNCAN, O. D., AND DUNCAN, B. Residential distribution and occupational stratification. *American journal of sociology 60*, 5 (1955), 493–503. doi:10.1086/221609.

[19] DWYER, R. E. Expanding homes and increasing inequalities: US housing development and the residential segregation of the affluent. *Social Problems 54*, 1 (2007), 23–46. doi:10.1525/sp.2007.54.1.23.

[20] FEICK, R., AND ROBERTSON, C. A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems 53* (2015), 96–109. doi:10.1016/j.compenvurbsys.2013.11.006.

[21] GAO, S., LIU, Y., WANG, Y., AND MA, X. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS 17*, 3 (2013), 463–481. doi:10.1111/tgis.12042.

[22] GAO, S., YAN, B., GONG, L., REGALIA, B., JU, Y., AND HU, Y. Uncovering the digital divide and the physical divide in Senegal using mobile phone data. In *Advances in geocomputation*. Springer, 2017, pp. 143–151.

[23] GRUZD, A., WELLMAN, B., AND TAKHTEYEV, Y. Imagining twitter as an imagined community. *American Behavioral Scientist 55*, 10 (2011), 1294–1318. doi:10.1177/0002764211409378.

[24] HALBERSTAM, Y., AND KNIGHT, B. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics 143* (2016), 73–88. doi:10.1016/j.jpubeco.2016.08.011.

[25] HARRIS, R. Measuring the scales of segregation: Looking at the residential separation of White British and other schoolchildren in England using a multilevel index of dissimilarity. *Transactions of the Institute of British Geographers 42*, 3 (2017), 432–444. doi:10.1111/tran.12181.

[26] HAWELKA, B., SITKO, I., BEINAT, E., SOBOLEVSKY, S., KAZAKOPOULOS, P., AND RATTI, C. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science 41*, 3 (2014), 260–271. doi:10.1080/15230406.2014.890072.

[27] HERDAĞDELEN, A., ZUO, W., GARD-MURRAY, A., AND BAR-YAM, Y. An exploration of social identity: The geography and politics of news-sharing communities in Twitter. *Complexity 19*, 2 (2013), 10–20. doi:10.1002/cplx.21457.

[28] HIMELBOIM, I., MCCREERY, S., AND SMITH, M. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication 18*, 2 (2013), 154–174. doi:10.1111/jcc4.12001.

[29] HUANG, Q., CAO, G., AND WANG, C. From where do tweets originate?: a GIS approach for user location inference. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (2014), ACM, pp. 1–8. doi:10.1145/2755492.2755494.

[30] HUANG, Q., AND WONG, D. W. Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science 30*, 9 (2016), 1873–1898. doi:10.1080/13658816.2016.1145225.

[31] JAKUBS, J. F. A consistent conceptual definition of the index of dissimilarity. *Geographical Analysis 11*, 3 (1979), 315–321. doi:10.1111/j.1538-4632.1979.tb00698.x.

[32] JIANG, B., MA, D., YIN, J., AND SANDBERG, M. Spatial distribution of city tweets and their densities. *Geographical Analysis 48*, 3 (2016), 337–351. doi:10.1111/gean.12096.

[33] JIANG, Y., LI, Z., AND YE, X. Understanding demographic and socioeconomic biases of geotagged twitter users at the county level. *Cartography and Geographic Information Science 46*, 3 (2019), 228–242.

[34] JOHNSON, J. B., AND OMLAND, K. S. Model selection in ecology and evolution. *Trends in Ecology & Evolution 19*, 2 (2004), 101–108. doi:10.1016/j.tree.2003.10.013.

[35] JURDAK, R., ZHAO, K., LIU, J., ABOUJAOUDE, M., CAMERON, M., AND NEWTH, D. Understanding human mobility from Twitter. *PloS one 10*, 7 (2015). doi:10.1371/journal.pone.0131469.

[36] KALLUS, Z., BARANKAI, N., SZÜLE, J., AND VATTAY, G. Spatial fingerprints of community structure in human interaction network for an extensive set of large-scale regions. *PloS one 10*, 5 (2015), e0126713. doi:10.1371/journal.pone.0126713.

[37] KAPLAN, D. H., AND HOLLOWAY, S. R. Scaling ethnic segregation: causal processes and contingent outcomes in Chinese residential patterns. *GeoJournal 53*, 1 (2001), 59–70. doi:10.1023/A:1015822117915.

[38] KLOOG, I., CHUDNOVSKY, A., KOUTRAKIS, P., AND SCHWARTZ, J. Temporal and spatial assessments of minimum air temperature using satellite surface temperature measurements in Massachusetts, USA. *Science of the total environment 432* (2012), 85–92. doi:10.1016/j.scitotenv.2012.05.095.

[39] KLOOG, I., CHUDNOVSKY, A. A., JUST, A. C., NORDIO, F., KOUTRAKIS, P., COULL, B. A., LYAPUSTIN, A., WANG, Y., AND SCHWARTZ, J. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment 95* (2014), 581–590. doi:10.1016/j.atmosenv.2014.07.014.

[40] KLOOG, I., KOUTRAKIS, P., COULL, B. A., LEE, H. J., AND SCHWARTZ, J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric environment 45*, 35 (2011), 6267–6275. doi:10.1016/j.atmosenv.2011.08.066.

[41] KLOOG, I., MELLY, S. J., COULL, B. A., NORDIO, F., AND SCHWARTZ, J. D. Using satellite-based spatiotemporal resolved air temperature exposure to study the association between ambient air temperature and birth outcomes in Massachusetts. *Environmental health perspectives 123*, 10 (2015), 1053–1058. doi:10.1289/ehp.1308075.

[42] KLOOG, I., NORDIO, F., COULL, B. A., AND SCHWARTZ, J. Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. *Remote sensing of environment 150* (2014), 132–139. doi:10.1016/j.rse.2014.04.024.

[43] KOLACZYK, E. D., AND CSÁRDI, G. *Statistical analysis of network data with R*, vol. 65. Springer, New York, 2014. doi:10.1007/978-1-4939-0983-4.

[44] KRINGS, G., CALABRESE, F., RATTI, C., AND BLONDEL, V. D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment 2009*, 07 (2009), L07003. doi:10.1088/1742-5468/2009/07/l07003.

[45] KRIVO, L. J., WASHINGTON, H. M., PETERSON, R. D., BROWNING, C. R., CALDER, C. A., AND KWAN, M.-P. Social isolation of disadvantage and advantage: The reproduction of inequality in urban space. *Social Forces 92*, 1 (2013), 141–164. doi:10.1093/sf/sot043.

[46] 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 of the Association of American Geographers 103*, 5 (2013), 1078–1086. doi:10.1080/00045608.2013.792177.

[47] LAZER, D., PENTLAND, A., ADAMIC, L., ARAL, S., BARABÁSI, A.-L., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., GUTMANN, M., JEBARA, T., KING, G., MACY, M., ROY, D., AND VAN ALSTYNE, M. Computational social science. *Science 323*, 5915 (2009), 721–723. doi:10.1126/science.1167742.

[48] LEE, J. Y., AND KWAN, M.-P. Visualisation of socio-spatial isolation based on human activity patterns and social networks in space-time. *Tijdschrift voor economische en sociale geografie 102*, 4 (2011), 468–485. doi:10.1111/j.1467-9663.2010.00649.x.

[49] LENGYEL, B., VARGA, A., SÁGVÁRI, B., JAKOBI, Á., AND KERTÉSZ, J. Geographies of an online social network. *PloS one 10*, 9 (2015), e0137248. doi:10.1371/journal.pone.0137248.

[50] LI, F., AND WANG, D. Measuring urban segregation based on individuals' daily activity patterns: A multidimensional approach. *Environment and Planning A: Economy and Space 49*, 2 (2017), 467–486. doi:10.1177/0308518X16673213.

[51] LIBEN-NOWELL, D., NOVAK, J., KUMAR, R., RAGHAVAN, P., AND TOMKINS, A. Geographic routing in social networks. *Proceedings of the National Academy of Sciences 102*, 33 (2005), 11623–11628. doi:10.1073/pnas.0503018102.

[52] LIU, Y., SUI, Z., KANG, C., AND GAO, Y. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS one 9*, 1 (2014), e86026. doi:10.1371/journal.pone.0086026.

[53] LOVELACE, R., BIRKIN, M., CROSS, P., AND CLARKE, M. From big noise to big data: Toward the verification of large data sets for understanding regional retail flows. *Geographical Analysis 48*, 1 (2016), 59–81. doi:10.1111/gean.12081.

[54] MA, D., SANDBERG, M., AND JIANG, B. A socio-geographic perspective on human activities in social media. *Geographical Analysis 49*, 3 (2017), 328–342. doi:10.1111/gean.12122.

[55] MALIK, M. M., LAMBA, H., NAKOS, C., AND PFEFFER, J. Population bias in geotagged tweets. In *Ninth international AAAI conference on web and social media* (2015).

[56] MARSDEN, P. V. Core discussion networks of Americans. *American sociological review* (1987), 122–131. doi:10.2307/2095397.

[57] MARSDEN, P. V. Homogeneity in confiding relations. *Social networks 10*, 1 (1988), 57–76. doi:10.1016/0378-8733(88)90010-X.

[58] MASSEY, D. S., AND DENTON, N. A. The dimensions of residential segregation. *Social Forces 67*, 2 (1988), 281–315. doi:10.2307/2579183.

[59] MCPHERSON, M., SMITH-LOVIN, L., AND COOK, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*, 1 (2001), 415–444. doi:10.1146/annurev.soc.27.1.415.

[60] MISLOVE, A., LEHMANN, S., AHN, Y.-Y., ONNELA, J.-P., AND ROSENQUIST, J. N. Understanding the demographics of Twitter users. *ICWSM 11*, 5th (2011), 25.

[61] MOK, D., WELLMAN, B., AND CARRASCO, J. Does distance matter in the age of the internet? *Urban Studies 47*, 13 (2010), 2747–2783. doi:10.1177/0042098010377363.

[62] MORALES, A. J., DONG, X., BAR-YAM, Y., AND 'SANDY' PENTLAND, A. Segregation and polarization in urban areas. *Royal Society Open Science 6*, 10 (2019), 190573. doi:10.1098/rsos.190573.

[63] MUKAKA, M. M. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal 24*, 3 (2012), 69–71.

[64] NELSON, G. D., AND RAE, A. An economic geography of the United States: From commutes to megaregions. *PloS one 11*, 11 (2016), e0166083. doi:10.1371/journal.pone.0166083.

[65] NGUYEN, Q. C., KATH, S., MENG, H.-W., LI, D., SMITH, K. R., VANDERSLICE, J. A., WEN, M., AND LI, F. Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography 73* (2016), 77–88. doi:10.1016/j.apgeog.2016.06.003.

[66] OKA, M., AND WONG, D. W. Capturing the two dimensions of residential segregation at the neighborhood level for health research. *Frontiers in Public Health 2* (2014), 118. doi:10.3389/fpubh.2014.00118.

[67] ONNELA, J.-P., ARBESMAN, S., GONZÁLEZ, M. C., BARABÁSI, A.-L., AND CHRISTAKIS, N. A. Geographic constraints on social network groups. *PLoS one 6*, 4 (2011). doi:10.1371/journal.pone.0016939.

[68] PAOLA, J. M. Unravelling invisible inequalities in the city through urban daily mobility. The case of Santiago de Chile. *Swiss Journal of Sociology 33*, 1 (2007).

[69] PFEFFER, J., MAYER, K., AND MORSTATTER, F. Tampering with Twitter's Sample API. *EPJ Data Science 7*, 1 (2018), 50. doi:10.1140/epjds/s13688-018-0178-0.

[70] PRESTBY, T., APP, J., KANG, Y., AND GAO, S. Understanding neighborhood isolation through spatial interaction network analysis using location big data. *Environment and Planning A: Economy and Space* (2019), 0308518X19891911.

[71] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[72] REARDON, S. F., MATTHEWS, S. A., O'SULLIVAN, D., LEE, B. A., FIREBAUGH, G., FARRELL, C. R., AND BISCHOFF, K. The geographic scale of Metropolitan racial segregation. *Demography 45*, 3 (2008), 489–514. doi:10.1353/dem.0.0019.

[73] REARDON, S. F., AND O'SULLIVAN, D. Measures of spatial segregation. *Sociological Methodology 34*, 1 (2004), 121–162. doi:10.1111/j.0081-1750.2004.00150.x.

[74] ROSSUM, G. Python reference manual. Tech. rep., Amsterdam, The Netherlands, The Netherlands, 1995.

[75] SENARATNE, H., MOBASHERI, A., ALI, A. L., CAPINERI, C., AND HAKLAY, M. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science 31*, 1 (2017), 139–167. doi:10.1080/13658816.2016.1189556.

[76] SHELTON, T., POORTHUIS, A., AND ZOOK, M. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning 142* (2015), 198–211. doi:10.1016/j.landurbplan.2015.02.020.

[77] STEIGER, E., DE ALBUQUERQUE, J. P., AND ZIPF, A. An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Transactions in GIS 19*, 6 (2015), 809–834. doi:10.1111/tgis.12132.

[78] STEIGER, E., RESCH, B., AND ZIPF, A. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science 30*, 9 (2016), 1694–1716. doi:10.1080/13658816.2015.1099658.

[79] STEPHENS, M., AND POORTHUIS, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems 53* (2015), 87–95. doi:10.1016/j.compenvurbsys.2014.07.002.

[80] TAKHTEYEV, Y., GRUZD, A., AND WELLMAN, B. Geography of Twitter networks. *Social networks 34*, 1 (2012), 73–81. doi:10.1016/j.socnet.2011.05.006.

[81] VESSELINOV, E., CAZESSUS, M., AND FALK, W. Gated communities and spatial inequality. *Journal of Urban Affairs 29*, 2 (2007), 109–127. doi:10.1111/j.1467-9906.2007.00330.x.

[82] WANG, Q., PHILLIPS, N. E., SMALL, M. L., AND SAMPSON, R. J. Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences 115*, 30 (2018), 7735–7740. doi:10.1073/pnas.1802537115.

[83] WELLMAN, B., AND HAMPTON, K. Living networked on and offline. *Contemporary Sociology 28*, 6 (1999), 648–654. doi:10.2307/2655535.

[84] WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2016. doi:10.1007/978-3-319-24277-4.

[85] WILLIAMS, D. R., AND COLLINS, C. Racial residential segregation: A fundamental cause of racial disparities in health. *Public health reports 116*, 5 (2001), 404. doi:10.1093/phr/116.5.404.

[86] WISSINK, B., AND HAZELZET, A. Bangkok living: Encountering others in a gated urban field. *Cities 59* (2016), 164–172. doi:10.1016/j.cities.2016.08.016.

[87] WISSINK, B., SCHWANEN, T., AND VAN KEMPEN, R. Beyond residential segregation: Introduction. *Cities 59* (2016), 126–130. doi:10.1016/j.cities.2016.08.010.

[88] WONG, D. W. Comparing traditional and spatial segregation measures: A spatial scale perspective. *Urban Geography 25*, 1 (2004), 66–82. doi:10.2747/0272-3638.25.1.66.

[89] WOOD, S. N. *Generalized Additive Models: An introduction with R*. Chapman and Hall/CRC, New York, 2017. doi:10.1201/9781315370279.

[90] YIN, J., SOLIMAN, A., YIN, D., AND WANG, S. Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geolocated Twitter data. *International Journal of Geographical Information Science 31*, 7 (2017), 1293–1313. doi:10.1080/13658816.2017.1282615.

[91] ZHAO, Z., SHAW, S.-L., XU, Y., LU, F., CHEN, J., AND YIN, L. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science 30*, 9 (2016), 1738–1762. doi:10.1080/13658816.2015.1137298.

[92] ZUUR, A., IENO, E. N., WALKER, N., SAVELIEV, A. A., AND SMITH, G. M. *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media, 2009. doi:10.1007/978-0-387-87458-6.