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

M. Dorman*[1], T. Svoray[1,2] and I. 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

*Email: dorman@post.bgu.ac.il

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. Weighted spatial networks representing tie strength between geographical areas are constructed, followed by tie formation modelling 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 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: followers, mobility, network, spatial segregation, Twitter.

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 (Barabási, 2011). In particular, Location-Based Social Networks (LBSN), such as Twitter, provide opportunities to study spatial dimensions of human behaviour in detail (Ma et al., 2017).

Segregation is the extent to which individuals of different groups occupy or experience different social-environments. Groups generally form distinct patterns of over- and under-representation across residential regions (Brown and Chung, 2006).

The most commonly used measure of spatial segregation is the Index of Dissimilarity (ID) (Duncan and Duncan, 1955), that quantifies the evenness with which two demographic groups are distributed between areal units. Massey and Denton (1988) further identified five dimensions of segregation: *unevenness*, *exposure*, *clustering*, *concentration*, and *centralisation*. Of these, unevenness and clustering are regarded as the most important (Oka and Wong, 2014; Reardon and O'Sullivan, 2004).

Studies of spatial segregation are mostly focused on measuring population distribution patterns in residential space, based on census data (Dwyer, 2007)—ignoring the fact that social isolation likely extends from residential place to other locations and to other dimensions of activity (Krivo et al. 2013). Recently, it has been recognised that segregation studies should go beyond residential place to daily activity space (Li and Wang, 2017), and shift from location-based to people-based analysis (Kwan, 2013). The degree of mobility and communication between areas of contrasting socio-economic background are important aspects of the formation and maintenance of spatial segregation (Paola, 2007). In reference to LBSN, these two complementary components (Croitoru et al., 2015) are thought to be the strength of *physical / mobility* ties and of *friendship / virtual* ties, respectively.

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 (Huang and Wong, 2016), or have over-simplified the representation of physical space (De Choudhury, 2011). Recently, Ma et al. (2017) spatially analysed a large-scale friendship tie network between users. Aggregating social ties between users, the authors created both location-location and city-city spatial networks, to reveal tie counts across the entire area of the continental US. In the present study, we take the next step and quantitatively evaluate these 'characteristics of place' in terms of their effect on follower and mobility ties. Accounting for distance and population size, we focus on the effects of SD characteristics due to spatial segregation, as reflected through the recorded behaviour 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 recognised as the next frontier in studying segregation (Wissink et al., 2016). We propose a comprehensive hypothesis-testing oriented methodology that operates on a large sample of regions covering a wide and heterogeneous spatial extent, using two tie metrics that represent both mobility and friendship, while bearing in mind SD 'characteristics of place' (income, age and race). Our *aim* 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) To construct 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 used for validation.
(2) Fitting models where tie strength between given areas in the latter five networks is explained with their distance, SD dissimilarity and their 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) and 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* (McPherson et al., 2001).

(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 (Wissink and Hazelzet, 2016).

2. Methods

2.1. Twitter Data

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 that each user *follows* (otherwise known as their 'friends'). The analysed dataset consisted of point locations (lon-lat) that each unique user posted from, coupled with an indication on whether or not a social tie exists between each pair of users.

Although geo-tagged tweets comprise only 1–2% of the total volume (Lovelace et al., 2016), thanks to the relatively prolonged collection period we were able to collect a large sample size of over 20M tweets sent by ~900K unique users (Table 1), that can serve as a good approximation of human activities for study purposes (Ma et al., 2017).

The process was repeated on two spatial scales:

(1) **County scale**, the contiguous United States (US) (~8,000,000 km^2)
(2) **Census tract scale**, a rectangular area of ~50,000 km^2 in the Greater Boston Area (GBA)

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.

2.2. Socio-economic and demographic (SD) data

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 (Almquist, 2010). Three key characteristics (Nguyen et al., 2016) were extracted: median household income (ACS code: B19013e1), median age (B01002e1), and total population of each racial/ethnic group (B02001e2, B02001e3, B02001e4, B02001e5, B02001e6, B02001e7) (see Table S1 and Fig. S1-S6 in Supplemental Materials).

Dissimilarity between each pairs of areas was expressed as Euclidian distance. For racial composition, multi-dimensional dissimilarity was calculated by means of the function `dist` in R, using `method="euclidean"`, after converting 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 origin and destination areas.

2.3. Commute data

To validate the results obtained using Twitter, we reproduced the analysis of socio-economic effects on mobility realisation through an external data source – the *2009-2013 5-Year ACS Commuting Flows* dataset[1]. This dataset contains counts of workers in commuting flows between each pair of counties in the USA (Nelson and Rae, 2016). To standardise the number of commuters by potential commuters count, we also obtained the county-level labour 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 – since estimates of SD characteristics are only available in the latter case.

Aggregation involved the following steps:

(1) Assigning each Twitter user to the areal unit where he/she is most active, which we labelled his/her *centre of activity*—defined as the areal unit with the greatest number of tweets for a given user (Ma et al., 2017). One of the reasons for this broad definition, rather than attempting to detect a *place of residence* (Huang and Wong, 2016; Huang et al., 2014), is that Twitter accounts for organisations, agencies, services, etc., rather than individuals, are increasingly more common (Senaratne et al., 2017), and for such users, the term *place of residence* is naturally irrelevant. 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.

(2) Calculating virtual and physical tie strength metrics between all possible pairs of areal units A and B in the study area, in accordance with the following algorithms for each tie type -

    1. **Follower** = The number of *follower ties* between a user from area A and one from 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).
    2. **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.
    3. **Commute** = The commuting flow count from area A to area B, divided by total labour force in area A.

(3) Assigning each pair of areas A and B with the distance between their centroids, as well as the corresponding set of SD dissimilarity metrics (see above).

(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).

Note that the tie strength indices are directional—i.e., tie strength for A→B is not necessarily the same as tie strength for B→A. Self-ties where the origin and destination are the same (e.g., A→A) were

---

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

excluded from analysis. Also note that the tie strength indices are inherently standardised by total network activity for removing group size bias (Lengyel et al., 2015).

2.5. Statistical analysis

A *preliminary* visual evaluation of 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 × two tie types) (Figure 3). The dependent variable was the tie strength estimate (Figure 1), while the independent variable was the distance between the respective areas. The observations comprised all network edges—i.e., all ties between pairs of areas.

The *main* statistical approach follows section *9.2–Modeling Network Flows: Gravity Models* in Kolaczyk and Csárdi (2014) (pp. 162-170). The input data for the statistical analysis comprised the five weighed networks: Twitter-derived networks for two spatial scales × two tie types + the US commute network. In the main analysis, we considered not only distance, but also SD dissimilarity, and the interactions of SD dissimilarity with distance. In each case, we statistically tested whether realisation of follower or mobility ties was associated with the latter variables, and if so – how.

Generalized Linear Models (GLM) 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 actual and potential network ties (Figure 1). The independent variables were: 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 between the four examined variables on the network edges – i.e., distance and dissimilarity of income, age and racial composition. Although spatial autocorrelation between network edges is not clearly defined, and pairwise neighbour weighting is computationally unfeasible for sample sizes of 3,097,600 or 9,659,664 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 (Bivand et al., 2013), with the 8-nearest-neighbours edge centroid criterion for defining neighbour weights. There was no significant autocorrelation in the follower network (p-value = 0.53), or in 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 four networks (Table 3). In each case, models were ordered by the AIC score – from lowest AIC (i.e., highest relative support) to highest (i.e., lowest relative support). The hypotheses underpinning the inclusion of variables present in the most parsimonious models (i.e. having the lowest AIC) were considered supported by the data (Johnson and Omland, 2004). The five most parsimonious models were eventually used to generate and visualise predicted tie strength in the studied parameter space (Figure 4), to characterise 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 (Johnson & Omland 2004).

Model predictions were calculated using non-standardised GLM model coefficients, thus allowing for interpretation in the original units for each variables (e.g. ‘kilometres’ for distance, or $ for income difference) (Figure 4). In addition, Table 3 reports the standardised model coefficients. Standardised regression coefficients basically refer to how many standard deviations a dependent variable will change per standard deviation increase in the predictor variable. Standardised coefficients are therefore useful when numerically comparing effect sizes between variables in the same model, or among different models based on the same data.

2.6. Software

Accessing the Twitter APIs for data collection was done by means the Python package *twarc*[2]. All other analyses were carried out in R (R Core Team, 2018). Spatial processing of the Twitter and census data were executed using R packages *sp* (Bivand et al., 2013) and *rgeos* (Bivand and Rundel, 2017). Network construction and statistical calculations were performed using package *igraph* (Csardi and Nepusz, 2006). Generalized Additive Models (GAM) were fitted using package *mgcv* (Wood, 2006). Model selection procedure of Generalized Linear Models (GLM) was done through package *MuMIn* (Barton, 2016). Figures were produced with package *ggplot2* (Wickham, 2016).

3. Results

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) than in the mobility networks (2.85% and 3.12%). In other words, a higher proportion of area pairs was characterised 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 commutes take place between a small subset of (adjacent) county pairs, out of all possible county pairs in the US.

Follower ties and mobility ties markedly differ in their form of distance decay (Figure 3). In the main analysis, according to the AIC-based model selection procedure, the full models had 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 × Distance’ effect – although in that case the full model came second, not far behind in terms of relative support (AIC_w of 0.31 vs. 0.69). Specifically, the hypothesis that distance and SD dissimilarity and their interactions (with the exception of the ‘Income × Distance’ interaction for follower ties in the GBA) affect follower and mobility tie formation was supported by the data.

The effect of distance on tie formation was: (a) consistently negative; and (b) larger for mobility ties than for follower ties. In terms of effect size, observing best models’ predicted values in the relevant parameter space (Figure 4) as well as standardised coefficients (Table 3) revealed that distance effect on mobility tie formation was stronger by an order of magnitude compared with follower tie formation. The effect of distance on commute frequency was higher still – however, it should be noted that the x-axis for commute predictions in Figure 4 does not show the full range of distances, but only distances up to 200 km, since commute realisation above that distance is practically nil.

---

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

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, Figure 4), with no substantial 'income × distance' interaction effect size in the studied parameter space. Age effect was consistently negative (Table 3, Figure 4) – i.e., follower and mobility tie realisation constantly increased when destination area had a *lower median age*. The effects of income and age dissimilarity thus maintained their direction (positive and negative, respectively) irrespective of distance, at both sites (Figure 4).

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 4a). In the four models, the highest tie realisation rates were associated with *low dissimilarity* (i.e. high similarity) in racial composition – as might be expected. Additionally, a strong 'race × distance' interaction effect was observed at 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 4b). Predicted follower tie formation in Boston, however, was highest at short distances and *high dissimilarity* in racial composition.

Explained deviance was 8.6% and 7.5% in follower tie models, and 17.1% and 23.5% in the mobility tie models, for the GBA and the US, respectively, and 54.7% in the US commute model. Effect sizes of examined variables were also larger when predicting mobility compared with follower ties – most notably for distance effect (Figure 3), but for the SD variables, as well (Table 3). Overall, the range of predicted tie realisation within the 0.05–0.95 inter-quantile parameter space (in all independent variables) was 0.004–0.031% and 0.002–0.006% for follower models, compared with 0.003–1.071% and <0.001–0.989% for mobility models, in the GBA and US areas, respectively.

4. Discussion

Our analysis bolsters the negative effect of distance on virtual (Stephens and Poorthuis, 2015) and physical (Liu et al., 2014) Twitter tie formation probabilities (*hypothesis 1*). We suggest that the relative weakness and low-cost of virtual tie formation (Wellman and Hampton, 1999) makes them less sensitive to distance, compared with physical ties. Nevertheless, even for virtual ties distance is not 'dead' (Mok et al., 2010), and proximity still makes a difference (Figure 3).

Physical ties were governed by distance – shrinking towards an average zero realisation rate above a distance of several hundred kilometres. Individuals may have more reasons to physically travel short distances, however the cost of physical travel is higher than the cost of creating a virtual follower tie on Twitter. We hypothesise that this additional cost is responsible for: (1) sharper decline and (2) towards a zero average rate, in case of mobility, as opposed to follower tie formation (Figure 3). Naturally, the cost of maintaining a regular commute is higher than that of conducting any given one-time travel (Figure 3). Indeed, regular commute travel to distances of over 160km accounts for only 2.6% of commuter flows in the USA (Nelson and Rae, 2016).

In addition to distance effect, previous studies demonstrated that both types of tie formation—virtual and physical—are affected by dissimilarity in populations-level characteristics, such as spoken language (Takhteyev et al., 2012), cultural barriers (Kallus et al., 2015), and political or other interests. Our work is the first to show that the effect of distance was stronger by an order of magnitude than that of SD characteristics dissimilarity – namely income, age and race differences: the distance × SD interactions had smaller and less consistent effect size than the main effects. Even so, much of the variation in tie formation probability remains unexplained. We hypothesise that other population characteristics and common interests (such as political views) (Halberstam and Knight, 2016) may explain some of the remaining variation.

The two study areas were generally characterised by similar patterns (Figure 4). The only substantial difference in tie formation determinants was observed in relation to the effect of racial composition. In other words, follower tie formation in the GBA was most frequent among nearby tracts of low racial composition similarity. We hypothesise 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 (Fig. S3), which may be characterised 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 (Marsden, 1987).

The fact that our Twitter-based findings were in agreement across the two analysed scales and with the results of applying the same procedure to an independent data source – the commute dataset – strengthens their validity (Figure 4). This suggests that our results do indeed reflect real-world human behaviour, rather than being an artefact of LBSN data (Pfeffer et al., 2018).

Using a bi-directional network approach and directional predictors (income and age difference) our results highlight 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, an asymmetric mobility pattern was observed between two contrasting socio-economic regions of Louisville, based on Twitter data (Shelton et al., 2015). We hypothesise that areas of a younger median age may be more economically influential, and therefore attract more network attention – be it virtual or physical. In addition, areas of a higher median income and lower median age may have a higher proportion of relevant 'experts' that provide specialised knowledge, advice and services, thereby attracting further physical and virtual attention in the network (Cornwell and Cornwell, 2008).

It should be noted that segregation patterns are not merely a direct outcome of SD population characteristics and physical distance, but are also shaped by the pre-existing spatial structure of cities. 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, because we analysed 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. (Huang and Wong, 2016).

Understanding social factors that shape spatial community formation may initiate progress beyond exploratory community delineation (Nelson and Rae, 2016), towards spatial segregation prediction. We examined spatial segregation in physical and virtual activity spaces, by applying a 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 distance. Nonetheless, SD effects were ubiquitous and consistent at both region- and country-scale, and in virtual and physical ties alike. Mapping intangible barriers for population movement in physical and virtual space can contribute to understanding the formation of such barriers as a first step towards reducing the negative effects of spatial segregation in human society.

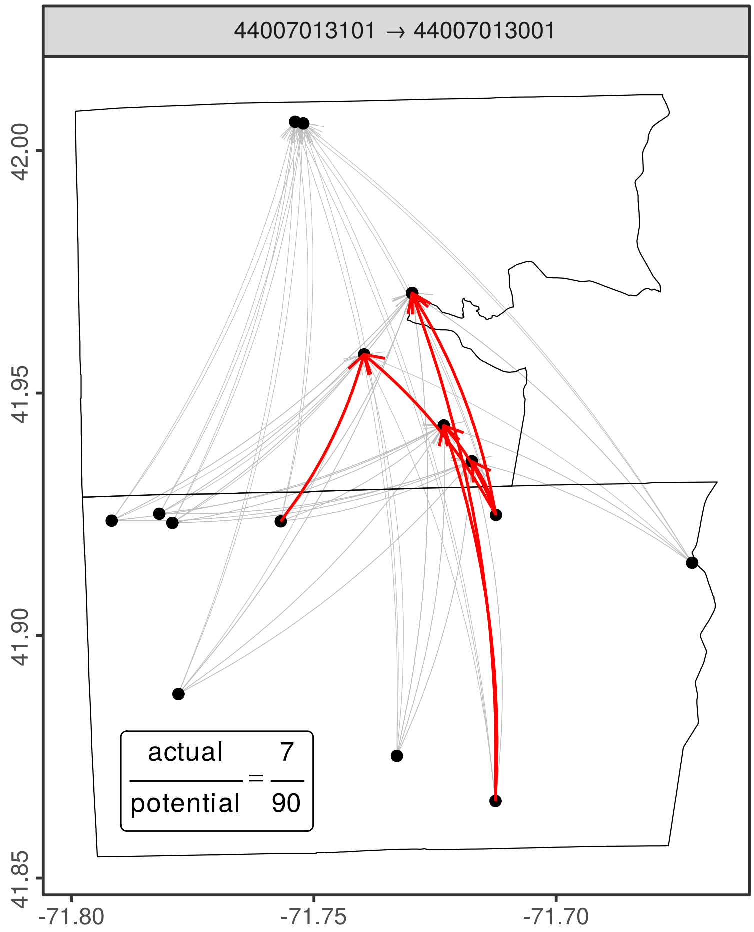


Figure 1: Calculation of the follower tie ratio between two census tracts in the Greater Boston Area (GBA) (44007013101 and 44007013001). Grey segments represent all 90 possible follower ties extending from Twitter users whose estimated ‘centre of activity’ is located in tract 44007013101 towards users in whose ‘centre of activity’ is in tract 44007013001. Red segments represent the 7 ties that are actually realised. Follower tie ratio for the 44007013101→44007013001 edge is therefore equal to 7/90, or 7.78%.

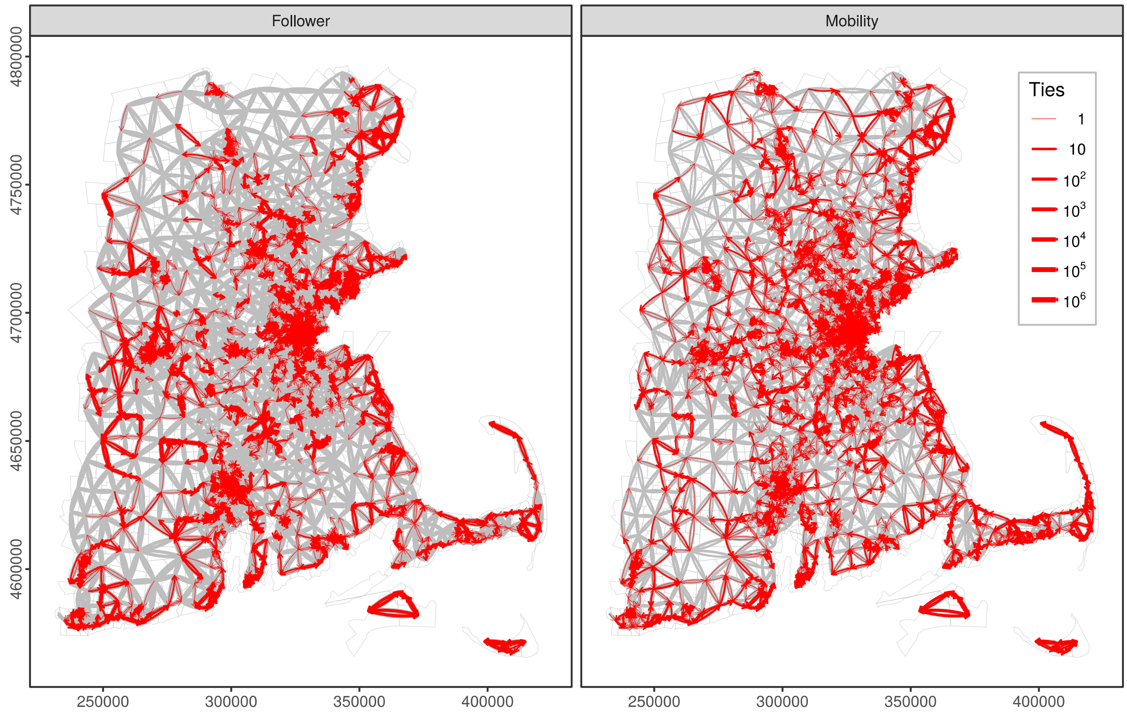


Figure 2: Observed follower and mobility ties between adjacent census tracts in the Greater Boston Area (GBA). Grey lines represent potential ties count, red lines represent actual count. Note that for visual clarity, these figures do not display the entire networks, but only sub-networks of adjacent 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.

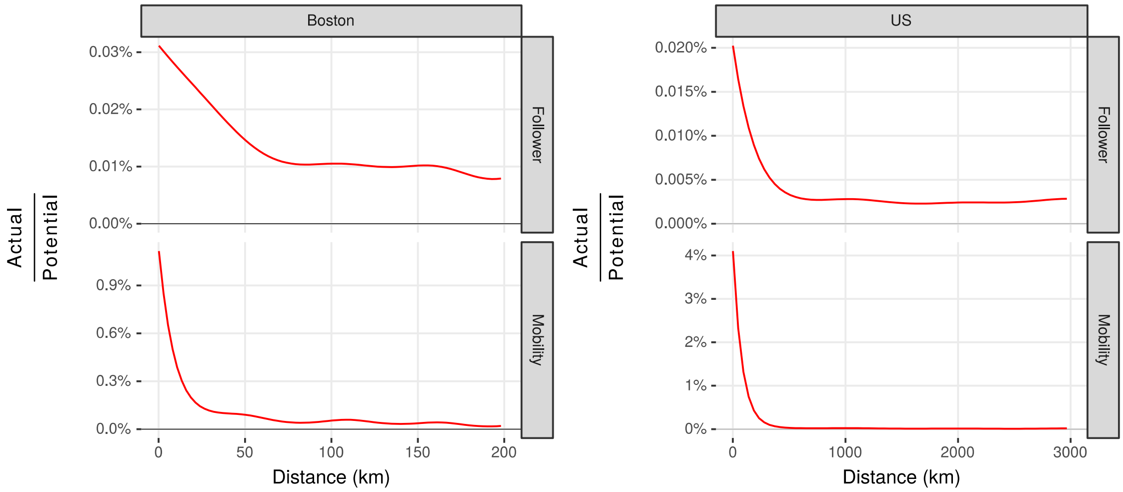
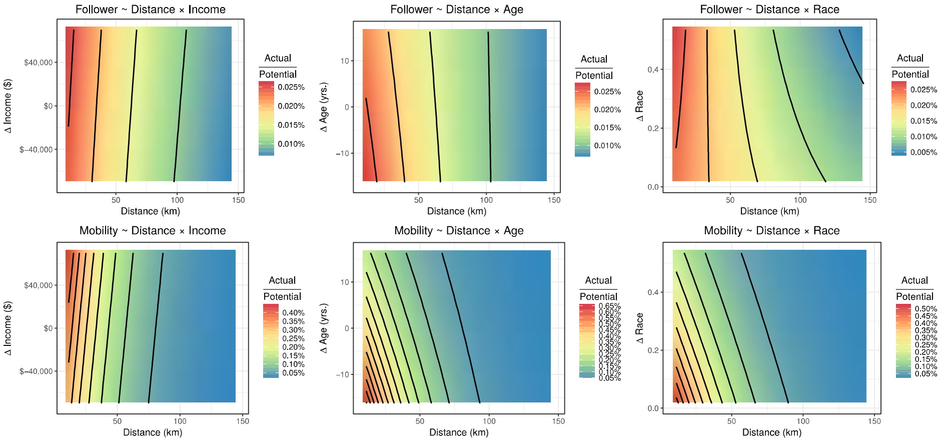
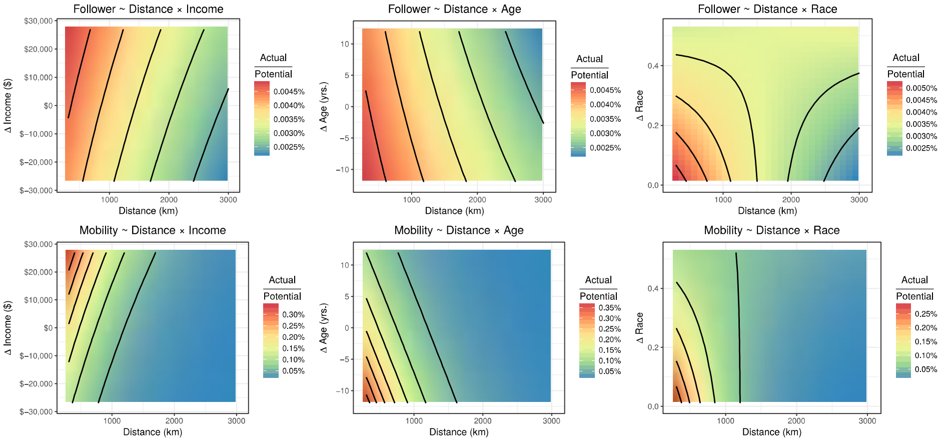


Figure 3: Follower and mobility tie proportions as function of geographical distance in the Greater Boston Area (GBA) and the US. Lines show the average trend based on a Generalised Additive Model (GAM).

(a) Follower and Mobility in the GBA



(b) Follower and Mobility in the US



(c) Commute in the US

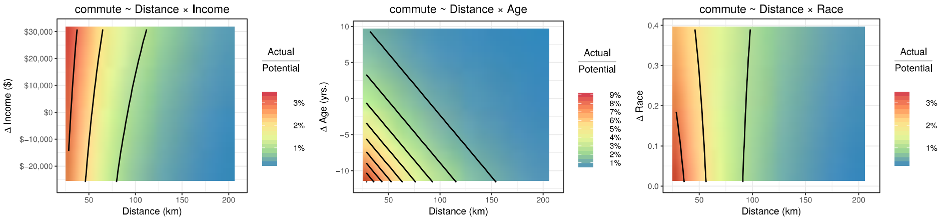


Figure 4: Predicted follower, mobility and commute tie proportions, as function of geographical distance and socio-economic dissimilarity in the Greater Boston Area (GBA) and the US, based on models described in Table 3. Explained deviance was 8.6% and 17.1% in the GBA (follower and mobility, respectively), and 7.5%, 23.5% and 54.7% in the US (follower, mobility and commute) respectively. Predicted values are shown as function distance and income dissimilarity (1st column), distance and age dissimilarity (2nd column) and distance and race dissimilarity (3rd column).

	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.

	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%	10.4%	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.

Area	Type	Intercept	Dist.	Income	Age	Race	Income×Dist.	Age×Dist.	Race×Dist.	AIC$_w$
GBA	Follower	**-8.863**	**-0.409**	**0.027**	**-0.021**	**-0.076**	**0.005**	**0.018**	**-0.082**	**0.996**
		-8.863	-0.409	0.025	-0.021	-0.076	-	0.019	-0.082	0.004
		-8.863	-0.41	0.027	-0.023	-0.076	0.013	-	-0.082	<0.001
		-8.863	-0.409	-	-0.015	-0.076	-	0.015	-0.082	<0.001
		-8.863	-0.41	0.019	-0.024	-0.076	-	-	-0.082	<0.001
		-8.863	-0.409	-	-0.019	-0.076	-	-	-0.082	<0.001
	Mobility	**-7.259**	**-1.183**	**0.104**	**-0.257**	**-0.308**	**-**	**0.056**	**0.032**	**0.69**
		-7.259	-1.183	0.102	-0.257	-0.309	-0.003	0.057	0.032	0.31
		-7.267	-1.191	0.099	-0.257	-0.339	-0.006	0.058	-	<0.001
		-7.267	-1.191	0.104	-0.257	-0.339	-	0.056	-	<0.001
		-7.269	-1.191	0.111	-0.298	-0.307	0.016	-	0.036	<0.001
		-7.269	-1.191	0.096	-0.299	-0.309	-	-	0.033	<0.001

US	Follower	**-10.243**	**-0.188**	**0.056**	**-0.049**	**-0.033**	**0.01**	**-0.013**	**0.1**	**1**
		-10.244	-0.188	0.056	-0.053	-0.033	0.01	-	0.1	<0.001
		-10.244	-0.187	0.06	-0.049	-0.033	-	-0.011	0.1	<0.001
		-10.244	-0.187	0.06	-0.053	-0.033	-	-	0.1	<0.001
		-10.243	-0.188	0.054	-	-0.033	0.009	-	0.099	<0.001
		-10.243	-0.187	0.057	-	-0.033	-	-	0.1	<0.001
	Mobility	**-7.857**	**-1.235**	**0.454**	**-0.396**	**0.068**	**0.072**	**0.005**	**0.386**	**0.998**
		-7.857	-1.234	0.454	-0.399	0.068	0.072	-	0.386	0.002
		-7.861	-1.237	0.407	-0.395	0.059	-	0.003	0.377	<0.001
		-7.861	-1.237	0.407	-0.397	0.059	-	-	0.377	<0.001
		-7.772	-1.165	0.436	-0.391	-0.124	0.057	0.037	-	<0.001
		-7.77	-1.16	0.435	-0.414	-0.125	0.054	-	-	<0.001
	Commute	**-36.912**	**-20.083**	**3.004**	**0.288**	**4.053**	**1.769**	**0.288**	**2.513**	**1**
		-36.982	-20.127	3.022	-	4.066	1.782	-	2.52	<0.001
		-36.897	-20.081	3.242	0.445	0.058	1.917	0.445	-	<0.001
		-37.003	-20.147	3.242	-	0.058	1.919	-	-	<0.001
		-36.925	-20.092	3.276	0.461	-	1.937	0.461	-	<0.001
		-37.034	-20.16	3.274	-	-	1.938	-	-	<0.001

Table 3: Model selection results for GLM models 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 their interactions. Only the six most highly supported models are shown per model selection procedure (i.e. area and network type). Models are ordered by decreasing AIC, starting from the most supported model (in **bold**). The last column shows Akaike weights (AIC_w), which express relative weight of evidence for each model, summing to 1 across all models. 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 model. The remaining columns show standardised coefficients of each independent variable in each model (when present).

3. References

Almquist, Z.W. (2010). US Census spatial and demographic data in R: the UScensus2000 suite of packages. J. Stat. Softw. 37, 1–31.

Barabási, A.-L. (2011). The network takeover. Nat. Phys. 8, 14.

Barton, K. (2016). MuMIn: Multi-Model Inference.

Bivand, R., and Rundel, C. (2017). rgeos: Interface to Geometry Engine - Open Source (GEOS).

Bivand, R.S., Pebesma, E., and Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R (New York: Springer).

Brown, L.A., and Chung, S.-Y. (2006). Spatial segregation, segregation indices and the geographical perspective. Popul. Space Place 12, 125–143.

Cornwell, E.Y., and Cornwell, B. (2008). Access to expertise as a form of social capital: An examination of race-and class-based disparities in network ties to experts. Sociol. Perspect. 51, 853–876.

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., and Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. Comput. Environ. Urban Syst. 53, 47–64.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Systems, 1695.

De Choudhury, M. (2011). 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, (IEEE), pp. 465–470.

Duncan, O.D., and Duncan, B. (1955). Residential distribution and occupational stratification. Am. J. Sociol. 60, 493–503.

Dwyer, R.E. (2007). Expanding homes and increasing inequalities: US housing development and the residential segregation of the affluent. Soc. Probl. 54, 23–46.

Halberstam, Y., and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. J. Public Econ. 143, 73–88.

Huang, Q., and Wong, D.W. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? Int. J. Geogr. Inf. Sci. 30, 1873–1898.

Huang, Q., Cao, G., and Wang, C. (2014). 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, (ACM), pp. 1–8.

Johnson, J.B., and Omland, K.S. (2004). Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108.

Kallus, Z., Barankai, N., Szuele, J., and Vattay, G. (2015). Spatial Fingerprints of Community Structure in Human Interaction Network for an Extensive Set of Large-Scale Regions. Plos One 10, e0126713.

Kolaczyk, E.D., and Csárdi, G. (2014). Statistical analysis of network data with R (Springer).

Krivo, L.J., Washington, H.M., Peterson, R.D., Browning, C.R., Calder, C.A., and Kwan, M.-P. (2013). Social isolation of disadvantage and advantage: The reproduction of inequality in urban space. Soc. Forces 92, 141–164.

Kwan, M.-P. (2013). Beyond space (as we knew it): toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and GIScience. Ann. Assoc. Am. Geogr. 103, 1078–1086.

Lengyel, B., Varga, A., Sagvari, B., Jakobi, A., and Kertesz, J. (2015). Geographies of an Online Social Network. Plos One 10, e0137248.

Li, F., and Wang, D. (2017). Measuring urban segregation based on individuals' daily activity patterns: A multidimensional approach. Environ. Plan. A 49, 467–486.

Liu, Y., Sui, Z., Kang, C., and Gao, Y. (2014). Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. Plos One 9, e86026.

Lovelace, R., Birkin, M., Cross, P., and Clarke, M. (2016). From Big Noise to Big Data: Toward the Verification of Large Data sets for Understanding Regional Retail Flows. Geogr. Anal. 48, 59–81.

Ma, D., Sandberg, M., and Jiang, B. (2017). A Socio-Geographic Perspective on Human Activities in Social Media. Geogr. Anal. 49, 328–342.

Marsden, P.V. (1987). Core discussion networks of Americans. Am. Sociol. Rev. 122–131.

Massey, D.S., and Denton, N.A. (1988). The dimensions of residential segregation. Soc. Forces 67, 281–315.

McPherson, M., Smith-Lovin, L., and Cook, J.M. (2001). Birds of a feather: Homophily in social networks. Annu. Rev. Sociol. 27, 415–444.

Mok, D., Wellman, B., and Carrasco, J. (2010). Does Distance Matter in the Age of the Internet? Urban Stud. 47, 2747–2783.

Nelson, G.D., and Rae, A. (2016). An Economic Geography of the United States: From Commutes to Megaregions. PLOS ONE 11, e0166083.

Nguyen, Q.C., Kath, S., Meng, H.-W., Li, D., Smith, K.R., VanDerslice, J.A., Wen, M., and Li, F. (2016). Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. Appl. Geogr. 73, 77–88.

Oka, M., and Wong, D.W. (2014). Capturing the two dimensions of residential segregation at the neighborhood level for health research. Front. Public Health 2, 118.

Paola, J.M. (2007). Unravelling Invisible Inequalities in the City through Urban Daily Mobility. The Case of Santiago de Chile. Swiss J. Sociol. 33.

Pfeffer, J., Mayer, K., and Morstatter, F. (2018). Tampering with Twitter's Sample API. EPJ Data Science, 7(1), 50.

R Core Team (2018). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).Reardon, S. F., and O'Sullivan, D. (2004). Measures of spatial segregation. Sociological methodology, 34(1), 121-162.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., and Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. Int. J. Geogr. Inf. Sci. 31(1), 139-167.

Shelton, T., Poorthuis, A., and Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. Landsc. Urban Plan. 142, 198–211.

Stephens, M., and Poorthuis, A. (2015). Follow thy neighbor: Connecting the social and the spatial networks on Twitter. Comput. Environ. Urban Syst. 53, 87–95.

Takhteyev, Y., Gruzd, A., and Wellman, B. (2012). Geography of Twitter networks. Soc. Netw. 34, 73–81.

Wellman, B., and Hampton, K. (1999). Living networked on and offline. Contemp. Sociol. 28, 648–654.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).

Wissink, B., and Hazelzet, A. (2016). Bangkok living: Encountering others in a gated urban field. Cities.

Wissink, B., Schwanen, T., and van Kempen, R. (2016). Beyond residential segregation: Introduction. Cities.

Wood, S. (2006). Generalized additive models: an introduction with R (CRC press).