scientific reports

Check for updates

OPEN

Universal patterns of long-distance commuting and social assortativity in cities

Eszter Bokányi[1,2]✉, Sándor Juhász[1,2], Márton Karsai[3,4] & Balázs Lengyel[1,2]

Millions commute to work every day in cities and interact with colleagues, partners, friends, and strangers. Commuting facilitates the mixing of people from distant and diverse neighborhoods, but whether this has an imprint on social inclusion or instead, connections remain assortative is less explored. In this paper, we aim to better understand income sorting in social networks inside cities and investigate how commuting distance conditions the online social ties of Twitter users in the 50 largest metropolitan areas of the United States. An above-median commuting distance in cities is linked to more diverse individual networks, moreover, we find that longer commutes are associated with a nearly uniform, moderate reduction of overall social tie assortativity across all cities. This suggests a universal relation between long-distance commutes and the integration of social networks. Our results inform policy that facilitating access across distant neighborhoods can advance the social inclusion of low-income groups.

Cities are champions of diversity[1–3]. Complex interaction networks of individuals in urban areas enabled by population density, co-location, and easy access together made cities the global engines of technological and economic progress[4–7]. However, cities are also known for high levels of segregation[8–10] where disparate neighborhoods are separated from each other in the urban space[11–15]. Furthermore, spatial segregation by income also fragments social networks, which can hinder progress and can deepen inequalities[16–20]. Given the importance of this problem, a growing community has investigated the patterns of mobility in cities to better understand mixing potentials across disparate and diverse neighborhoods[21–24], which may increase economic prosperity[25]. Yet, less is known whether mobility mixing has any imprint on the social connections of people.

Commuting covers a large share of urban mobility[26] and by connecting home with work locations, the places where people spend most of their time, it plays an important role in the spatial formation of social connections[27–29]. Since aggregated social networks form spatially bounded communities across neighborhoods[17], the further one commutes, the higher the likelihood that commuting-related social connections will introduce diversity in the egocentric network of the commuter[30,31]. Due to spatial segregation, economically disparate neighborhoods tend to be far from each other[32], thus long commutes are more likely to link places with different social status[33,34]. Nevertheless, it is not trivial that long commutes should facilitate social inclusion, because social interactions might remain assortative even at places far from home[21,23,35]. Meanwhile, the time to develop new social connections is especially limited for low-income workers who travel to work during rush hours[36,37].

The spatial distribution of high versus low-income households determines the length of travel that can bridge disparate neighborhoods. Since the scale of socio-economic isolation greatly varies across cities[12], one may expect that the mobility of people also enables a different degree of social mixture. However, the assortativity of urban mobility is a universal feature across cities: individuals have been recently reported to visit locations that are similar to their home neighborhood[21,23,38–40]. Yet, how assortativity of commuting and social networks are related and how this relation is modified by the length of commute in cities is still largely uncovered.

In this paper, we aim to better understand how mixing in urban social networks is facilitated by commuting. To answer this question, we use a unique dataset on 348,850 Twitter users living in the 50 largest metropolitan areas of the US and track their home and work locations as well as their mutual followership ties on the platform, which from now on, we call the social network of users. We project these social networks in the urban space and attribute users with an average income based on their home locations on an income map extracted from census

[1]Laboratory for Networks, Technology and Innovation, Corvinus University of Budapest, Budapest 1093, Hungary. [2]Agglomeration and Social Networks Lendület Research Group, ELKH Centre for Economic and Regional Studies, Budapest 1097, Hungary. [3]Department of Network and Data Science, Central European University, 1100 Vienna, Austria. [4]Rényi Alfréd Institute of Mathematics, Budapest 1053, Hungary. ✉email: bokanyi.eszter@krtk.hu

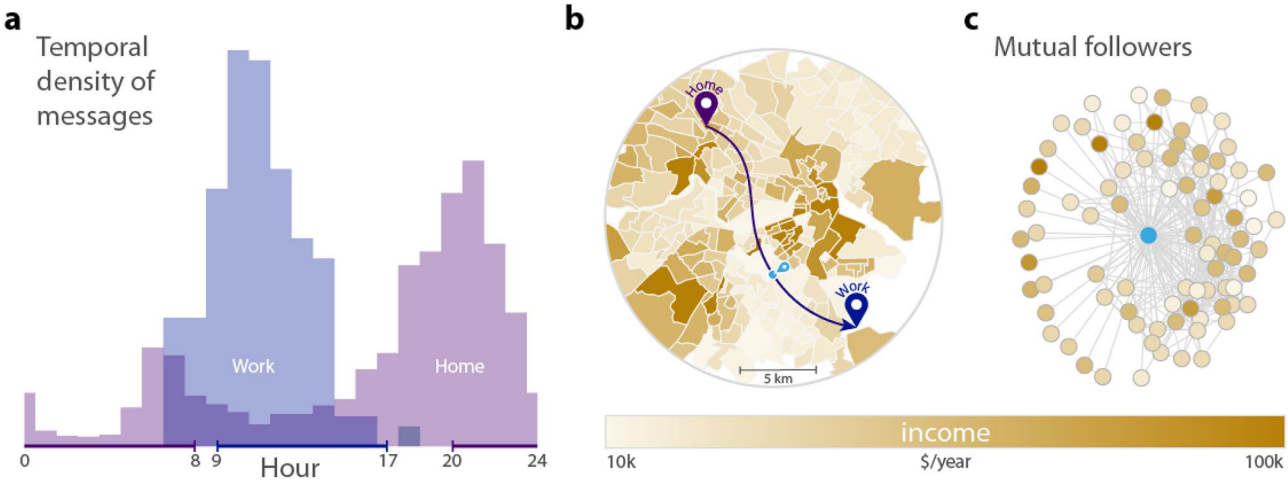

Figure 1. Combination of spatial, temporal and social network data of geolocated Twitter messages. (**a**) Home and work locations of users are identified through the distribution of timestamps on all their collected tweets within their most frequently visited spatial clusters. We assign a possible home location (8 p.m.–8 a.m.) and a possible work location (9 a.m.–5 p.m.) to each user[45,46] as their most frequently visited location in the given period. The histogram represents the timeline of tweets for the clusters of a sample user. (**b**) Commuting is defined as the overhead distance between users' home and work locations. The colorbar of the map indicates the income level of census tracts. Census tract shapes have been downloaded from https://www.census.gov/data/developers/data-sets/acs-5year.html, the figure is the authors' own creation using the `geopandas` library in Python (`geopandas` version 0.6.1, https://pypi.org/project/geopandas/0.6.1/, Python 3.7.2). (**c**) Twitter ego network of a sample user based on mutual followership. The coloring of nodes also corresponds to the level of income in the home tract of users.

data. By comparing ego network indicators between people commuting to different distances, we find that long commuting is associated with lower levels of transitivity, the tendency that friends of friends know each other, and higher levels of income diversity among friends. These results are consistent across the 50 largest US cities and suggest that long commutes can indeed facilitate social mixing.

Our results suggest a universal relation between commuting and integration of disparate social networks. The paper contributes to the discussion on the importance of commuting in cities and shows that longer commutes have a measurable even though moderate influence on establishing diverse and less segregated social connections. The findings imply that supporting access to distant work can help the inclusion of lowest income groups and to a certain degree the richest as well, regardless of the urban context.

Results

We use a unique Twitter database that contains all messages and profile information of 348,850 Twitter users in the top 50 metropolitan areas of the United States. The data was collected between 2012 and 2015 and due to the sample selection method described in[41], the database contains a considerable amount of individuals who allowed automatic GPS data collection for all their messages. This dataset was used in previous research to detect dominant language use and temporal patterns connected to socio-economic indicators such as ethnicity or unemployment in the US, to establish world-wide communities of users reflecting political and cultural boundaries, and to model the spreading of viral content[14,42–44].

Figure 1 illustrates how commuting and social network information is retrieved from the data. Home and work locations are detected by the most frequent locations of tweets in the morning and evening hours or during daytime as depicted in Fig. 1a (and as explained in "Materials and methods" section). This process enables us to identify the census tract of home and work locations and attach socio-economic status, measured by the average household income of census tracts from the 2012 American Community Survey. Commuting is characterized by the Euclidean distance between home and work and the socio-economic status of both locations. Finally, we construct the ego network for every user from mutual followership of Twitter profiles and characterize egos and alters by the socio-economic status of their home location. This enables us to quantify social mixing in terms of commuting and social ties in cities.

Figure 1b shows the census tracts of inner Boston colored by the average annual household incomes and the home and work locations of a sample user. The user's ego network is depicted in panel (c), with colors indicating the income of the neighbors inferred from their home census tract. Each user in our sample has at least 1 mutual followership-based connection and has identifiable home and work locations that are at least 100 m away from each other. The distribution of users across the 50 selected cities is illustrated in Supplementary Information (SI) 1 and 2. For a more detailed description, see "Materials and methods" section.

To characterize the relation between d, the distance of commutes, and the social network of individuals, we compare the social networks of people commuting to $d > median$ with $d < median$ commuting distances in each of the 50 largest US metropolitan areas. Median commuting distances are calculated on the basis of the sampled

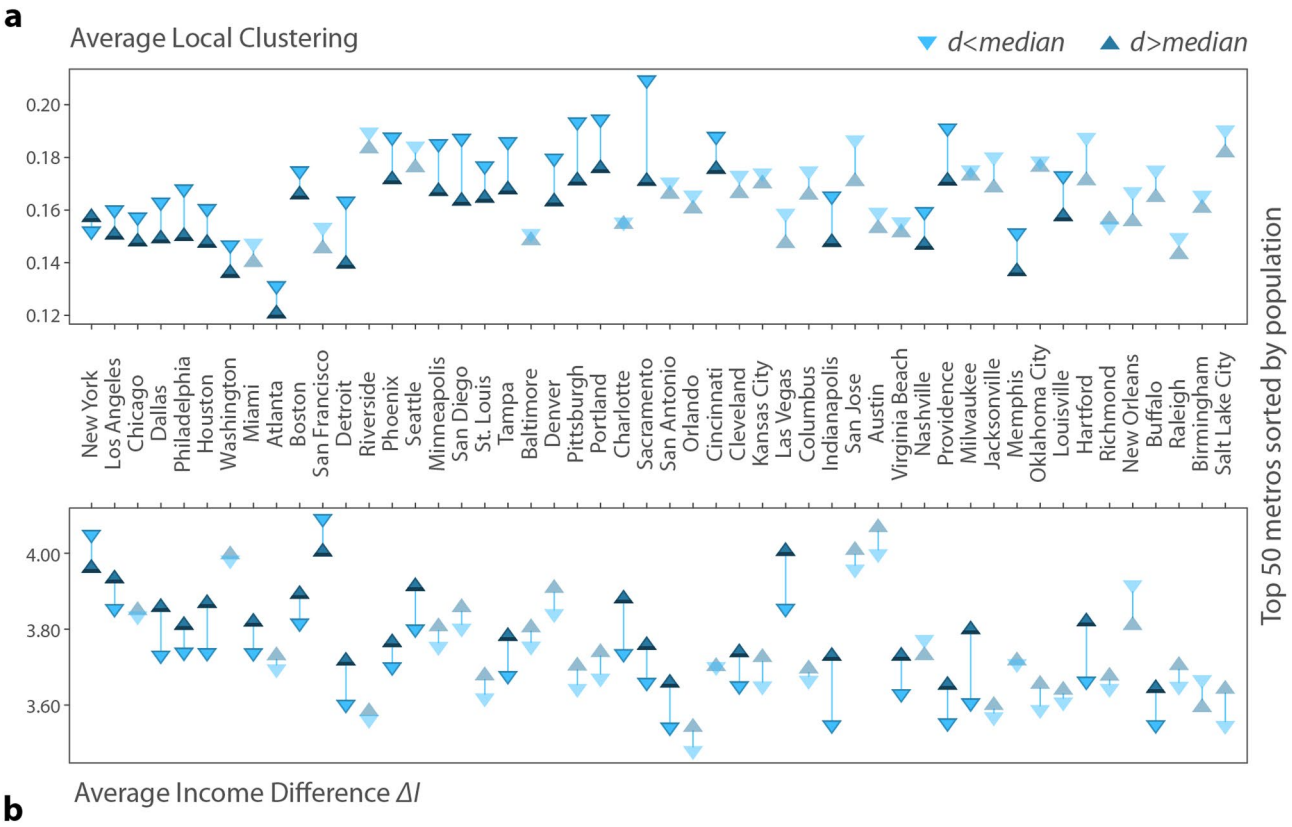

Figure 2. Network characteristics of users and commuting distance in the top 50 metropolitan areas of the United States. (**a**) Network closure measured by the local clustering coefficient is lower in most cities for those users who commute further than the local median distance. (**b**) Income mixture, measured by average income difference from friends, is higher of those who commute above the local median distance in the majority of metropolitan areas. Non-transparent symbols indicate that t-tests suggest significantly different means for the groups ($p < 0.05$).

users in each city as illustrated in SI 3. Our expectation is that commuting may induce more out-of-community independent social ties for commuters, in turn decreasing the transitivity of their egocentric networks. We observe this effect by measuring the local clustering coefficient[47] for each user, which quantifies the tendency that an individuals' friends know each other. Another assumption of ours is that these out-of-community ties introduce stronger diversity in ego networks in terms of socioeconomic status of neighbors. We quantify this effect via the average income difference from friends in users' ego networks, which measures the income similarity of online social connections (for a formal definition see Eq. (1) in "Materials and methods" section).

Figure 2a reports the average of local clustering coefficient and (b) the average income differences of users commuting above and below the local median distance in the 50 largest metropolitan areas in the USA. These findings suggest that, with a few exceptions, an above-median distance commute is associated with lower local clustering (Fig. 2a), and with greater income difference in the commuters' ego networks (Fig. 2b). This implies that working further away from home helps people to develop less cohesive and income-wise more diverse social networks in most metro areas. Note that here metropolitan areas are sorted in decreasing population order and non-transparent markers denote significant differences ($p < 0.05$) between averages.

While these results suggest clear trends, they also highlight the heterogeneity of cities. To support these observations, in SI 4, we compute the degrees for below and above median distance commuters, illustrate the underlying distributions, and we also repeat the measurements and find them to be robust for various distance thresholds. A multivariate regression analysis using continuous variables in SI 5 provides further evidence that commuting distance correlates negatively with local clustering even when controlling for the number of connections and income. These regressions also inform us that commuting distance facilitates mixing in social networks by enabling commuters to make more friendships.

For a more detailed insight into the structure of social and mobility assortativity in these cities, next, we analyze social mixing through commuting and online social ties between income groups. We sort all census tracts into income deciles based on the income distribution across all census tracts in the metro area in question and assign an income decile ranging from 1 to 10 to home and work locations. For each metro area, we construct a commuting assortativity matrix C and a social network assortativity matrix S to represent connection probabilities between these income deciles. The elements of the commuting assortativity matrix C_{ij} measure the probability that a user with home census tract in income decile i commutes to work in a census tract of income decile j. Similarly, elements of the social network assortativity matrix S_{ij} represent the average probability that a person

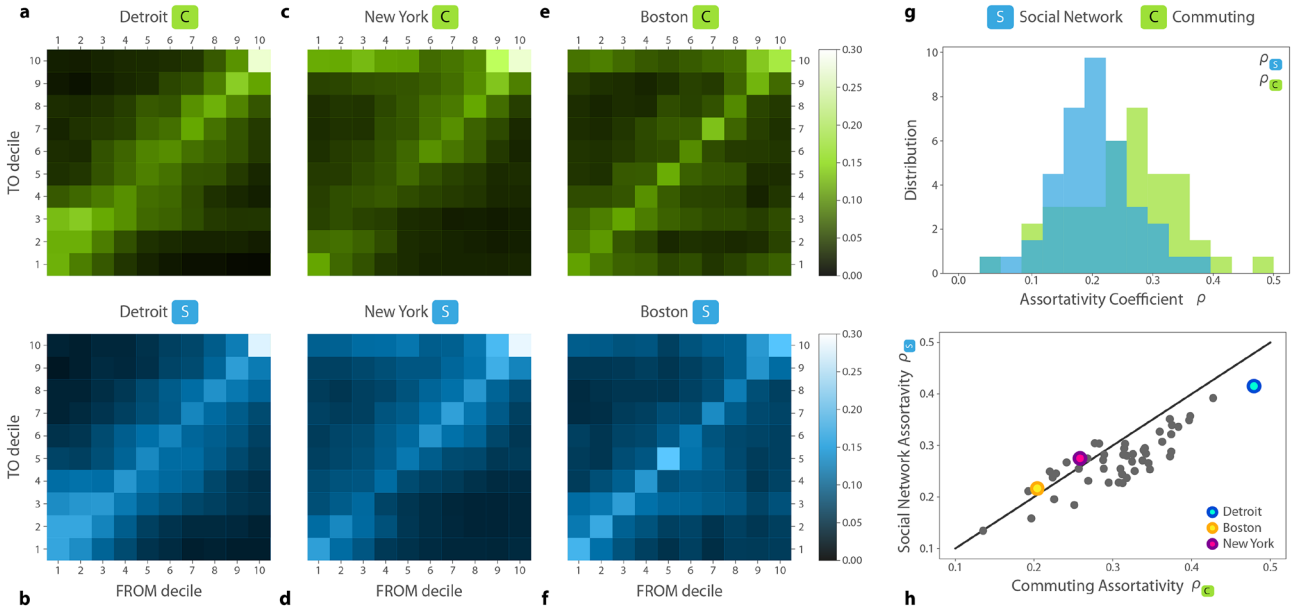

Figure 3. (**a**) Commuting assortativity matrix C and (**b**) social network assortativity matrix S between the 10 income deciles for Detroit, New York (**c**,**d**) and Boston (**e**,**f**). (**h**) Distribution of Pearson correlations ρ_C (green) and ρ_S (blue) for the assortativity matrices C and S of the top 50 metropolitan areas of the US. (**g**) Commuting assortativity and social network assortativity are strongly correlated across cities. Solid line represent $\rho_C = \rho_S$.

living in a tract with income decile i has a mutual followership tie with a user living in a tract with income decile j. For more details on the construction of the matrices, see "Materials and methods" section.

The aggregated patterns of commuting C and friendship ties S are presented in Figures 3a-f for three example metropolitan areas, Detroit, New York, and Boston. Unlike previous studies[23,35], we do not observe universal assortativity patterns over all cities in these networks. In some of the cities, such as Detroit, the strong diagonal component features strong segregation patterns, meaning that people tend to commute to neighborhoods with similar annual household incomes as their home neighborhood, and they tend to form social ties with people living in neighborhoods with similar income, as also found in[24]. In cities like Boston, patterns of mobility and online social ties are less assortative with higher likelihood for diverse, off-diagonal connections. All commuting and social network matrices are available in the SI 6 for the 50 metropolitan areas.

To explore this heterogeneity further, we computed the Pearson correlation coefficient of the above matrices (see "Materials and methods" section Eq. (4)). We use these correlation coefficients as a single-number measure of assortativity in the metropolitan-level networks denoted by ρ_C for the commuting, and ρ_S for the social network assortativity matrix. We show the ρ_C and ρ_S distributions in Fig. 3g. We see here that the level of assortativity varies remarkably across the 50 metro areas, but judging by their averages, commuting in metro areas ($\bar{\rho}_C = 0.31 \pm 0.07$) are more income assortative than online social ties ($\bar{\rho}_S = 0.27 \pm 0.05$). Interestingly, our observations in Fig. 3a–f further suggest that the measured commuting and social network assortativity matrices are not independent from each other. Indeed, Fig. 3h illustrates that ρ_C and ρ_S pairs are strongly correlated (ρ=0.84) suggesting a substantial relationship that social networks are segregated in cities where home-work commuting patterns are assortative.

To investigate the association between long-distance commute and social mixing on the aggregate city-level in more detail, we separate the baseline sample of the C and S matrices by commuting distance. Thus, we create a C and S matrix from users commuting to a distance $d < median$ and $d > median$, as in the example in Fig. 4a–d, where we show these four matrices (two for both C and S) for Detroit. These matrices indicate that for users commuting an above median distance, matrices are less diagonal, and reflect more diverse and less segregated commuting and social connections. Panels (e) and (f) from Fig. 4 present the distributions of ρ_C and ρ_S for the two subgroups of users in all 50 metropolitan areas. As expected, longer commuting distance is associated with less assortativity because distant workplaces are likely to be located in socio-economically different environments as compared to home location. This might be due to spatial clustering of tracts with similar annual household incomes[12], leading to shorter commute patterns landing in places with similar income level. In parallel, we observe that longer commutes are also associated with lower levels of assortativity of online social network ties such that off-diagonal social ties are relatively more likely for $d > median$ distances than for $d < median$. However, while ρ_C falls sharply for $d > median$ distances compared to $d < median$, the difference of ρ_S is moderate in Fig. 4e,f. This finding indicates that although long-distance commutes can link disparate neighborhoods, not all of the diversity generated by commuting has imprints on social connections. Instead, income homophily remains a major yet weaker factor of social tie selection for long commuters as well.

Despite the heterogeneity of metro areas, results in Fig. 4g show general patterns in two regards. First, assortativity of both commuting and social networks are lower for long-distance commuters in every metropolitan area. Second, the assortativity reduction between shorter and longer than median commutes is decreasing sharply,

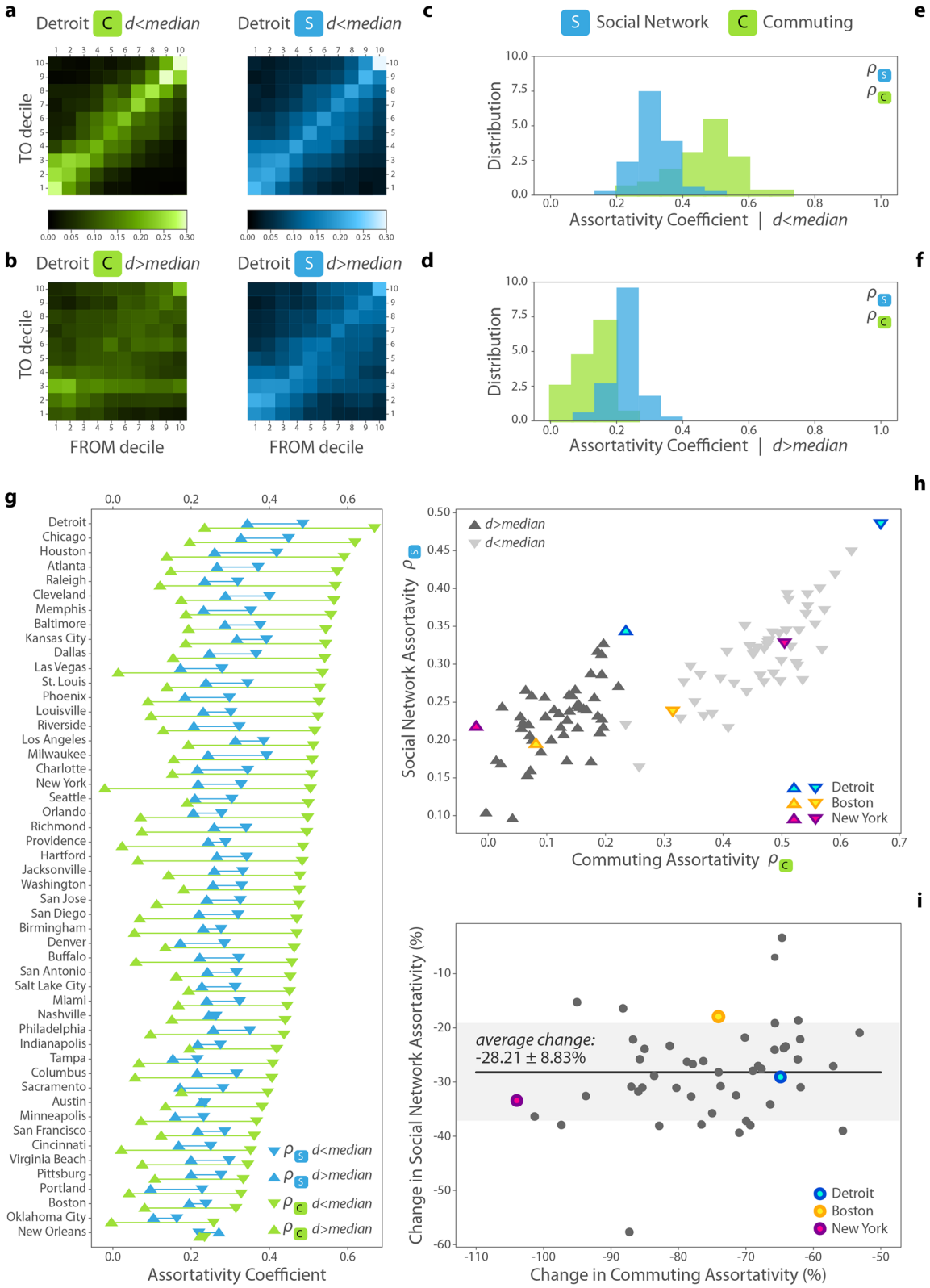

Figure 4. Panels (**a**–**d**) show the *C* and *S* assortativity matrices for the below ($d < median$) and above median ($d > median$) commuting users in a selected metropolitan area, Detroit. (**e**,**f**) The corresponding distributions of ρ_C (green) and ρ_S (blue) for all 50 metropolitan areas for users with $d > median$ and $d < median$. (**g**) Pairwise values of ρ_C and ρ_S for users with $d > median$ and $d < median$ by metropolitan areas. Metropolitan areas are sorted in decreasing order by ρ_C for easier representation. (**h**) Social network assortativity versus commuting assortativity for below and above median commuters with selected cities from Fig. 3 labeled. (**i**) Decrease in the commuting assortativity and the social network assortativity measured in percentage. Black horizontal line corresponds to the average change in social network assortativity. Grey shaded area marks the standard deviation.

while the reduction of social network assortativity is moderate and takes similar values for every metropolitan area. The Pearson correlation coefficient between the two assortativity values ρ_S and ρ_C is 0.80 for short commuters and 0.72 for long commuters, thus they signify a strong relationship between mobility and social network assortativity patterns for both user groups (Fig. 4h). To understand the magnitudes of change, we calculate the percentage of social network assortativity reduction by $((\rho_{S,d>median} - \rho_{S,d<median})/\rho_{S,d<median})$ and the percentage of commuting assortativity reduction by $((\rho_{C,d>median} - \rho_{C,d<median})/\rho_{C,d<median})$ for each city. Illustrating these metrics, Fig. 4i shows that the decrease in commuting assortativity ranges on a wide scale between − 50 and − 100%. However, the decrease in the social network assortativity concentrates around the average value of $-28 \pm 9\%$. Remarkably, this signals a universal pattern of social mixing potentials across very different urban settings and it explains a general trend of how mixing through commutes manifests in social inclusion. SI 7 illustrates that the uniform $\sim$ 30% decrease disappears if we separate two user groups randomly instead of by commuting distance, but this observation remains consistent across multiple absolute distance thresholds (3 km, 5 km, and 10 km). In addition, in SI 8, we show that assortativity reduction by long-distance commute is a result of increasing social mixing of users from poorest and to some extent, from the richest neighborhoods.

Discussion

Understanding the complex behavioral patterns of people is crucial to develop more liveable, equal and sustainable urban environments. Our study contributes to this challenge by using large-scale geolocated Twitter data to study the role of commuting in the composition and assortativity of social interaction. We illustrate that long-distance commuting acts against structural closure and income homophily of social relationships and reduces segregation between remote income classes by facilitating connections and mixing. We show that home-work commutes and online social ties are not equally assortative in every metropolitan area, but in most cases, commuting is even more likely to point to places with similar income level than online social connections. Our findings suggest that longer commutes are more likely to connect places with different income levels, which contributes to the development of more diverse and less assortative social ties. Moreover, working further away from home results in more heterogeneous social connections in every metropolitan area.

Our results suggest that urban mobility has a fundamental role in fulfilling the promise of social inclusion and reduction of social segregation in cities. The association between commuting distance and social networks is remarkably stable across all metropolitan areas with different size and spatial structure[48]. This universal pattern highlights that commuting-enabled social mixing follows similar mechanisms regardless of the urban context. We find that facilitating the access between distant neighborhoods can reduce segregation in metropolitan areas, while gains in social inclusion are limited to a 30% reduction of assortativity. These results signal that providing access across disparate neighborhoods cannot erase mechanisms of social network segregation but can mitigate the divide between rich and poor.

The methodology applied in this paper could easily be extended to other cities with large populations of geolocated Twitter users, and where granular census data with similar spatial resolution is available. However, this approach is not without limitations. While we are confident in our approach to identify home and work locations of users, we cannot confirm whether the identified work locations are actual workplaces or any other facility that people visit frequently during daytime (such as restaurants, schools, etc.). We measure commuting distances as the Euclidean distance between the home location and the work location, whereas in multiple cities, physical obstacles such as rivers might considerably increase travel times or change the socio-economic segregation patterns of settlements[49]. We are not aware of the available modalities to reach work destinations, but we admit that it would also introduce a large variability into travel times. We choose this simplification because both travel times with a car or public transportation might depend on the exact time of the day and varying traffic conditions. Both the underestimation of commuting distances and the inclusion of users who might not have a regular workplace can result that the observed commuting in our case (see SI Figure 3) falls behind the commuting distances reported in the American Community Survey.

Because we do not use an absolute threshold to distinguish long and short commutes, and we use the city-wise median to divide the users into categories, we believe that the aforementioned biases do not affect our results drastically. However, we test both the results of Figs. 2 and 4i for different absolute distance thresholds, 3 km, 5 km and 10 km, where our results still hold (see SI 4 and SI 7).

Even though the fraction of users present in the analysis is proportional to the population size of the 50 metropolitan areas (see SI Figure 2), we have to highlight that our dataset is not fully representative for the US population and results have to be interpreted accordingly. Hargittai and Litt[50] finds that African American users are overrepresented on the platform, and Twitter users are predominantly young, well-educated[51,52] and unrepresentative of other ethnicities[53,54]. Therefore, we cannot generalize our findings to the whole population of these metropolitan areas. Another limitation of the study could be that the free 1% sample from Twitter Streaming API was used for the initial data collection. Joseph et al. and Morstatter et al.[55,56] confirms that tweets filtered to containing GPS coordinates are retrieved to almost 90% of the time compared to the full dataset. By imposing strict count limits, spatio-temporal constraints and mutual followership for ties, we believe that our sample is less distorted from bot activity than what[57] would suggest.

Despite the imperfection of the data, we believe that the presented exercise offers useful insights to the structure of social connections within urban areas. Such large-scale, micro-level analysis enables us to uncover the fundamental patterns behind segregation, inequality or the lack of inclusion inside cities. Publicly available online social network data can complementing official census reports or surveys and can provide opportunities to detect and react to societal patterns and changes.

Materials and methods

Data collection and combination methods. We focus on users of the online social networking site of Twitter who posted tweets frequently containing precise geographic information. More specifically, we use a unique, historical database rich in tweets containing GPS coordinates[41,58]. These tweets originate from users who enabled the exact geolocation option on their smartphones. Overall, we detect the three most frequent tweeting locations of users as spatial clusters of their locations in the 50 most populated metropolitan areas of the United States. We use the Friend-of-Friend algorithm[59] to cluster the spatial coordinates for each user. This algorithm is a paralellizable, scalable clustering algorithm known from astronomy, and it is widely used to identify galaxy clusters[60]. In our case, any two tweet coordinates of the same person are considered to belong to the same spatial cluster if their separation is less than 1 km. For each cluster, we determine the first two moments of the coordinate distribution. Before calculating the mean coordinates of the cluster, we trim data points until all points are inside a 3σ radius to eliminate outliers. We keep the aforementioned three highest cardinality clusters per user[41,42].

To determine the possible home and work locations of users, we follow the approach proposed by[46]. We assume that the home and work locations of users are within the previously detected three clusters. We select users for whom at least two out of the three clusters are within the same metropolitan area from the top 50 metropolitan areas of the United States and one of these clusters is their top cardinality location. First, we calculate the daily timeline of clusters for each user based on the timestamp of the tweets with hourly aggregation, converting all UTC tweet timestamps to local times across the whole US. We only consider users with more than 15 tweets on weekdays (Monday to Friday) in total. Local aggregated weekday timelines of two clusters for a sample user are presented in Fig. 1a. We calculate the share of tweets sent between 9 a.m. and 5 p.m. on weekdays to capture messages predominantly sent during the working hours. Similarly, we calculate the share of tweets sent between 8 p.m. and 8 a.m. on weekdays contributing to a possible home tweeting fraction. Then, the cluster with the highest work tweet share or home tweet share becomes the work and home cluster of the user.

Commuting of users is characterized by the overhead distance between their home and work locations. We restricted our sample to users with at least 0.1 km commutes to avoid those ambiguous cases where detected home and work clusters are the same. Thus, we have 975,492 users in our sample. The distribution of observed commuting distances for each metro area are presented in SI 3. Additionally, we attach socio-economic data to each home and work location in the observed metropolitan areas from the 2012 American Community Survey. More precisely, we map the home locations of users into the census tracts of the top 50 US metropolitan areas and attribute the average annual household income of the census tract to each user living there. After that, we sort users into city-wise income deciles based on the average annual household incomes, and we apply the same approach to determine the average income and the income decile of their workplaces. Figure 1b shows the commute of the same sample user and the income level of the surrounding census tracts.

Social connections of users are defined as their mutual followership relations on Twitter as they represent relative stronger ties in context of online social networks[61]. Figure 1c represents a sample ego network that we construct for every user from our home-work sample who has at least 1 mutual followership tie within the same metropolitan area. In the end, we have 348,850 users for whom we have both the home and work location information, and a mutual followership ego network. The composition and spatial distribution of our final sample is presented in SI 1. Through the home location of the user's friends, we can infer their income, thus, we are able to characterize the socio-economic status of the neighbors in the ego networks by identifying their income deciles. Figure 1c shows this characterization by using the same colorscale for both the ego and its first neighbors as the choropleth map in Fig. 1b.

At the individual level, commuting and online social ties of our users are characterized by multiple different indicators. We measure user commutes by the Euclidean distance d between their inferred home and work locations. We calculate degree and local clustering coefficient from their ego networks. We also measure the average income difference between their own home income and the home income of their friends, following the formula below:

$$\Delta I = \frac{1}{\#\text{neighbors}} \sum_{f \in \text{neighbors}} \log_{10} \left| I_f - I_{ego} \right| \tag{1}$$

Assortativity metrics. At the aggregated, metropolitan area level, we create multiple different assortativity matrices between income deciles D for each metropolitan area. First, an assortativity matrix of commuting is constructed, where we capture the probability C_{ij} that a user u belonging to a home census tract in income decile $D = i$ commutes to a tract with income decile $D = j$ to work. Second, we measure the conditional probabilities of social ties across home census tracts in different income deciles, the social network assortativity matrix S. The element S_{ij} of this matrix measures the probability that a user u from income decile $D = i$ has a mutual followership tie to a user in income decile $D = j$. Formally, the two matrices can be calculated as

$$C_{ij} = \frac{\sum_{\{u \in U | D_{u,\text{home=j}}, D_{u,\text{work=i}}\}} 1}{\sum_{\{u \in U | D_{u,\text{home=j}}\}} 1} \tag{2}$$

$$S_{ij} = \frac{\sum_{\{u \in U | D_{u,\text{home}=j\}}} \frac{1}{k_u} \sum_{\{e_{uf} \in E_u | D_{f,\text{home=i}}\}} 1}{\sum_{\{u \in U | D_{u,\text{home=j}}\}} 1}, \tag{3}$$

where U is the user set within a metropolitan area for which we calculate the matrices, E_u is the set of edges connected to the user u, k_u is the degree of ego user u in the ego network, e_uf is the undirected edge between user u and f, D_u and D_f are the (home or work) deciles of users u and f, respectively. We also measure two additional friendship and commuting assortativity matrices, $S^{d>\text{median}}$, $S^{d<\text{median}}$, $C^{d>\text{median}}$ and $C^{d<\text{median}}$, for users commuting more or less than the median commute in the given metropolitan area. In these cases, the set U is what is different in the matrices from Eq. (3).

We measure assortativity in these matrices by calculating the Pearson correlation coefficient ρ of the matrix entries. If we normalize the elements of matrix X such that $\tilde{X}_{ij} = X_{ij}/n$, where $n = \sum_{i,j} X_{ij}$, the sum of the elements of a matrix, then ρ captures how diagonal these matrices are:

$$\rho_X = \frac{\sum_{i,j} ij\tilde{X}_{ij} - \sum_{i,j} i\tilde{X}_{ij} \sum_{i,j} j\tilde{X}_{ij}}{\sqrt{\sum_{i,j} i^2\tilde{X}_{ij} - \left(\sum_{i,j} i\tilde{X}_{ij}\right)^2}\sqrt{\sum_{i,j} j^2\tilde{X}_{ij} - \left(\sum_{i,j} j\tilde{X}_{ij}\right)^2}}, \tag{4}$$

where the summation for i and j both go over all of the income deciles $D = 1, \ldots, 10$. An assortativity value $\rho = +1$ would mean a completely diagonal, thus, completely assortative matrix, whereas $\rho \approx 0$ values indicate the lack of any preference for people following others from the very same income class of their own.

Received: 20 May 2021; Accepted: 8 October 2021
Published online: 21 October 2021

References

1. Jacobs, J. *The Death and Life of Great American Cities* (Vintage, 2016).
2. Glaeser, E. Cities, productivity, and quality of life. *Science* **333**(6042), 592–594 (2011).
3. Bettencourt, L. M. A. The origins of scaling in cities. *Science* **340**(6139), 1438–1441. https://doi.org/10.1126/science.1235823 (2013).
4. Duranton, G. & Puga, D. The economics of urban density. *J. Econ. Perspect.* **34**(3), 3–26. https://doi.org/10.1257/jep.34.3.3 (2020).
5. Storper, M. & Venables, A. J. Buzz: Face-to-face contact and the urban economy. *J. Econ. Geogr.* **4**(4), 351–370. https://doi.org/10.1093/jnlecg/lbh027 (2004).
6. Calabrese, F. *et al.* Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE* **6**(7), e20814. https://doi.org/10.1371/journal.pone.0020814 (2011).
7. Chong, S. K. *et al.* Economic outcomes predicted by diversity in cities. *EPJ Data Sci.* **9**(1), 17. https://doi.org/10.1140/epjds/s13688-020-00234-x (2020).
8. Sampson, R. J. Moving to inequality: Neighborhood effects and experiments meet social structure. *Am. J. Sociol.* **114**(1), 189–231. https://doi.org/10.1086/589843 (2008).
9. Glaeser, E. L., Resseger, M. & Tobio, K. Inequality in cities. *J. Reg. Sci.* **49**(4), 617–646. https://doi.org/10.1111/j.1467-9787.2009.00627.x (2009).
10. Florida, R. & Mellander, C. *Segregated City: The Geography of Economic Segregation in America's Metros* (Martin Prosperity Institute, 2015).
11. Ananat, E. O. The wrong side(s) of the tracks: The causal effects of racial segregation on urban poverty and inequality. *Am. Econ. J. Appl. Econ.* **3**(2), 34–66. https://doi.org/10.1257/app.3.2.34 (2011).
12. Chodrow, P. S. Structure and information in spatial segregation. *Proc. Natl. Acad. Sci. USA* **114**(44), 11591–11596. https://doi.org/10.1073/pnas.1708201114 (2017).
13. Fry, R. & Taylor, P. *The Rise of Residential Segregation by Income* (Pew Research Center, 2012).
14. Bokányi, E. *et al.* Race, religion and the city: Twitter word frequency patterns reveal dominant demographic dimensions in the United States. *Palgrave Commun.* **2**(1), 16010. https://doi.org/10.1057/palcomms.2016.10 (2016).
15. Massey, D. S. & Denton, N. A. The dimension of residential segregation. *Soc. Forces* **67**(2), 281–315 (1988).
16. Eagle, N., Pentland, A. & Lazer, D. Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci. USA* **106**(36), 15274–15278. https://doi.org/10.1073/pnas.0900282106 (2009).
17. Bailey, M. *et al.* Social connectedness in urban areas. *J. Urban Econ.* **118**, 103264. https://doi.org/10.1016/j.jue.2020.103264 (2020).
18. Norbutas, L. & Corten, R. Network structure and economic prosperity in municipalities: A large-scale test of social capital theory using social media data. *Soc. Netw.* **52**, 120–134. https://doi.org/10.1016/j.socnet.2017.06.002 (2018).
19. Abitbol, J. L. & Karsai, M. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nat. Mach. Intell.* **2**(11), 684–692 (2020).
20. Tóth, G. *et al.* Inequality is rising where social network segregation interacts with urban topology. *arXiv* **12**(1), 1–9. https://doi.org/10.1038/s41467-021-21465-0 (2019).
21. Wang, Q. *et al.* Urban mobility and neighborhood isolation in America's 50 largest cities. *Proc. Natl. Acad. Sci.* **115**(30), 7735–7740. https://doi.org/10.1073/pnas.1802537115 (2018).
22. Pappalardo, L. *et al.* Using big data to study the link between human mobility and socioeconomic development. In *Proceedings—2015 IEEE International Conference on Big Data, IEEE Big Data 2015* 871–878. IEEE. https://doi.org/10.1109/BigData.2015.7363835 (2015).
23. Dong, X. *et al.* Segregated interactions in urban and online space. *EPJ Data Sci.* **9**(1), 20. https://doi.org/10.1140/epjds/s13688-020-00238-7 (2020).
24. Heine, C. *et al.* Analysis of mobility homophily in Stockholm based on social network data. *PLoS ONE* **16**(3), 1–14. https://doi.org/10.1371/journal.pone.0247996 (2021).
25. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**(5981), 1029–1031. https://doi.org/10.1126/science.1186605 (2010).
26. Jiang, S. *et al.* The TimeGeo modeling framework for urban motility without travel surveys. *Proc. Natl. Acad. Sci. USA* **113**(37), E5370–E5378. https://doi.org/10.1073/pnas.1524261113 (2016).
27. Dahlin, E., Kelly, E. & Moen, P. Is work the new neighborhood? Social ties in the workplace, family, and neighborhood. *Sociol. Q.* **49**(4), 719–736. https://doi.org/10.1111/j.1533-8525.2008.00133.x (2008).
28. Calabrese, F. *et al.* Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE* **6**(7), e20814. https://doi.org/10.1371/journal.pone.0020814 (2011) (**Ed. by E. Scalas**).
29. Small, M. L. & Adler, L. The role of space in the formation of social ties. *Ann. Rev. Sociol.* **45**, 111–132. https://doi.org/10.1146/annurev-soc-073018-022707 (2019).

30. Viry, G. Residential mobility and the spatial dispersion of personal networks: Effects on social support. *Soc. Netw.* **34**(1), 59–72. https://doi.org/10.1016/j.socnet.2011.07.003 (2012).
31. Blumenstock, J., Chi, G. & Tan, X. Migration and the value of social networks (2019).
32. Roberto, E. The spatial proximity and connectivity method for measuring and analyzing residential segregation. *Sociol. Methodol.* **48**(1), 182–224. https://doi.org/10.1177/0081175018796871 (2018).
33. van Ham, M., Tammaru, T. & Janssen, H. J. A multi-level model of vicious circles of socioeconomic segregation. *Divided Cities* **615159**(8774), 135–153. https://doi.org/10.1787/9789264300385-8-en (2018) (**OECD**).
34. Nieuwenhuis, J. *et al.* Does segregation reduce socio-spatial mobility? Evidence from four European countries with different inequality and segregation contexts. *Urban Stud.* **57**(1), 176–197. https://doi.org/10.1177/0042098018807628 (2020).
35. Morales, A. J. *et al.* Segregation and polarization in urban areas. *R. Soc. Open Sci.* **6**(10), 190573. https://doi.org/10.1098/rsos.190573 (2019).
36. Florez, M. A., *et al.* Measuring the impacts of economic well being in commuting networks|A case study of Columbia. In *Transportation Research Board, 96th Annual Meeting*, Vol. 17 03745 (2016).
37. Dannemann, T., Sotomayor-Gómez, B. & Samaniego, H. The time geography of segregation during working hours. *R. Soc. Open Sci.* **5**(10), 180749. https://doi.org/10.1098/rsos.180749 (2018).
38. Bora, N., Chang, Y.-H. & Maheswaran, R. Mobility patterns and user dynamics in racially segregated geographies of US cities. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* 11–18. https://doi.org/10.1007/978-3-319-05579-4_2 (2014).
39. Leo, Y. *et al.* Socioeconomic correlations and stratification in social-communication networks. *J. R. Soc. Interface* **13**(125), 20160598. https://doi.org/10.1098/rsif.2016.0598 (2016).
40. Yip, N. M., Forrest, R. & Xian, S. Exploring segregation and mobilities: Application of an activity tracking app on mobile phone. *Cities* **59**, 156–163. https://doi.org/10.1016/j.cities.2016.02.003 (2016).
41. Dobos, L. *et al.* A multi-terabyte relational database for geo-tagged social network data. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)* 289–294. IEEE. https://doi.org/10.1109/CogInfoCom.2013.6719259 (2013).
42. Kallus, Z. *et al.* Spatial fingerprints of community structure in human interaction network for an extensive set of large-scale regions. *PLoS ONE* **10**(5), e0126713. https://doi.org/10.1371/journal.pone.0126713 (2015) (**Ed. by B. Jiang**).
43. Kallus, Z. *et al.* Video pandemics: Worldwide viral spreading of Psy's Gangnam Style Video. In *ICT Innovations 2017: Data-Driven Innovation*, Vol. 778 (eds Trajanov, D. & Bakeva, V.) 3–12. (Springer, 2017). https://doi.org/10.1007/978-3-319-67597-8_1.
44. Bokányi, E., Lábszki, Z. & Vattay, G. Prediction of employment and unemployment rates from Twitter daily rhythms in the US. *EPJ Data Sci.* **6**(1), 14. https://doi.org/10.1140/epjds/s13688-017-0112-x (2017).
45. Lambiotte, R. *et al.* Geographical dispersal of mobile communication networks. *Physica A* **387**(21), 5317–5325. https://doi.org/10.1016/j.physa.2008.05.014 (2008).
46. McNeill, G., Bright, J. & Hale, S. A. Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Sci.* **6**(1), 24. https://doi.org/10.1140/epjds/s13688-017-0120-x (2017).
47. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world9 networks'. *Nature* **393**(6684), 440–442. https://doi.org/10.1038/30918 (1998).
48. Boeing, G. Urban spatial order: Street network orientation, configuration, and entropy. *Appl. Netw. Sci.* **4**(1), 67. https://doi.org/10.1007/s41109-019-0189-1 (2019).
49. Tóth, G. *et al.* Inequality is rising where social network segregation interacts with urban topology. *Nat. Commun.* **12**(1), 1143. https://doi.org/10.1038/s41467-021-21465-0 (2021).
50. Hargittai, E. & Litt, E. The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New Media Soc.* **13**(5), 824–842. https://doi.org/10.1177/1461444811405805 (2011).
51. Webster, T. Twitter usage in America: 2010. In *Edison Research/Arbitron Internet and Multimedia Study* (2010).
52. Sloan, L. *et al.* Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user metadata. *PLoS ONE* **10**(3), e0115545. https://doi.org/10.1371/journal.pone.0115545 (2015) (**Ed. by T. Preis**).
53. Mislove, A. *et al.* Understanding the demographics of Twitter users. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)* 554–557 (2011).
54. Malik, M. M. *et al.* Population bias in geotagged tweets. In *AAAI Workshop—Technical Report WS-15-18* 18–27 (2015).
55. Joseph, K., Landwehr, P. M. & Carley, K. M. Two 1%s don't make a whole: Comparing simultaneous samples from Twitter's streaming API. In *Association of the Advanced of Artificial Intelligence* 75–83 (2014). https://doi.org/10.1007/978-3-319-05579-4_10.
56. Morstatter, F., Pfeffer, J. & Liu, H. When is it biased? In *Proceedings of the 23rd In ternational Conference on World Wide Web—WWW '14 Companion* 555–556 (ACM Press, 2014). https://doi.org/10.1145/2567948.2576952.
57. Pfeffer, J., Mayer, K. & Morstatter, F. Tampering with Twitter's sample API. *EPJ Data Sci.* **7**(1), 50. https://doi.org/10.1140/epjds/s13688-018-0178-0 (2018).
58. Kondor, D. et al. Effcient classification of billions of points into complex geographic regions using hierarchical triangular mesh. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management—SSDBM '14* 1–4 (ACM Press, 2014). https://doi.org/10.1145/2618243.2618245.
59. Huchra, J. P. & Geller, M. J. Groups of galaxies. I—Nearby groups. *Astrophys. J.* **257**, 423. https://doi.org/10.1086/160000 (1982).
60. Kwon, Y. *et al.* Scalable clustering algorithm for N-body simulations in a shared-nothing cluster. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6187 LNCS* 132–150 (2010). https://doi.org/10.1007/978-3-642-13818-8_11.
61. Szüle, J. *et al.* Lost in the city: Revisiting Milgram's experiment in the age of social networks. *PLoS ONE* **9**(11), e111973. https://doi.org/10.1371/journal.pone.0111973 (2014).

Acknowledgements

Eszter Bokányi was supported by the ÚNKP-20-4 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund of Hungary. Márton Karsai acknowledges support from the H2020 SoBigData++ project (H2020-871042) and the DataRedux ANR project (ANR-19-CE46-0008). Balázs Lengyel and Sándor Juhász acknowledge support from the Hungarian Scientific Research Fund (OTKA K-138970). We thank for the usage of ELKH Cloud (https://science-cloud.hu/) that significantly helped us achieving the results published in this paper. We thank József Stéger for helping in the maintenance of the Twitter database, and Szabolcs Tóth-Zs. (https://bandart.eu/) for figure design.

Author contributions

E.B. and S.J. analyzed the data and prepared the figures. All authors took part in designing the study, and writing the manuscript. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-00416-1.

Correspondence and requests for materials should be addressed to E.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.