Connecting Digital and Physical Segregation

Do Online Activity and Social Networks Mirror Residential Patterns?

Joan Wang

Graham MacDonald

Solomon Greene

February 2018

The authors welcome feedback on this working paper. Please send all inquiries to gmacdonald@urban.org.

URBAN INSTITUTE

2100 M Street NW
Washington, DC 20037
www.urban.org

Abstract

This paper seeks to explore the connection between segregation in digital and physical spaces. We perform an exploratory analysis using a sample of geocoded Twitter data for the metropolitan area of Chicago to address two specific questions:

1. Are the spatial mobility patterns of Twitter users and their tweets similar to the patterns of physical segregation along the lines of race, income, and education?
2. Are Twitter users' friend networks related to the demographic characteristics of the neighborhoods in which they reside?

Our results suggest that although the spatial movements of Twitter users tend to align with known residential segregation patterns in Chicago, the friend networks of these users do not seem to be related to these demographic patterns in any obvious way. We also use descriptive statistics to investigate the potential bias in our Twitter sample, and find that, on average, higher-income, highly educated, and whiter neighborhoods tend to be overrepresented.

Key words: Twitter, social network analysis, social media, segregation

Copyright © February 2018. Joan Wang, Graham MacDonald, and Solomon Greene. All rights reserved.

Introduction

Physical segregation, or the separation of groups of people into defined physical spaces, has been studied thoroughly over the past few decades. This body of research has documented large disparities in the concentrations of people by race, ethnic group, and class that persist in most large urban areas today (Massey and Denton 1987; Rivkin 1994; Iceland, Weinberg, and Steinmetz 2002; Acs et al. 2017).

Digital segregation, or the separation of groups of people into defined online spaces and information "communities," has been less studied although the proliferation of data from online social networks has led to new empirical work on digital segregation in recent years (Smith 2017). Recent research has found, among other things, differential access to and use of the internet by race, income, and other demographic factors (Hoffman and Novak 1998; Hargittai and Hinnant 2016). In addition, the concept of homophily, or the tendency of individuals to associate with people of similar racial or ethnic groups in online spaces, has emerged (Thelwall 2009; Hofstra et al. 2017; Wimmer and Lewis 2010; Cesare et al. 2017).

Yet, despite the research interest in both physical and digital segregation, and the meteoric rise in the use of the internet and social media platforms, little effort has been made to study the connection between digital and physical segregation, leaving important questions unanswered. For example, do patterns of physical segregation manifest themselves online? And do residents of segregated communities face similar segregation in digital spaces?

In this paper, we begin to examine these questions at the intersection of digital and physical segregation, by combining "big data" from online social networks matched with traditional demographic data from the Census. This analysis is meant to be exploratory, and therefore before discussing our results, we provide a descriptive analysis of the potential bias in the data we use. We acknowledge that the data we use have limitations, additional research is necessary, and that researchers in general need to better understand the bias in these data sources to produce more reliable results going forward.

Literature Review

Limited research has been conducted on the connection between digital and spatial segregation, but a few recent studies have explored some aspects of the two in tandem.

Early studies connecting digital and physical segregation focus on the "digital divide," that is, the gap between those who do and do not have access to digital technology and the Internet. These studies explore the physical aspects of the digital divide by comparing Internet access in rural and urban areas (Stern, Adams, and Elsasser 2009) or examining the importance of the number of access points (Hassani 2006). These studies suggest that differentiated access to the internet is influenced by and could further contribute to social inequities.

Other researchers take a sociological approach to studying the connection between digital and physical segregation. Historically, researchers have noted two non-mutually exclusive factors contributing to segregation: structural constraints, such as redlining or racial steering, and individual agency based on personal preference (Cesare et al. 2017). Structural constraints are more difficult to demonstrate in the context of online social interactions because of the seemingly "borderless" nature of the Internet, though they do exist. For example, Edelman and colleagues (n.d.) found that hosts on the online rental site Airbnb were 16 percent less likely to accept bookings requests from people with African American names. Most sociological studies of online segregation, however, are centered on the notion of homophily, the idea that individuals tend to associate with others who are similar to themselves. They focus on the manifestation of homophily by race and ethnicity in online social networks such as MySpace, Facebook, and Twitter (Thelwall 2009; Hofstra et al. 2017; Wimmer and Lewis 2010; Cesare et al. 2017).

A few recent studies using large, digital data sources have advanced innovative methods for detecting the relationship between user-generated spatial data and the social characteristics of their locations. Cranshaw and colleagues (2013) investigate the relationship between users' mobility patterns and the structure of their underlying social network, using data from Facebook friendships and a Facebook location-sharing network called Locaccino. They demonstrate that users' location histories can predict friendships and provide insight into the diversity levels of the locations they visit. Another study, called the Livehoods Project (Cranshaw et al. 2010), uses location data to capture social connectedness on a larger scale. Researchers used Foursquare

check-ins to identify venue clusters—what some would call "neighborhoods"—that characterize a city beyond officially defined borders. They also note signs of the digital divide, as low-income, predominately black neighborhoods were underrepresented in their data. Shelton and colleagues (2015) focus their efforts on the socio-spatial relations around the "9th Street Divide," an informal border between the predominantly affluent, whiteEast End and predominantly poor, black West End of Louisville, Kentucky. They use geotagged tweets to observe the mobility patterns of users "belonging" to each area. These papers begin to measure digital and physical segregation by using online data to understand offline phenomena.

In other cases, user-generated spatial data are paired with demographic information to associate race and socioeconomic characteristics with user behavior. Li and colleagues (2013) combine geocoded Twitter and Flickr data with Census data to describe the relationship between placemark densities and the demographic features of local neighborhoods. They focus on California counties and find that well-educated people from certain occupations are more likely to generate geocoded tweets. In a different study of user-generated placemarks in the aftermath of Hurricane Katrina, researchers explored the role of race in the adoption of digital technologies (Crutcher and Zook 2009). The researchers found that neighborhoods with a higher share of African American residents were much less likely be represented in the user-generated data, suggesting a racial divide.

We build upon this body of work in the remainder of the report, using a combination of traditional and nontraditional data and methods to explore additional possibilities for observing the relationship between physical and digital segregation.

Data and Methods

In this study, we explore the relationship between Twitter users' activity patterns and the demographic characteristics of the neighborhoods in which they reside or traverse. We seek to address two narrow research questions:

1. Are the spatial mobility patterns of Twitter users and their tweets similar to the patterns of physical segregation along the lines of race, income, and education?
2. Are Twitter users' friend networks related to the demographic characteristics of the neighborhoods in which they reside?

We use the Chicago Metropolitan Statistical Area (MSA) as a case study. Chicago has been and remains one of the most diverse and segregated urban regions in the country. Previous studies have highlighted the prevalence of physical segregation in Chicago neighborhoods, and a recent report further emphasizes the detrimental effects of segregation in Chicago (Acs et al. 2017; Grabinsky and Reeves 2015; Glaeser and Vigdor 2012).

Within the Chicago MSA, we use census tracts as proxies for neighborhoods. Census tracts are sufficiently granular to capture the demographic particularities of different neighborhoods, but large enough to produce reliable estimates for this analysis.

Data Sources

To understand the demographic makeup of Chicago neighborhoods, we use tract-level data from the 2011–15 American Community Survey (ACS) 5-Year Estimates. We focus on three demographic characteristics: race, family income, and educational attainment (see Appendix A for the list of subcategories that comprise each characteristic). For each of the three demographic characteristics, we calculate the percent distribution for each subcategory——for example, percent with bachelor's degree——within each tract. We limit our analysis of the demographic characteristics of residents to census tracts within Chicago MSA that have nonzero population, households, and adults over 25 years old ("populated census tracts"), which eliminates just 13 of the 2,215 tracts in the metro area, leaving us with 2,202 tracts.

To explore the digital activity of individuals in Chicago, we use a sample of geotagged tweets. Twitter is a popular social media platform that has a publicly accessible API. Its geolocation feature allows users to include geographic information when posting a tweet. Depending on account settings and geolocation method, users may share precise GPS coordinates or a general place. We obtained a random sample of geotagged tweets that were posted between November 2016 and June 2017 (excluding 13 days for which at least some tweets were lost due to technical issues). Since we are interested in connecting tweets to the specific tract from which they were posted, we limit our dataset to only those tagged with exact coordinates from populated census tracts within Chicago MSA, excluding the most common check-in locations with an exceptional number of Tweets, which represent the city of Chicago as a whole and not a specific location within the city. This leaves us with a sample of 1.2 million tweets that serve as the basis for our analysis.

We also obtained the friends and followers lists of Twitter users within our analysis. These provide insight into an individual user's social network, consisting of accounts that the user follows and accounts that follow the user. Approximately 3 percent of users treated as residents of a tract in our dataset (see below) have their account privacy set to protected mode, which prevents public access to their friends and followers lists. These users' data are not included in our analysis.

Definitions

Since a user may tweet from any number of physical locations, we use two methods to assign Twitter users to census tracts.

Twitter User-Tracts

Twitter "user-tracts" are distinct Twitter user and tract combinations. A Twitter user contributes to the unique user-tract count of a particular tract if the dataset contains any posts from that user within that tract. User-tracts do not distinguish between users who are residents, workers, or tourists of a tract. Our dataset contains 246,770 unique Twitter user-tracts, representing 2,198 of the Chicago MSA's 2,215 different tracts (note this is different than the 2,202 tracts used to calculate demographic characteristics only, as described in the previous section).

Twitter Residents

We are ultimately interested in connecting Twitter users to the demographic characteristics of the neighborhood in which they reside. Therefore, we associate each user with a single tract by applying a series of logical restrictions to determine users' likely tract of residence.

We first identify and filter out tweets that are likely to have originated from a user's workplace rather than home. We define "workplace tweets" as those posted during typical business hours from commercial tracts, that is, tracts with a high ratio of jobs relative to residential population. Using 2014 data from the Longitudinal Employer-Household Dynamics (LEHD) program, we calculate the ratio of the number of primary jobs to the total population within each tract. We define a tract as "commercial" if it has a job to population ratio over 1.5; this encompasses about 5 percent of tracts that house about 5 percent of the Chicago MSA population. We are unaware of previous studies that attempted to classify residential tracts using this method, so we compared a number of different thresholds and chose a threshold that would

include at least 95 percent of residents and tracts while eliminating the most potential workplace tweets (see Appendix C for the range of values tested and population excluded). Typical business hours are defined as Monday through Friday between 8:00 am and 6:00 pm local time.

In total, these restrictions categorize about 15 percent of tweets as workplace tweets. Note that this approach fails to capture tweets from individuals working in less commercial areas or during non-typical business hours, and may misclassify tweets from individuals who reside in commercial areas.

A user is considered a "resident" of a tract (called "Tract A" here for instructive purposes) if the following criteria are met:

1. There is at least a ten-day difference between any two non-workplace tweets from Tract A. This prevents us from labeling temporary visitors as residents.
2. Non-workplace tweets from Tract A make up over half of the user's total geotagged tweets (including workplace tweets). This ensures that each user is only assigned to one tract of residence. We use a majority and not a plurality to ensure we capture a more definitive place of residence.
3. The user's friends and followers lists are publicly accessible. This ensures that we can include the user's social network in our analysis.

We further limit the analysis to tracts where three or more users have been assigned as residents. Twitter data are not representative of the underlying population, as we will discuss in the next section, but we impose a minimum number of observations to improve representativeness. The final dataset consists of 2,358 "residents", representing 428 different tracts.

Limitations and Representation

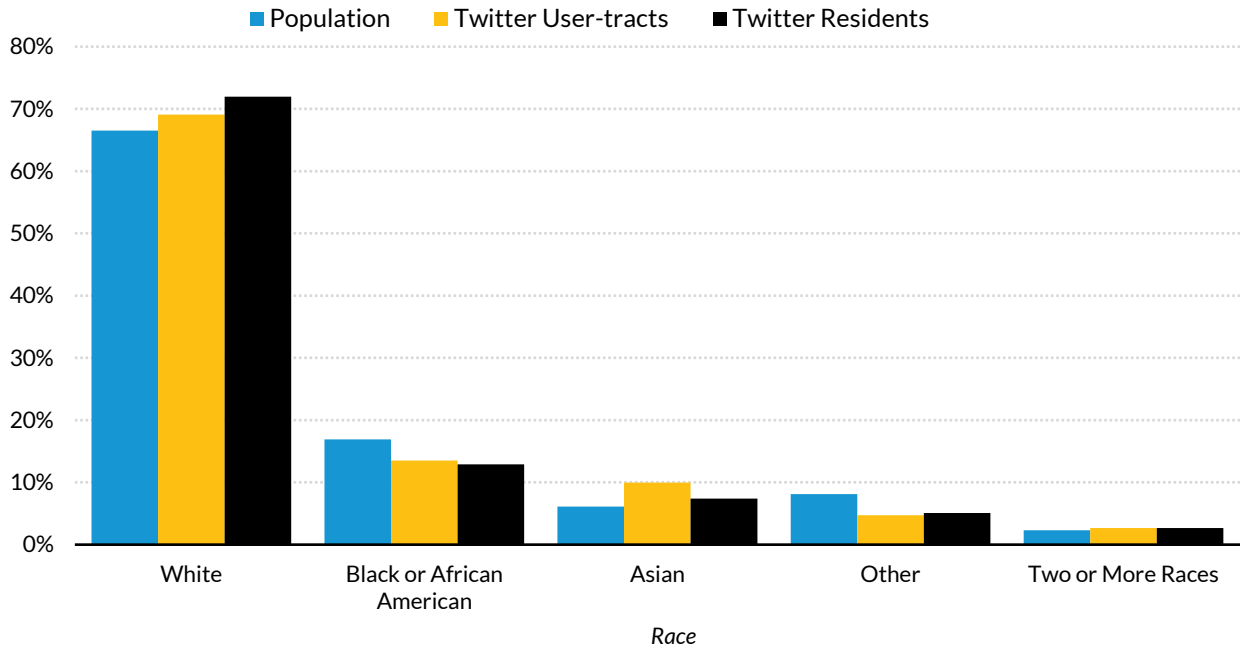
Before presenting the results of our analysis, we must address the crucial limitations to using Twitter data for research focusing on digital and spatial segregation. Unlike many surveys and administrative data, the demographic profile of any particular Twitter user is unknown. Although we can connect geocoded tweets to the census tracts from which they originated and describe the demographic characteristics of those tracts, we cannot determine or infer demographic characteristics of individual users directly. Some research has been conducted on inferring the

age, race, and gender of Twitter users based on tweet text and profile information, but we do not attempt that here (Mislove et al. 2011; Culotta et al. 2015.

There are a number of reasons the Twitter data and analyses that use these data may be skewed toward certain demographic groups. First and foremost, the Twitter user base is not representative of the general population. A 2016 report from Pew Research Center found that only 24 percent of online adults use Twitter, and this audience tends to be younger and more educated (Smith 2017; Pew 2017). Furthermore, the geolocation feature in Twitter is an opt-in service that certain user groups may be more inclined to enable. Previous studies found that only 3 percent of users actually geotagged their tweets, and that geocoded tweets account for less than 1 percent of all tweets (Sloan 2015; Graham 2013). Frequency of activity on Twitter also differs significantly by user. Highly active Twitter users contribute to more user-tract counts, as they are counted each time they tweet from a different tract. Active users are also more likely to meet the residency criteria and thus be included in the residential analysis. Previous analyses of Twitter data have found that tweet frequency follows the classic power law in statistics—a small number of users post the majority of tweets on the site (Rao et al. 2010).

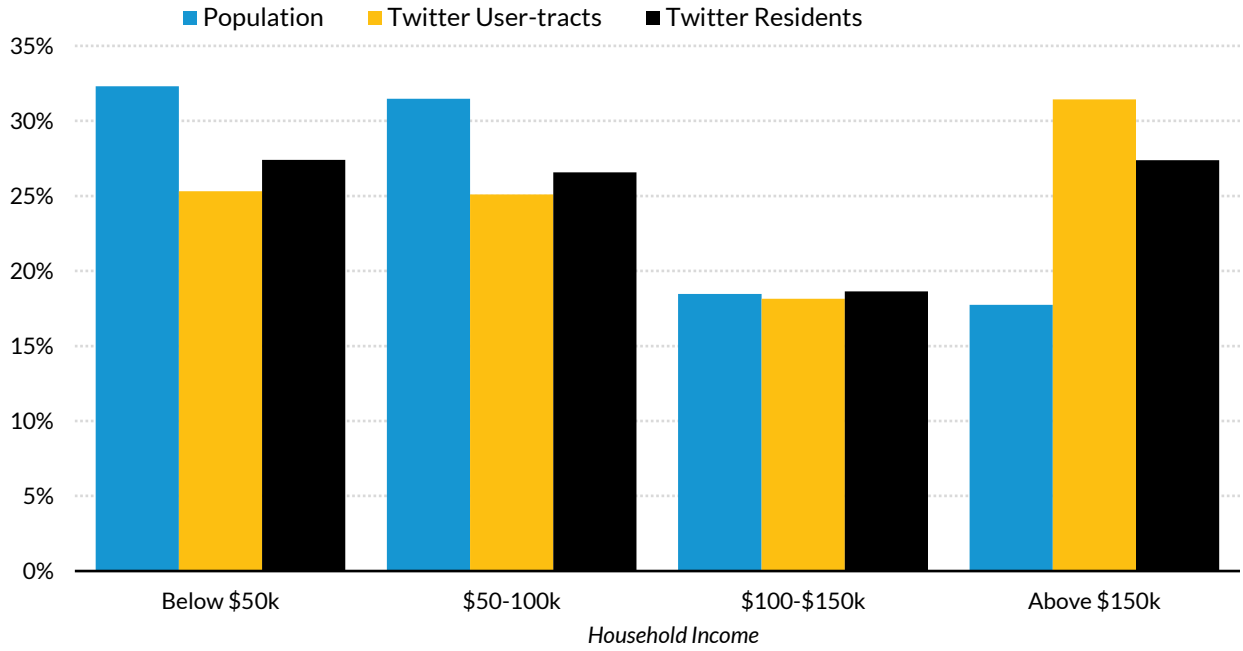
To better understand user representation in our Twitter dataset, we compare the demographic characteristics of the overall population to those represented by the user-tracts and resident tracts in our sample using a weighted average. The results are captured in Figures 1 through 3. Considering the possible sources of demographic imbalances in Twitter data described earlier, the true degree of over or underrepresentation may be more pronounced than suggested by the below figures.

Figure 1. Tracts with Higher White and Asian Populations Are Slightly Overrepresented, while Tracts with Higher Black Populations Are Slightly Underrepresented
Tract-level population statistics, share of all people



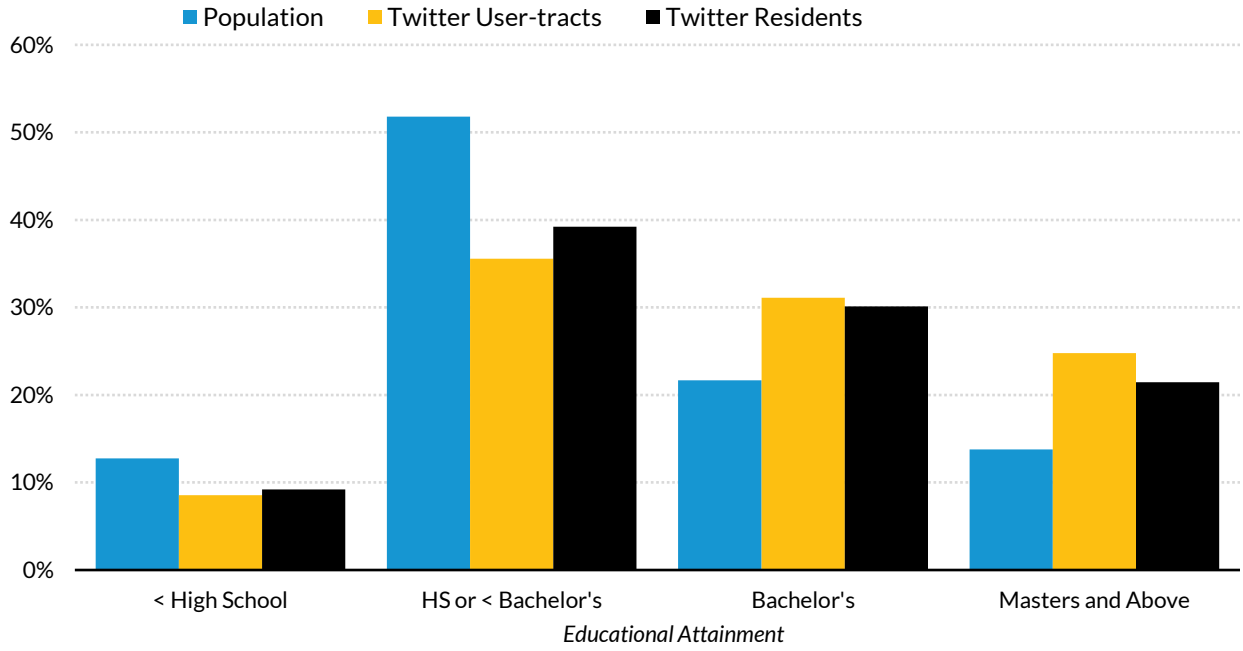
Source: Urban Institute analysis of American Community Survey 2011–15 5-year data, Twitter sample data collected from November 16, 2016, to June 30, 2017, and LEHD Origin-Destination Employment Statistics 2014 data.
Note: Our initial analyses used Census race designations only. Future analyses will include ethnicity reporting to understand patterns for Hispanic and Latino populations.

Figure 2. Tracts with More Higher Income Households are Overrepresented
Tract-level population statistics, share of all households



Source: Urban Institute analysis of American Community Survey 2011–15 5-year data, Twitter sample data collected from November 16, 2016, to June 30, 2017, and LEHD Origin-Destination Employment Statistics 2014 data.

Figure 3. Tracts with More People of Higher Educational Attainment are Overrepresented
Tract-level population statistics, share of all people age 25 or older



Source: Urban Institute analysis of American Community Survey 2011–15 5-year data, Twitter sample data collected from November 16, 2016, to June 30, 2017, and LEHD Origin-Destination Employment Statistics 2014 data.

The demographic overrepresentation of educated, higher income areas in our dataset is similar to a recent Pew Research Center (2017) survey, which found college educated and higher income individuals were much more likely to use Twitter than their less educated and lower-income counterparts. Despite the limitations of using public Twitter data, we believe our analysis has the potential to provide valuable insights. However, further research is required to determine the difference between the demographic profiles of Twitter users and the overall population.

Spatial Segregation of Twitter Users

The first research question we address is whether the spatial mobility patterns of Twitter users and their tweets are similar to the patterns of physical segregation. If the geospatial behavior of Twitter users reflects existing residential segregation patterns in the Chicago metro area, we would expect geotagged Tweets from certain groups of users to cluster in demographically-similar tracts.

Method

We use "user overlap" as a measure of spatial mobility. User overlap between any two tracts is the number of unique user tract combinations (user-tracts) that have tweeted from both tracts during the period captured in our dataset. A pair of tracts with a high volume of user overlap suggests that a large number of Twitter users have physically visited both tracts. For this analysis, we do not distinguish between residents, tourists, and workers; we are interested in the movement patterns of all users across the city.

Since the volume of user overlap between a pair of tracts depends on user density and Tweet frequency, issues of representation prevent us from measuring a direct relationship between the demographic similarity of tracts and the volume of user overlap. Instead, we adopt a network analysis approach to modeling tracts and user overlaps. We build a graph where nodes represent census tracts and weighted edges connecting the nodes represent the number of user overlaps between two census tracts. We then apply the NetworkX implementation of the Louvain community detection algorithm to identify communities among census tracts. The Louvain

method is an algorithm that identifies clusters, or communities, within large networks by optimizing for modularity, a measure of the difference between the proportion of edges falling within clusters and the expected proportion in a random network. The modularity score of a particular graph partition has a maximum value of 1, and a higher modularity score suggests a stronger, less random community structure.

"Communities" are clusters of tracts that are more densely connected to each other than to tracts outside the community. In other words, communities that the algorithm identifies have more user overlaps inside than outside the cluster. Two tracts are more likely to belong to the same community if there are more users traversing and tweeting between them, that is, if the number of user overlaps is higher.

Results

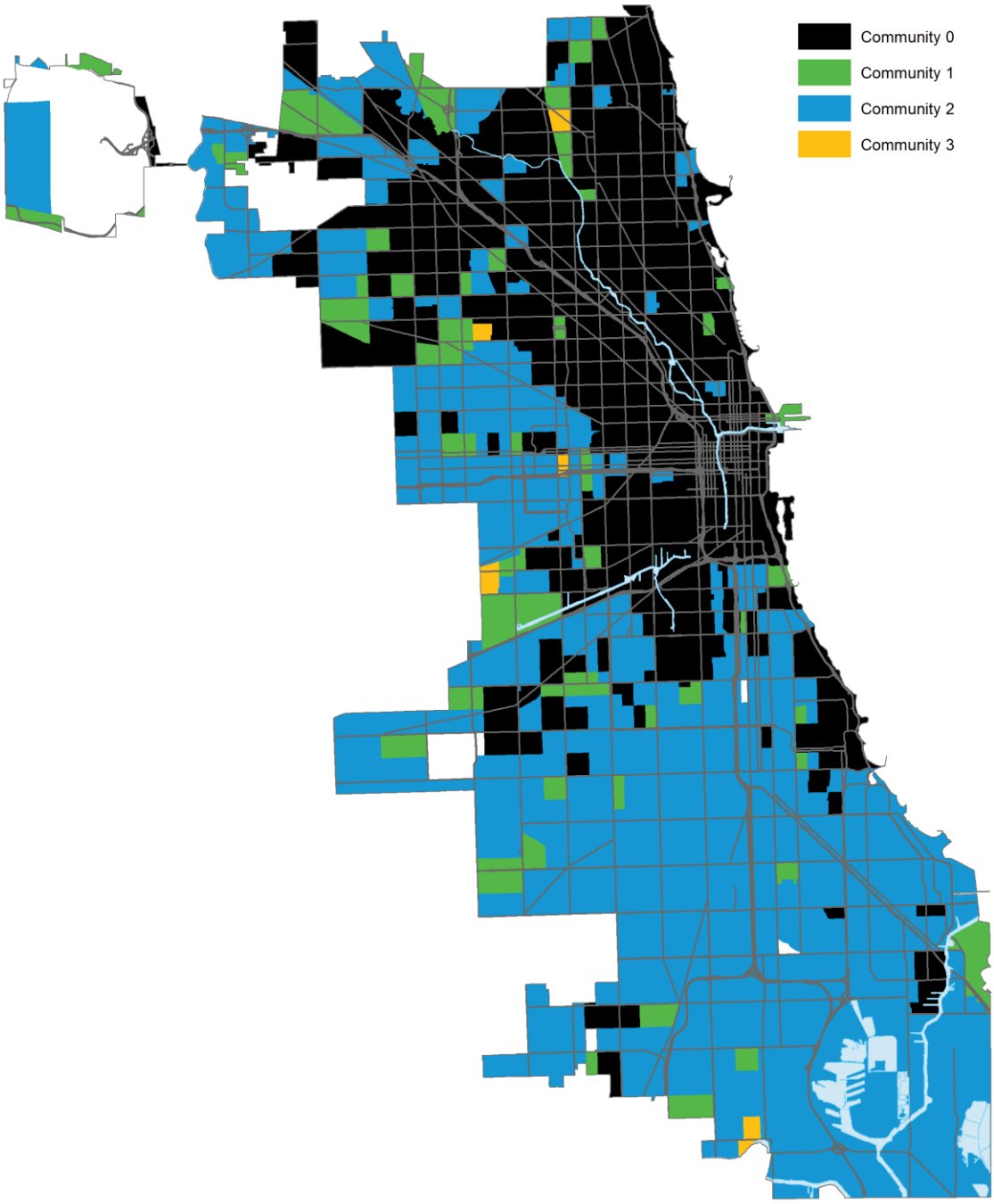
The algorithm identified four communities in our dataset, resulting in a modularity score of 0.24. Table 3 shows the number of tracts classified into each community, and Figure 4 visualizes the locations of tracts belonging to each community.

Table 3. Four Communities Identified by the Louvain Algorithm

	Tracts	
	Number	**Percent**
Community		
0	469	21%
1	698	32%
2	839	38%
3	191	9%

Source: Urban Institute analysis of Twitter sample data collected from November 16, 2016, to June 30, 2017.
Note: The number of tracts sums to 2,197 because one tract in our sample did not have any connections with other tracts, and was therefore not included in the graph model.

Figure 4. Locations of the Four Communities Identified by the Louvain Algorithm, City of Chicago



Source: Urban Institute analysis of Twitter sample data collected from November 16, 2016, to June 30, 2017. Map data sourced from the Census, City of Chicago, and © OpenStreetMap contributors.

To determine whether the algorithmically defined groups of tracts reflect existing neighborhood demographics, we performed a multinomial logistic regression to determine the predictive power of demographic characteristics on community membership of a tract. We chose three regressor variables that represent traditional demographic and geographic divisions in the Chicago metro area:

1. Percent black population
2. Percent families with annual income below $50,000
3. Percent adults age 25 and up with no bachelor's degree

We then calculated the marginal effects of each regressor on each outcome. This is the relative difference in probability of being in the community due to a one percentage point increase in the value of the regressor. Although the predictors do not explain the majority of variation in the independent variable (pseudo R squared score was 0.15), there do seem to be patterns in community membership based on race, income, educational attainment.

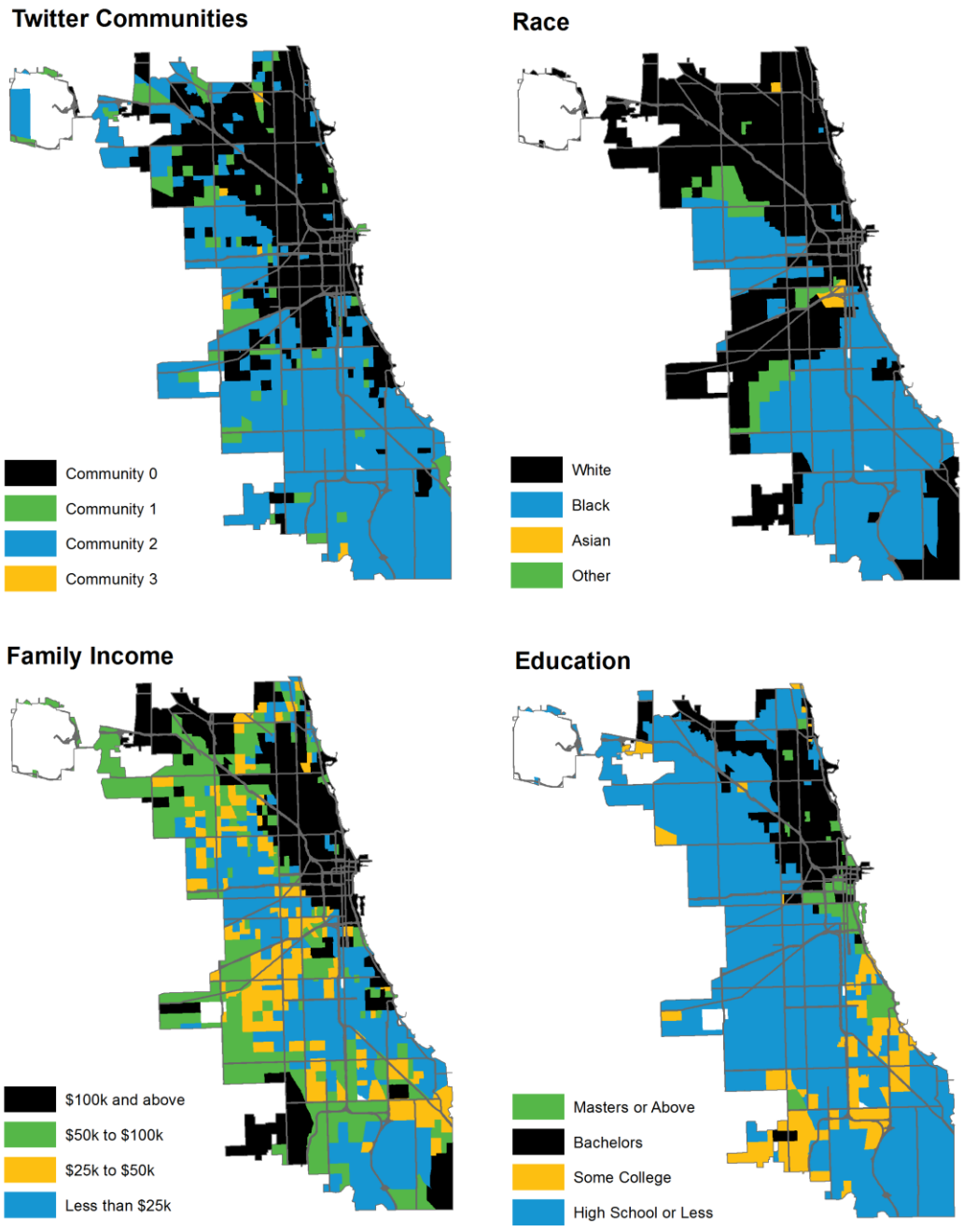
- Lower income and higher education have the most predictive power for membership in Community 0. Tracts with more families below $50,000 annual income and tracts with more individuals with bachelor's degrees are more likely to be in Community 0. Note that this may include tracts with families below $50,000 but lower average education levels, and tracts with families with bachelor's degrees and higher incomes.
- Higher income has the most predictive power for membership in Community 1. Tracts with more families below $50,000 annual income are less likely to be in Community 1.
- Race and lower educational attainment have the most predictive power for membership in Community 2. Tracts with higher percent black and individuals without Bachelor's degrees are more likely to be in Community 2.
- No significant results were found for Community 3, which represented a much smaller and randomly distributed set of tracts in our analysis.

We observed similar results when we run the regression using binomial rather than continuous regressor variables:

1. Majority of population is black
2. Majority of families have annual income below $50,000
3. Majority of adults age 25 and up have no bachelor's degree

The complete regression output can be found in Appendix B.

Figure 5. Locations of the Four Communities Identified by the Louvain Algorithm, City of Chicago, Compared with Basic Demographic Characteristics (Color Represents Plurality)



Source: Urban Institute analysis of 2011–15 5-year American Community Survey data and Twitter sample data collected from November 16, 2016, to June 30, 2017. Map data sourced from the Census, City of Chicago, and © OpenStreetMap contributors.
Notes: Race, Income, and Education categories show the color for a tract when the given category represents a plurality within the tract.

A closer look into the City of Chicago reveals community membership that mirrors the well-known divide between the North and South sides of Chicago. The predominantly white tracts in the North side tend to belong to Community 0, while the predominantly black and Latino tracts in the South and West sides tend to belong to Community 2. These results suggest that the spatial movement of Twitter users suggested by geotagged tweets aligns in part with existing segregation patterns in Chicago.

Digital Segregation in Twitter Networks

The second research question we seek to address is whether the segregation of Twitter users' friend networks is related to the demographic characteristics of the underlying neighborhoods. If the segregation of friend networks on Twitter are related to existing demographic patterns in the city, we would expect users who reside in demographically similar tracts to have more similar social networks on Twitter. Essentially, we seek to measure whether neighborhoods that are demographically similar tend to house residents with similar social networks.

Method

Using the residency definition described earlier in the paper, we explore the relationship between the similarity in demographic makeup of tracts ("demographic similarity") and the similarity in tract residents' social networks ("network similarity").

The demographic similarity score between two tracts quantifies their similarity along a subset of demographic characteristics. For each pair of tracts in the Chicago MSA, we calculate four demographic similarity scores: one for each demographic characteristic (racial categories, income brackets, educational attainment levels) and the combination of the three. The similarity is measured using the cosine similarity metric, which is a number between 0 and 1 that indicates the resemblance between two nonzero vectors. In this case, the vectors represent the share of people in each demographic category along a given demographic characteristic and sum to one.

A high demographic similarity score between two tracts on some characteristic—for example, race—indicates that the two tracts have similar racial compositions.

The network similarity score between a pair of tracts quantifies the similarity of the Twitter friend networks of residents from the two tracts. For each pair of tracts in the Chicago MSA, we calculate two types of network similarity measures, one based on the Jaccard index and another based on the Hub Promoted Index (HPI). Given a pair of tracts, Tract A and Tract B for example, we first form all possible pairs consisting of a resident from Tract A and a resident from Tract B. For each of these resident pairs, we calculate the Jaccard and HPI scores, which are two ways of estimating the similarity between two sets of objects. In this case, the sets being compared are the Twitter networks (followers and followings) of the two residents. The network similarity score between Tract A and Tract B is calculated as the average Jaccard or HPI score among all resident pairs.

See Appendix A for the complete methodological details.

Results

The distribution of demographic similarity scores varies by demographic characteristic. Similarity scores by race are most concentrated; many tract pairs are highly similar in their racial makeup, with cosine similarity scores close to 1. Similarity scores by income bracket and education are most evenly distributed; pairs of tracts exhibit varying degrees of similarity, indicating neighborhood divides on income and education may be less stark than divisions by race.

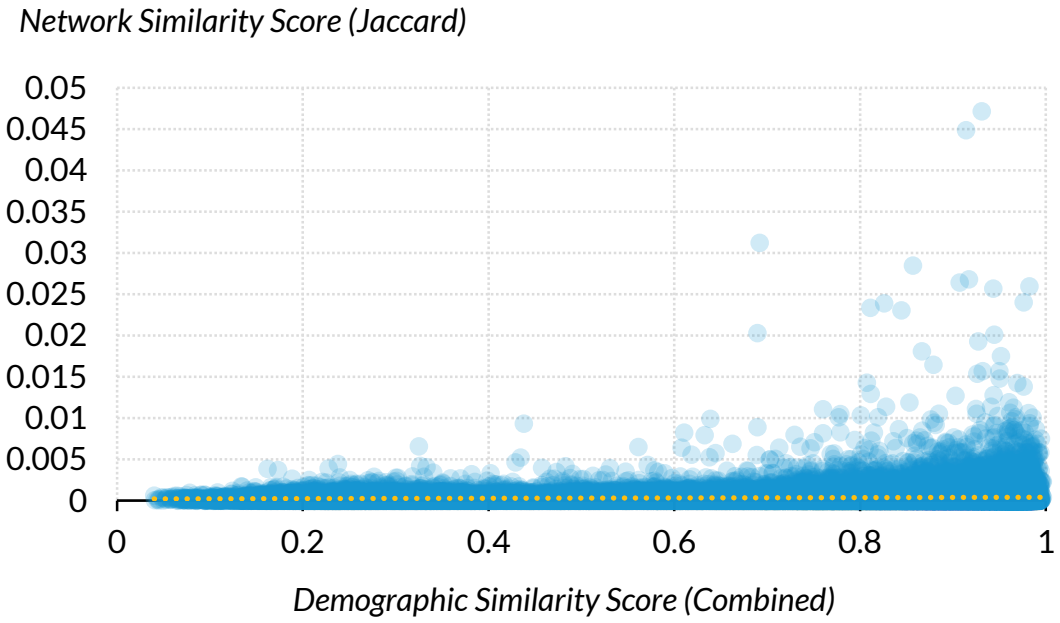
The distribution of network similarity scores was heavily concentrated around zero. Since most pairs of Twitter users have little to no friends in common, the Jaccard and HPI scores for most pairs of residents are zero.

We perform simple linear regressions of network similarity scores on demographic similarity scores to discover the relationship between the two measurements. The results suggest that demographic similarity explains almost none of the variance in network similarity - there is no clear linear relationship between demographic similarity and network similarity scores. The same pattern holds when the three demographic characteristics are evaluated separately.

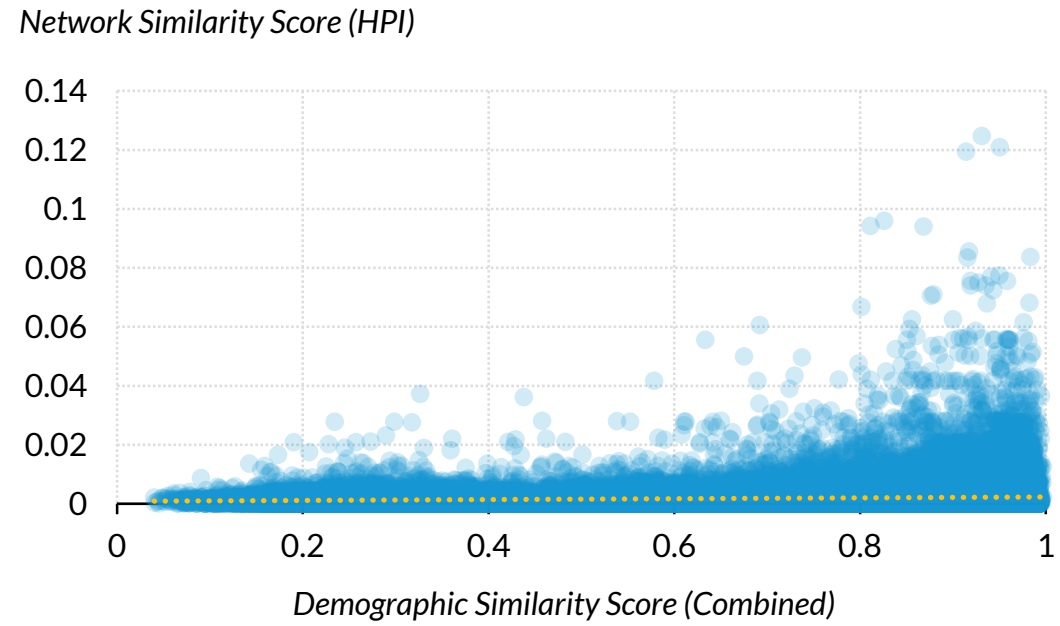
See Appendix C for details of demographic similarity score distributions and sample regression output.

Figure 6. Demographic Similarity Explains Almost None of the Variance in Network Similarity

Jaccard Network Similarity Score



HPI Network Similarity Score



Source: Urban Institute analysis of American Community Survey 2011-15 5-year data, Twitter sample data collected from November 16, 2016, to June 30, 2017, and LEHD Origin-Destination Employment Statistics 2014 data.

Suggested Findings

We set out to answer two fairly narrow questions, using the Chicago metro area as a case study:

1. Are the spatial mobility patterns of Twitter users and their tweets similar to the patterns of physical segregation along the lines of race, income, and education?
2. Are Twitter users' friend networks related to the demographic characteristics of the neighborhoods in which they reside?

We found that, although the spatial mobility patterns of Twitter users and their tweets *were similar* to the patterns of physical segregation in the Chicago metro area, the similarity of Twitter users' friend networks *did not seem to be related to* the demographic characteristics of their underlying neighborhoods. These findings suggest that Twitter geolocation data confirms existing physical segregation patterns but that the open nature of the internet create fewer barriers to the creation of online social networks.

However there are a number of important caveats to consider, so we caution drawing strong conclusions from these results. Twitter data only represents approximately 24 percent of online adults, and among that group, we studied the small portion of public Twitter users who enabled the geolocation feature on their account and exclude the small portion of public Twitter users with private friend networks. Our analysis shows that this group is more likely to be white, high income, and more highly educated than the Chicago metro population as a whole. Further research must be done in order to better understand the representativeness of Twitter data and its utility for understanding the connection between segregation in the physical and digital spaces. We encourage others to experiment with these methods, expand this relatively new field, and provide feedback.

Appendix A. Digital Segregation Methodology

Demographic Similarity

The cosine similarity metric measures the normalized dot product between two sets of data points. We use it to quantify the likeness of each pair of tracts along the lines of the following demographic characteristics (D): race, income, educational attainment, and all three combined.

The demographic similarity score for demographic characteristic D between Tract A and Tract B is calculated as follows:

$$demographic_similarity(D_A, D_B) = \frac{D_A \cdot D_B}{||D_A|| ||D_B||},$$

where D_A = a vector containing percent distributions along demographic characteristic D in tract A. The number of subcategories in demographic characteristic D determines the length of D_A. The subcategories in each demographic characteristic are defined as follows:

Race: white, black or African America, American Indian and Alaska Native, Asian, Native Hawaiian or other Pacific Islander, other race, two or more races

Income (family income in past 12 months, in 2015 inflation-adjusted dollars): Less than $10,000, $10,000 to $14,999, $15,000 to $19,999, $20,000 to $24,999, $25,000 to $29,999, $30,000 to $34,999, $35,000 to $39,999, $40,000 to $44,999, $45,000 to $49,999, $50,000 to $59,999, $60,000 to $74,999, $75,000 to $99,999, $100,000 to $124,999, $125,000 to $149,999, $150,000 to $199,999, $200,000 or more

Education (educational attainment for the population 25 years and over): No schooling completed, Nursery school, Kindergarten, 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, High school diploma, GED or alt credential, Some college < 1 year, Some college >= 1 yr no degree, Associates, Bachelors, Masters, Professional school, Doctorate

Network Similarity

The network similarity score between Tract A and Tract B is calculated as follows:

$$network_similarity(A, B) = \frac{1}{|R|} \sum_{i,j \in R} Similarity(i, j),$$

where R = all resident pairs (i, j) formed from the Cartesian product $R_A \times R_B$, where

R_A = all residents of Tract A

R_B = all residents of Tract B

The Jaccard similarity score and Hub Promoted Index (HPI) are two ways to measure the similarity between two sets. We use them to quantify the likeness of the friend networks between two residents. $Similarity(i, j)$ differs depending on which set similarity metric we choose to use:

$$Jaccard_Similarity(i, j) = \frac{|\mathrm{F}_i \cap \mathrm{F}_j|}{|\mathrm{F}_i \cup \mathrm{F}_j|}$$

$$HPI_Similarity(i, j) = \frac{|\mathrm{F}_i \cap \mathrm{F}_j|}{\min(|\mathrm{F}_i|, |\mathrm{F}_j|)}$$

$F_i = E_i \cup G_i$ = friends of User i, where

E_i = followings of User i, i.e., users whom User i follows

G_i = followers of User i, i.e., users who follow User i

Appendix B. Spatial Segregation Analysis

This appendix provides detailed regression results for the community detection analysis. In both sets of analyses, we performed a multinomial logistic regression using demographic variables to predict community membership.

Continuous Demographic Variables

```
Optimization terminated successfully.
         Current function value: 1.083027
         Iterations 6
                          MNLogit Regression Results
==============================================================================
Dep. Variable:              community   No. Observations:                 2197
Model:                        MNLogit   Df Residuals:                     2185
Method:                           MLE   Df Model:                            9
Date:                Thu, 17 Aug 2017   Pseudo R-squ.:                  0.1498
Time:                        11:31:13   Log-Likelihood:                -2379.4
converged:                       True   LL-Null:                       -2798.8
                                        LLR p-value:                9.683e-175
==============================================================================
 community=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      -1.4331      0.183     -7.839      0.000      -1.791      -1.075
black_pct       0.0163      0.003      4.733      0.000       0.010       0.023
below50k_pct   -0.0994      0.006    -15.725      0.000      -0.112      -0.087
nobach_pct      0.0827      0.005     16.716      0.000       0.073       0.092
------------------------------------------------------------------------------
 community=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      -2.5320      0.209    -12.097      0.000      -2.942      -2.122
black_pct       0.0344      0.003     11.135      0.000       0.028       0.040
below50k_pct   -0.0815      0.006    -13.510      0.000      -0.093      -0.070
nobach_pct      0.0862      0.005     16.888      0.000       0.076       0.096
------------------------------------------------------------------------------
 community=3       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      -2.8570      0.286     -9.995      0.000      -3.417      -2.297
black_pct       0.0122      0.005      2.502      0.012       0.003       0.022
below50k_pct   -0.0949      0.009    -11.168      0.000      -0.112      -0.078
nobach_pct      0.0833      0.007     12.396      0.000       0.070       0.096
==============================================================================
```

Average marginal effect of variables on community placement

MNLogit Marginal Effects

Dep. Variable:	community
Method:	dydx
At:	overall

community=0	dy/dx	std err	z	P>|z|	[0.025	0.975]
black_pct	-0.0030	0.000	-8.459	0.000	-0.004	-0.002
below50k_pct	0.0114	0.001	20.581	0.000	0.010	0.012
nobach_pct	-0.0105	0.000	-26.592	0.000	-0.011	-0.010

community=1	dy/dx	std err	z	P>|z|	[0.025	0.975]
black_pct	-0.0009	0.000	-2.112	0.035	-0.002	-6.62e-05
below50k_pct	-0.0075	0.001	-9.916	0.000	-0.009	-0.006
nobach_pct	0.0041	0.001	7.050	0.000	0.003	0.005

community=2	dy/dx	std err	z	P>|z|	[0.025	0.975]
black_pct	0.0045	0.000	13.807	0.000	0.004	0.005
below50k_pct	-0.0022	0.001	-2.965	0.003	-0.004	-0.001
nobach_pct	0.0052	0.001	8.406	0.000	0.004	0.006

community=3	dy/dx	std err	z	P>|z|	[0.025	0.975]
black_pct	-0.0006	0.000	-1.937	0.053	-0.001	7.16e-06
below50k_pct	-0.0017	0.000	-3.388	0.001	-0.003	-0.001
nobach_pct	0.0012	0.000	3.231	0.001	0.000	0.002

Demographic Dummy Variables

Optimization terminated successfully.
Current function value: 1.154109
Iterations 7

MNLogit Regression Results

Dep. Variable:	community	No. Observations:	2197
Model:	MNLogit	Df Residuals:	2185
Method:	MLE	Df Model:	9
Date:	Thu, 17 Aug 2017	Pseudo R-squ.:	0.09404
Time:	10:11:04	Log-Likelihood:	-2535.6
converged:	True	LL-Null:	-2798.8
		LLR p-value:	1.267e-107

community=1	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.2076	0.103	-2.011	0.044	-0.410	-0.005
majority_black	0.6074	0.258	2.354	0.019	0.102	1.113
majority_below50k	-1.6032	0.180	-8.926	0.000	-1.955	-1.251
majority_nobach	1.3468	0.142	9.513	0.000	1.069	1.624

community=2	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.8316	0.124	-6.682	0.000	-1.076	-0.588
majority_black	2.0424	0.220	9.264	0.000	1.610	2.474
majority_below50k	-0.6921	0.161	-4.289	0.000	-1.008	-0.376
majority_nobach	1.7126	0.159	10.778	0.000	1.401	2.024

community=3	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-1.4889	0.161	-9.226	0.000	-1.805	-1.173
majority_black	-0.3873	0.449	-0.863	0.388	-1.267	0.492
majority_below50k	-1.1101	0.246	-4.512	0.000	-1.592	-0.628
majority_nobach	1.3039	0.205	6.372	0.000	0.903	1.705

Average marginal effect of variables on community placement

MNLogit Marginal Effects

Dep. Variable:	community
Method:	dydx
At:	overall

community=0	dy/dx	std err	z	P>|z|	[0.025	0.975]
majority_black	-0.1648	0.032	-5.084	0.000	-0.228	-0.101
majority_below50k	0.1741	0.022	8.054	0.000	0.132	0.216
majority_nobach	-0.2264	0.018	-12.893	0.000	-0.261	-0.192

community=1	dy/dx	std err	z	P>|z|	[0.025	0.975]
majority_black	-0.0777	0.037	-2.128	0.033	-0.149	-0.006
majority_below50k	-0.2183	0.027	-7.965	0.000	-0.272	-0.165
majority_nobach	0.0538	0.022	2.393	0.017	0.010	0.098

community=2	dy/dx	std err	z	P>|z|	[0.025	0.975]
majority_black	0.3477	0.025	13.859	0.000	0.299	0.397
majority_below50k	0.0619	0.024	2.616	0.009	0.016	0.108
majority_nobach	0.1596	0.025	6.471	0.000	0.111	0.208

community=3	dy/dx	std err	z	P>|z|	[0.025	0.975]
majority_black	-0.1052	0.032	-3.306	0.001	-0.168	-0.043
majority_below50k	-0.0177	0.017	-1.062	0.288	-0.050	0.015
majority_nobach	0.0130	0.013	0.967	0.334	-0.013	0.039

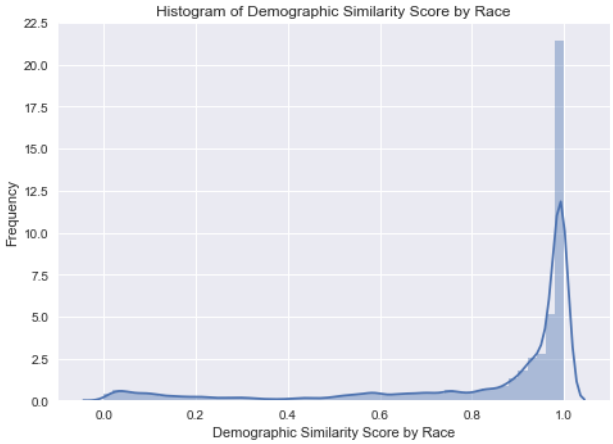
Appendix C. Digital Segregation Analysis

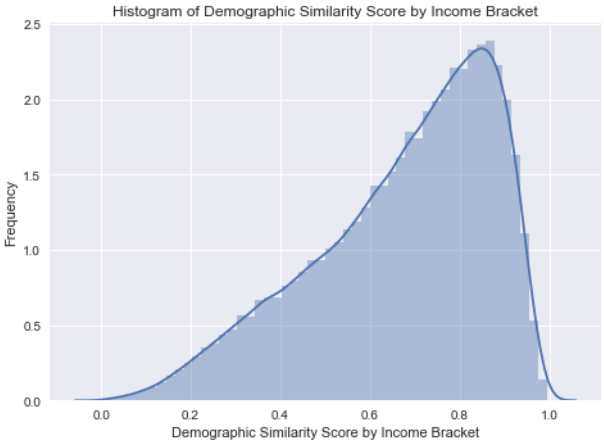
This appendix provides details on the job to population thresholds we explored, distributions for the demographic and network similarity scores, and regression results comparing the relationship between the demographic and network similarity scores.

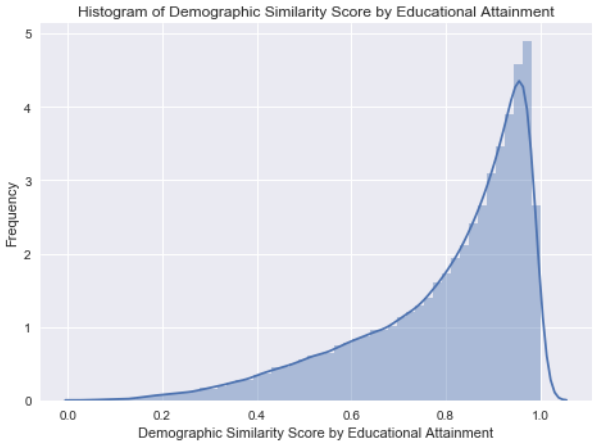
Jobs Population Threshold

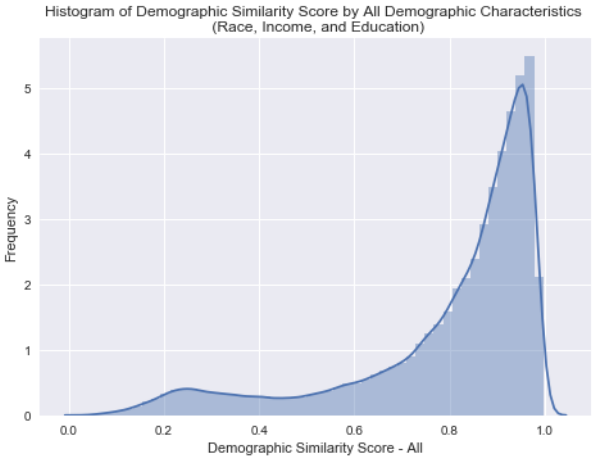
Ratio Cutoff	Share of population living in commercial tracts	Share of tracts excluded
1	8.8%	9.4%
1.5	4.6%	5.0%
2	2.8%	3.1%
2.5	2.1%	2.4%
3	1.4%	1.6%

Demographic Similarity Scores

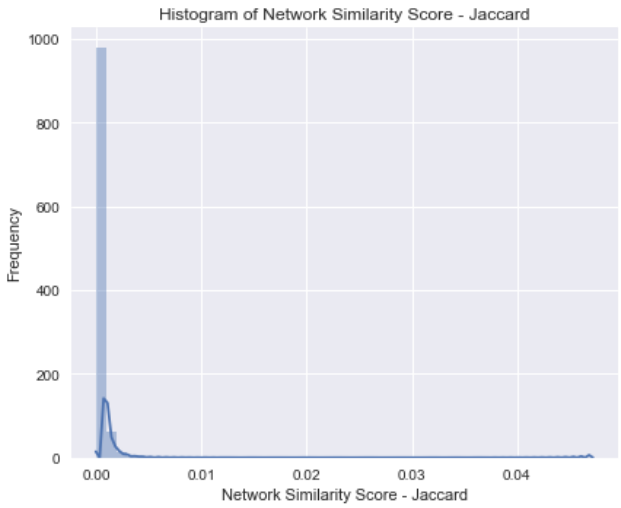


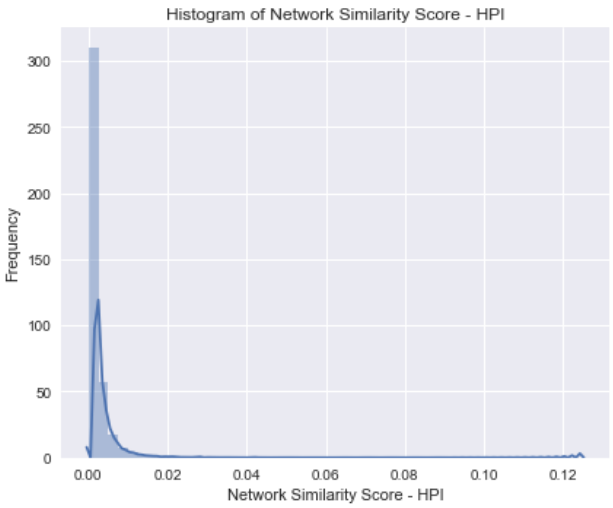






Network Similarity Scores





Regression Results

We ran simple bivariate regressions of Jaccard and HPI similarity scores against demographic similarity scores. The example below regresses Jaccard similarity scores on all demographic characteristics included in this study: race, household income, and education.

OLS Regression Results

Dep. Variable:	jaccard	**R-squared:**	0.004
Model:	OLS	**Adj. R-squared:**	0.004
Method:	Least Squares	**F-statistic:**	328.2
Date:	Tue, 15 Aug 2017	**Prob (F-statistic):**	3.22e-73
Time:	14:34:30	**Log-Likelihood:**	5.3384e+05
No. Observations:	91378	**AIC:**	-1.068e+06
Df Residuals:	91376	**BIC:**	-1.068e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0002	9.22e-06	21.445	0.000	0.000	0.000
all	0.0002	1.12e-05	18.116	0.000	0.000	0.000

Omnibus:	186947.582	**Durbin-Watson:**	1.530
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1706090175.026
Skew:	16.792	**Prob(JB):**	0.00
Kurtosis:	671.557	**Cond. No.**	7.97

References

Acs, G., Pendall, R., Treskon, M., Khare, A. (2016). *The Cost of Segregation.* The Urban Institute.

Cesare, N., Lee, H., McCormick, T., and Spiro, E. S. (2017). Redrawing the "Color Line": Examining Racial Segregation in Associative Networks on Twitter (Submitted) .

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2013). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *Proceedings of the Sixth International AAAI Conference on Weblogs an Social Media*.

Cranshaw, J., Toch, E., Hong, J., Kittur, A., and Sadeh, N. (2010). Bridging the gap between physical location and online social networks. *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 119-128.

Crutcher, M., and Zook, M. (2009). Placemarks and waterlines: Racialized cyberscapes in post-Katrina Google Earth. *Geoforum*, 40(4):523-534.

Culotta, A., Kumar, N. R., Cutler, J. (2015). Predicting the Demographics of Twitter Users from Website Traffic Data. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.*

Edelman, B., Luca, M., and Svirsky, D. n.d.. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. Retrieved from Benjamin Edelman, personal website. www.benedelman.org/publications/airbnb-guest-discrimination-2016-09-16.pdf.

Glaeser E. and Vigdor, J. (2012). The End of the Segregated Century: Racial Separation in America's Neighborhoods, 1890-2010. Manhattan Institute Civic Report 66. New York, NY: Manhattan Institute.

Grabinsky, J. and Reeves, R.V. (2015, December 21). The Most American City: Chicago, Race, and Inequality. Retrieved from Brookings: https://www.brookings.edu/blog/social-mobility-memos/2015/12/21/the-most-american-city-chicago-race-and-inequality/

Graham, M. H. (2013). Where in the World are You? Geolocation and Language Identification in Twitter.

Hargittai, E., and Hinnant, A. (2008). Digital inequality: Differences in young adults' use of the Internet. *Communication Research* , Vol 35, Issue 5, pp. 602 - 621.

Hassani, S. N. (2006). Locating digital divides at home, work, and everywhere else. *Poetics*, The digital divide in the twenty-first century, 34(4):250-272.

Hoffman, D. L., and Novak, T. P. (1998). Bridging the Racial Divide on the Internet. *Science*, 280(5362):390-391.

Hofstra, B., Corten, R., van Tubergen, F., and Ellison, N. B. (2017). Sources of Segregation in Social Networks: A Novel Approach Using Facebook. *American Sociological Review*, Vol. 82 Issue 3, pp. 625-656.

Iceland, J., Weinberg, D. H., and Steinmetz, E. (2002, August). Racial and Ethnic Segregation in the United States: 1980-2000. Census 2000 Special Reports. Washington, DC: U.S. Bureau of the Census.

Massey, D. S. and Denton, N. A. (1987). Trends in the Residential Segregation of Blacks, Hispanics, and Asians: 1970-1980. *American Sociological Review*, Vol. 52 No. 6, pp. 802-825.

Mislove, A., Jørgensen, S. L., Ahn, Y., Onnela, J., and Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.*

Pew Research Center (2017). Social Media Fact Sheet. Retrieved from Pew Research Center: https://web.archive.org/web/20180104134229/http://www.pewinternet.org/fact-sheet/social-media/

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010, October). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (pp. 37-44). ACM.

Rivkin, S. G. (1994). Residential Segregation and School Integration. *Sociology of Education*, 67(4):279-292.

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

Sloan L, M. J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *Plos ONE*, 10(11).

Smith, A. (2017, January 12). Record shares of Americans now own smartphones, have home broadband. Retrieved from Pew Research Center: http://www.pewresearch.org/fact-tank/2017/01/12/evolution-of-technology/

Stern, M. J., Adams, A. E., and Elsasser, S. (2009). Digital Inequality and Place: The Effects of Technological Diffusion on Internet Proficiency and Usage across Rural, Suburban, and Urban Counties. *Sociological Inquiry*, Nov2009, Vol. 79 Issue 4, p391-417.

Thelwall, M. (2009). Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, Vol. 60 Issue 2, p219-231.

Wimmer, A., and Lewis, K. (2010). Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook. *American Journal of Sociology*, Vol. 116, No. 2, pp. 583-642.