Journal of Urban Economics 118 (2020) 103264

Contents lists available at ScienceDirect

Journal of Urban Economics

journal homepage: www.elsevier.com/locate/jue

Social connectedness in urban areas☆

Michael Bailey[a], Patrick Farrell[a], Theresa Kuchler[b], Johannes Stroebel[b,*]

[a] *Princeton University*
[b] *New York University, Stern School of Business, NBER, and CEPR*

A R T I C L E I N F O

JEL classification:
R3
R4

Keywords:
Social Connectedness
Agglomeration Externalities
Transportation Infrastructure

A B S T R A C T

We use de-identified and aggregated data from Facebook to explore the spatial structure of social networks in the New York metro area. We find that a substantial share of urban residents' connections are to individuals who are located nearby. We also highlight the importance of transportation infrastructure in shaping urban social networks by showing that social connectedness declines faster in travel time and travel cost than it does in geographic distance. We find that areas that are more socially connected with each other have stronger commuting flows, even after controlling for geographic distance and ease of travel. We also document significant heterogeneity in the geographic breadth of social networks across New York zip codes, and show that this heterogeneity correlates with access to public transit. Zip codes with geographically broader social networks also have higher incomes, higher education levels, and more high-quality entrepreneurial activity. We also explore the social connections between New York zip codes and foreign countries, and highlight how these are related to past migration movements.

Social networks influence many aspects of our lives, with social ties providing access to a wide range of new ideas and employment opportunities (see Granovetter, 2005; Jackson, 2014; Bramoulle et al., 2016). Theories of positive externalities from agglomeration feature prominently in the study of urban economics, which frame the ability to interact with many different people as a key force behind the high productivity of cities (e.g., Jacobs, 1969; Bairoch, 1991; Glaeser, 2011; Barwick et al., 2019). For example, Glaeser et al. (1992) describe how *"the cramming of individuals, occupations, and industries into close quarters provides an environment in which ideas flow quickly from person to person."* Similarly, many models of urban economies focus on the specific role of social interactions in generating the positive externalities of agglomeration, and on influencing the size, structure, and location of cities (Brueckner and Largey, 2008; Fujita and Thisse, 2013; Sato and Zenou, 2015; Mossay and Picard, 2011; Helsley and Strange, 2014; Duranton and Puga, 2015). At the same time, a literature has debated whether technological advances in the past four decades has led to a "death of distance" as the increased prevalence of technologies that facilitate connectivity without geographic limits changes social organization (Cairncross, 2001; Green, 2002; Farazmand, 1999). Taken to its end, this argument predicts the positive externalities of agglomeration to weaken dramatically, leading to a "death of cities" (Gilder, 1995). While a number of more recent studies suggest that proximity continues to shape individuals' personal and online networks (Kim et al., 2017; Mok et al., 2010; Goldenberg and Levy, 2009; Takhteyev et al., 2012; Scellato et al., 2010), data challenges in measuring social connections have limited researchers' ability to empirically study the full geography of urban residents' present-day social networks.

In this paper we introduce a novel dataset that allows us to explore — with both high coverage and high granularity — the geography of social networks in and around urban environments. We provide new insights on the role of social interactions in creating agglomeration externalities by investigating the relationship between social networks, transportation infrastructure, and economic outcomes within cities. We measure social networks using aggregated and de-identified data from Facebook, a global online social network. At the end of 2017, Facebook had 239 million monthly active users in the U.S. and Canada and about 2.1 billion such users globally. We observe a de-identified snapshot of all Facebook users with location history enabled as of March 2018. For these users, we observe their locations at the zip code level as well as

☆ This version: May 28, 2020. We thank the Center for Global Economy and Business at NYU Stern for generous research support. We also thank Leah Boustan, Ed Glaeser, Paul Goldsmith-Pinkham, Matt Jackson, Stijn van Nieuwerburgh, Eleonora Patacchini, Nico Stier, and Maisy Wong, as well as seminar and conference participants at Facebook, NYU, the NYC Real Estate Conference, the 2019 Ohlstadt Workshop, the GEA Conference in Frankfurt, and the NBER Summer Institute, for their helpful comments. We thank Drew Johnston, Sung Lee, Hongbum Lee, and Dominic Russel for outstanding research assistance. This research was facilitated through a research consulting agreement between the academic authors (Farrell, Kuchler, and Stroebel) and Facebook. Bailey is an employee at Facebook.

* Corresponding author.
E-mail addresses: mcbailey@fb.com (M. Bailey), pwfarrell@princeton.edu (P. Farrell), tkuchler@stern.nyu.edu (T. Kuchler), johannes.stroebel@nyu.edu (J. Stroebel).

https://doi.org/10.1016/j.jue.2020.103264
Received 21 June 2019; Received in revised form 18 May 2020
Available online 29 May 2020

their connections to other individuals on Facebook. We use these data to explore the local, domestic, and international networks of Facebook users in both New York City (NYC) and the wider New York Combined Statistical Area (New York CSA).

The density, diversity, and large population of New York, combined with its geography and extensive public transportation infrastructure, present an ideal setting for investigating the factors that influence social network structure in urban settings. Indeed, we believe that our study brings the most comprehensive data to date to measure and explore the social structure of cities.[1] Our empirical approach complements an exciting recent literature that has used cell phone call records to better understand the geography of social connectedness (e.g., Schläpfer et al., 2014; Herrera-Yague et al., 2015; Büchel and von Ehrlich, 2016). Relative to that literature, the Facebook data capture many more links per individual, allowing us to measure the prevalence and distribution of potentially weak ties that have been shown to be important in the dissemination of information and ideas (Granovetter, 1977).[2] While we are unable to make any conclusive causal inferences on the determinants and effects of the observed social structures, we hope that the novel patterns presented in this paper can help advance our understanding of social connectedness in urban areas.

In the first part of the paper, we explore the role of geography and public transit infrastructure as potential determinants of social networks in urban areas. We first discuss a number of case studies that show that the social networks of urban zip codes are distributed along transit routes that connect these zip codes to other parts of the city. We then explore the relationship between social connectedness, geography, and transportation infrastructure more formally. We find that social connectedness declines in the geographic distance between locations: within NYC, a 10% greater geographic distance between zip codes is associated with 8.7% lower social connectedness, an elasticity of social connectedness to distance of -0.87. Notably, these results are of similar magnitude to those in Kim et al. (2017) who estimate a model in which a one kilometer increase in the geographic dispersion of individuals is associated with a 5-6% decrease in average social interactions. We next calculate the travel times on public transit between each pair of NYC zip codes and find that social connectedness declines even more strongly in the travel time between locations. Within NYC, the elasticity of social connectedness to travel time is -1.42, which is about 60% larger in magnitude than the elasticity of social connectedness to distance. This finding suggests that public transit can help facilitate the maintenance and formation of social links across individuals living in geographically distant parts of the same city. These estimates are thus consistent with public transit infrastructure being an important determinant of the magnitude of positive agglomeration forces within cities, which will depend on the extent to which individuals living in the same city actually interact with one another.[3]

More broadly, this section contributes to a research effort that finds that — despite new communications technologies — a substantial share of individuals' connections remain physically nearby (e.g., Goldenberg and Levy, 2009; Bailey et al., 2018b). The scale and coverage of our data allow us to add a note of reconciliation toward those who argue that technological advances should lead to more dispersed connections. In particular, while our results are generally consistent with studies that find communication technologies to serve as a complement to in-person interaction (e.g., Gaspar and Glaeser, 1998; Mok et al., 2010), they do highlight a role for developments in transportation infrastructure and technology to lead (albeit gradually and incompletely) to a lessening of the importance of strict geographic distance in the formation and maintenance of social links.

In addition to the role played by geographic distance and public transit travel time in forming and maintaining social ties between geographies, we find that zip codes that are more similar along demographic measures such as race, education, and income are more likely to be socially connected. This is consistent with previous studies that have documented that social ties are generally more common between similar individuals and regions, a feature that is often referred to as "homophily" (Lazarsfeld and Merton, 1954; Zipf, 1949; Verbrugge, 1983; Marmaros and Sacerdote, 2006; Bailey et al., 2018b; 2018a). We show that short public transit travel times are more important for connecting zip codes with different incomes than they are for connecting zip codes with similar incomes. This finding highlights that public transit might not just facilitate social connections between far-away zip codes in general, but does so particularly across zip codes with different demographics.

We also run a hierarchical agglomerative linkage clustering algorithm to construct hypothetical "connected communities" of zip codes that maximize within-community social connectedness. We find that although all the communities are contiguous at the CSA level, some communities are non-contiguous when focusing on zip codes within NYC. This finding reinforces the earlier observation that geographic distance might not be as relevant a measure to understand social ties within the context of an urban environment with a complex transportation network.

We then explore the interaction of social connectedness and two measures of across-zip code economic flows: home-to-work commuting patterns and total taxi trips. A 10 percent increase in connectedness between a pair of zip codes is associated with a 3 percent increase in the number of commuting flows and a 6 percent increase in the number cab trips, even after controlling for socioeconomic factors, geographic distance, and travel time. This relationship is the same when we only exploit variation in connectedness across zip codes that is driven by friendship links with family members or individuals from the same high school. The relationship is thus unlikely to be the result of a causal effect of commuting on forming social ties. Instead, our findings could be the result of individuals sharing job opportunities in their location with their friends, consistent with the relationship between information flows and worker flows found in Barwick et al. (2019) and the importance of referral networks found in Schmutte (2015). More generally, our results provide evidence consistent with models such as Kim et al. (2017), in which the geography of social networks plays an important role in shaping economic interactions, and our estimates can help parameterize such models.

We next provide a descriptive analysis of the geographic concentration of social networks. We find substantial heterogeneity in social network concentration across NYC zip codes. For residents of the median zip code, 29.0% of U.S.-based friends live within 5 miles, but this number ranges from 19.5% to 39.6% between the 5th and the 95th percentiles of the zip code distribution. Similarly, for the median NYC zip code, 22.0% of U.S.-based friends live among the nearest 1 million people, while the 5-95 percentile range is 13.1% to 32.7%. Various components of the social networks of zip codes' residents (e.g., close friends, high school friends, recent friends) are generally highly correlated in their geographic concentrations.

Consistent with the results described above, the geographic concentration of social networks is highly correlated with access to public transportation infrastructure (measured, for example, by the share of a zip code's population that lives within a quarter mile of a rail transit station). Quantitatively, a 15 minute increase in the average travel time to all other zip codes is associated with a 4.0 percentage point increase

[1] The zip code-level social connectedness data that we compile and use in this project is accessible to researchers and policy makers by emailing sci_data@fb.com.

[2] In addition, interactions via phone are often substitutes to in-person interactions. One might therefore worry that researchers' ability to observe a social link in phone records is systematically related to the frequency of the two individuals interacting in person. The latter should correlate both with geographic distance and the ease of travel via public transport.

[3] This result also aligns with recent findings that suggest that transportation infrastructure allows individuals to visit restaurants that are farther away, thereby lowering the segregation of consumption patterns (Davis et al., 2017).

in the share of friends living within 10 miles, even after controlling for population density and zip code demographics. The geographic concentration of social networks also correlates with socioeconomic outcomes: the share of friends living within various distances is decreasing in zip code income and entrepreneurial quality and increasing in the fraction of population without a high school degree. Our findings are consistent with studies of social networks wherein well-connected individuals optimally locate closer to central nodes, and access to jobs contacts and referrals through social interactions play a key role in shaping economic and labor market outcomes (Helsley and Zenou, 2014; Greve and Salaff, 2003; Schmutte, 2015); most directly, we provide evidence consistent with research such as Barwick et al. (2019), which highlights the economic importance of the diversity of social links across a variety of dimensions. Although our data do not allow us to make statements about the causal connection between social connectedness and socioeconomic outcomes, these findings are also consistent with the urban economics literature that points to social interactions as a primary channel for generating agglomeration externalities that improve residents' economic outcomes (see, for example, Fujita and Thisse, 2013).

In the final part of the paper, we study the social connectedness of New York zip codes to foreign countries. We find strong heterogeneity in the degree to which different zip codes are connected to different countries. We show that past migration movements are a strong determinant of connections abroad, which is suggestive of immigrants' desire to live in areas near the existing ethnic enclaves or areas with transportation accessible to these communities. Therefore, the clustering of ethnicities in a region plays a key role in explaining the presence of international friendship links.

In addition to the research already discussed, this paper contributes to a recent literature that has used data from online services such as Yelp and Twitter to better understand various elements of social and economic activity within cities (e.g., Davis et al., 2017; Glaeser et al., 2017). We also build on a literature that has studied the unique properties of urban social networks (see Glaeser and Kahn, 2004; Glaeser, 2011; Kowald et al., 2013; Ioannides, 2013; Herrera-Yague et al., 2015; Ioannides, 2015; Picard and Zenou, 2018; Kim et al., 2017; Sato and Zenou, 2015). Our novel data allow us to document that public transit infrastructure likely is a crucial determinant of the formation and maintenance of social ties in urban areas, in particular across locations with different demographic makeups. This suggests a mechanism through which transit infrastructure affects social network formation, which in turn can influence economic outcomes. In this sense, our work contributes to an important literature that has shown that transit investments generate immediate economic effects and cause long-term changes to the structure of cities. For instance, Perlman (2016) finds that transportation improvements had significant impact on increases in patenting, especially for counties that were not previously well-connected, and Glaeser (2005) finds that New York has become America's largest city due to its initial dominance as a hub of the transportation system (see also Glaeser and Shapiro, 2001; Glaeser and Gottlieb, 2009; Baum-Snow, 2013; Ioannides, 2013; Brooks and Lutz, 2014; Glaeser and Steinberg, 2016). We hope that the increasing availability of social network data from online social networking services such as Facebook will further boost research efforts that explore the determinants and effects of the social structures of cities.

1. Data

We construct our measures of the social connectedness across locations using de-identified administrative data from Facebook, a global online social networking service. Facebook was created in 2004 and as of the end of 2017 had 2.1 billion monthly active users globally and 239 million such users in the U.S. and Canada. An independent survey of Facebook users from 2015 found that more than 68% of the U.S. adult population and 79% of online adults in the U.S. used Facebook (Duggan et al., 2016). That same survey shows that Facebook usage rates among U.S.-based online adults were relatively constant across income groups, education levels, and race, and among urban, rural, and suburban residents; usage rates were slightly declining in age (from 88% of individuals aged 18 to 29, to 62% of individuals aged 65 and older).

Establishing a connection on Facebook requires the consent of both individuals, and there is an upper limit of 5,000 on the number of connections a person can have. As a result, Facebook connections are primarily between real-world acquaintances. Indeed, a second independent survey of Facebook users revealed that only 39% of users reported being Facebook friends with someone they had never met in person (Duggan et al., 2015). In contrast, Facebook users generally reported that they were Facebook friends with real-life friends: 91% said they were Facebook friends with current friends and 87% said they were connected to past friends, such as former classmates. Furthermore, most users reported that they were Facebook friends with their family members: 93% of Facebook users said they were Facebook friends with family members other than parents or children, 45% said they were Facebook friends with their parents, and 43% said they were Facebook friends with their children. Finally, Facebook networks often capture other important social ties: 58% of users said that they were Facebook friends with co-workers and 36% of users reported that they were Facebook friends with their neighbors (Duggan et al., 2015). As a result, networks formed on Facebook more closely resemble real-world social networks than those on other online platforms, such as Twitter and Instagram, where uni-directional links to non-acquaintances, such as celebrities, are common (see Bailey et al., 2018a, 2018b, 2019a, 2019b, 2020a, 2020b, Kuchler et al., 2020a, 2020b, for additional evidence that friendships observed on Facebook serve as a good proxy for real-world U.S. social connections).

We observe a de-identified snapshot of all active Facebook users from March 2018. We focus on those users who had location history enabled and who had interacted with Facebook over the 30 days prior to the date of the snapshot. We match those users who reside within the New York Combined Statistical Area (CSA) to their zip code locations. The New York CSA consists of 35 counties across the states of Connecticut, New Jersey, New York, and Pennsylvania. We count as within the New York CSA all zip codes that fall at least partly within a county making up the New York CSA. Users within the United States but not within the New York CSA are mapped to their county of residence. Users outside of the United States are mapped to their country of residence. From these data, we obtain a count of the number of connections between each zip code i in the New York CSA and each other region j, where j is either another zip code within the New York CSA, a U.S. county outside of the New York CSA, or a foreign country.

We only include zip codes in our analysis that have a total population of at least 500 people and that are above the 5th percentile in the number of eligible Facebook users within the New York CSA. These restrictions are intended to preserve user anonymity as well as to reduce the improper matching of users to officially unpopulated or unusual zip codes, such as individual non-residential buildings (e.g., post offices) or abnormal locations (e.g., JFK airport). Our final data set includes 182 zip codes in NYC and 1,181 zip codes across the entire New York CSA.

We combine these data on social networks with information on the population and demographics of zip codes from the 2015 Census Bureau 5-year American Community Survey (ACS) and the 2014 Internal Revenue Service (IRS) Individual Income Tax Statistics. In particular, information on total population, racial composition, and educational attainment comes from the ACS, and information on average income is calculated from IRS data.

Measuring Social Connectedness. To compare the intensity of social connectedness between zip codes with varying populations, we construct our measure of $SocialConnectedness_{i,j}$ as the total number of connections between individuals living in zip code i and individuals living in zip code j, which we refer to as $FB\_Connections_{i,j}$, divided by the

Table 1
Correlation of Social Connectedness Constructed from Select Friendship Pairs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
(1) All	1.00																
(2) Added <1 Year Ago	0.94	1.00															
(3) Added 1-5 Years Ago	0.98	0.95	1.00														
(4) Added <5 Years Ago	0.98	0.98	1.00	1.00													
(5) Added >5 Years Ago	0.98	0.86	0.93	0.92	1.00												
(6) Same High School	0.85	0.82	0.83	0.83	0.83	1.00											
(7) Same College Age 30+	0.77	0.76	0.75	0.76	0.76	0.74	1.00										
(8) Both Female	0.97	0.87	0.95	0.94	0.97	0.79	0.71	1.00									
(9) Both Male	0.94	0.95	0.94	0.95	0.88	0.86	0.77	0.84	1.00								
(10) Both age <30	0.73	0.78	0.76	0.77	0.64	0.77	0.51	0.65	0.81	1.00							
(11) Both age 30-54	0.93	0.87	0.92	0.92	0.90	0.84	0.77	0.91	0.87	0.66	1.00						
(12) Both age >54	0.75	0.61	0.70	0.68	0.79	0.43	0.50	0.80	0.58	0.25	0.54	1.00					
(13) Ages Within 5 Years	0.96	0.92	0.96	0.96	0.91	0.87	0.72	0.92	0.93	0.87	0.91	0.60	1.00				
(14) Family	0.67	0.62	0.63	0.63	0.67	0.57	0.64	0.61	0.64	0.37	0.65	0.44	0.57	1.00			
(15) Interaction in Last Month	0.96	0.88	0.92	0.92	0.96	0.78	0.78	0.95	0.86	0.56	0.88	0.80	0.86	0.73	1.00		
(16) Top Half Friendship	0.99	0.92	0.96	0.96	0.97	0.81	0.79	0.96	0.91	0.63	0.92	0.77	0.91	0.72	0.99	1.00	
(17) Top Decile Friendship	0.96	0.91	0.93	0.93	0.95	0.80	0.80	0.93	0.89	0.60	0.91	0.74	0.87	0.78	0.99	0.99	1.00

Note: Table presents correlations between *SocialConnectedness* across NY CSA zip codes constructed in Eq. 1 and similar measures constructed from restricted sets of connections. Row and column 1 is the original measure. Rows and columns 2-5 limit to connections made less than a year ago, between 1 and 5 years ago, less than five years ago, and more than 5 years ago. Rows and columns 6-7 limit to connections between individuals with the same high school and college information. Rows and columns 8-9 limit to connections between females and males. Rows and columns 10-13 limit to connections between individuals under 30, individuals between 30 and 54, individuals over 54, and individuals with ages 5 or fewer years apart. Row and column 14 limits to family connections. Rows and columns 15-17 limit to individuals that have interacted on Facebook in the past month, and individuals that are in the top half and decile of a measure of friendship strength.

product of the number of eligible Facebook users in those zip codes, as in Eq. 1 (see Bailey et al., 2018a, for the first use of this *Social Connectedness Index*, or SCI). This measure represents the relative probability of a Facebook friendship link between a given user in zip code i and a given user in zip code j:

$$SocialConnectedness_{i,j} = \frac{FB\_Connections_{i,j}}{FB\_Users_i \times FB\_Users_j}. \tag{1}$$

We can also construct similar measures that are restricted to certain sets of friendships, for example friends who went to the same high-school. Doing so allows us to better understand the types of connections underlying our original measure and the geography of the different components of individuals' social networks. Table 1 presents the cross-correlation of a number of these friendship-restricted measures of $SocialConnectedness_{i,j}$ for zip code-pairs in the New York CSA. These include restrictions on length of friendship (e.g., friends for <1 year), shared characteristics (e.g., same high school, same gender), and friendship strength (e.g., interacted on Facebook in the past month, in the top decile of friendship strength). Generally speaking, these measures are highly correlated with each other as well as with our baseline measure that is based on all friendship links. This provides evidence that the different components of individuals' social networks follow similar geographic patterns. We explore the geographic concentration of these different components in additional detail in Section 4. This result also provides more evidence that our measure captures real-world networks.

2. Determinants of urban social connectedness

In this section, we explore a number of factors that might shape the formation of social connections in urban environments. In Section 2.1, we present a number of case studies that show the geography of social networks between urban zip codes and the distribution of certain transit routes. In Section 2.2, we estimate the elasticity of social connectedness with respect to geographic distance, transit availability, and demographic similarity. In Section 2.3, we apply an agglomerative clustering algorithm to generate communities that maximize within-group social connectedness.

2.1. Social connectedness in urban areas: Case studies.

Panels A to D of Fig. 1 map the percentile ranks of $SocialConnectedness_{i,j}$ of all zip codes j in NYC to four zip codes i covering portions of the Upper East Side (10021), East Harlem (10035), Little Neck (11363), and Oakland Gardens (11364), respectively. Relevant transit links are included for illustration. In each panel, relatively more of the connections are to geographically close zip codes; beyond this general pattern, there is substantial heterogeneity in the social networks across the four zip codes. The focal zip codes in panels A and B are roughly two miles apart in uptown Manhattan. The distributions of their respective social networks differ considerably, but essentially all regions with strong social connectedness to these zip codes are linked via direct or one-transfer subway trips. Panel C maps the social network of residents of Little Neck, Queens, a neighborhood on the eastern edge of NYC with easy access to the Long Island Railroad (LIRR) into midtown Manhattan. Little Neck has strong social connectedness to residential areas in midtown Manhattan near the LIRR terminus. Panel D shows the social network of zip code 11364, covering the neighborhood of Oakland Gardens in Queens. While adjacent to Little Neck, which has two LIRR stops, Oakland Gardens does not itself have a LIRR stop. Its social network differs from that of Little Neck in that none of the zip codes in the top quartile of connectedness with Oakland Gardens are in Manhattan. The spatial distributions of the social networks presented in Fig. 1 provide the first suggestive evidence that NYC's public transit system plays an important role in enabling the formation and maintenance of social ties across geographic distances. Indeed, it appears as if transit links can effectively "shrink" the geographic distances between locations within the city.

The spatial distribution of social networks of zip codes across the New York CSA also exhibits patterns consistent with those explored for NYC zip codes. Panels E and F of Fig. 1 map the percentile rank of $SocialConnectedness_{i,j}$ for two zip codes i to all zip codes j in the New York CSA. Panel E shows the social connectedness to zip code 06511 in New Haven, CT. The social network of New Haven exhibits a strong state border effect along the New York-Connecticut border; it also has a notable instance of long-distance connectivity: over 100 miles away in New Jersey there is a cluster of strongly connected zip codes surrounding the town of Princeton, a feature that

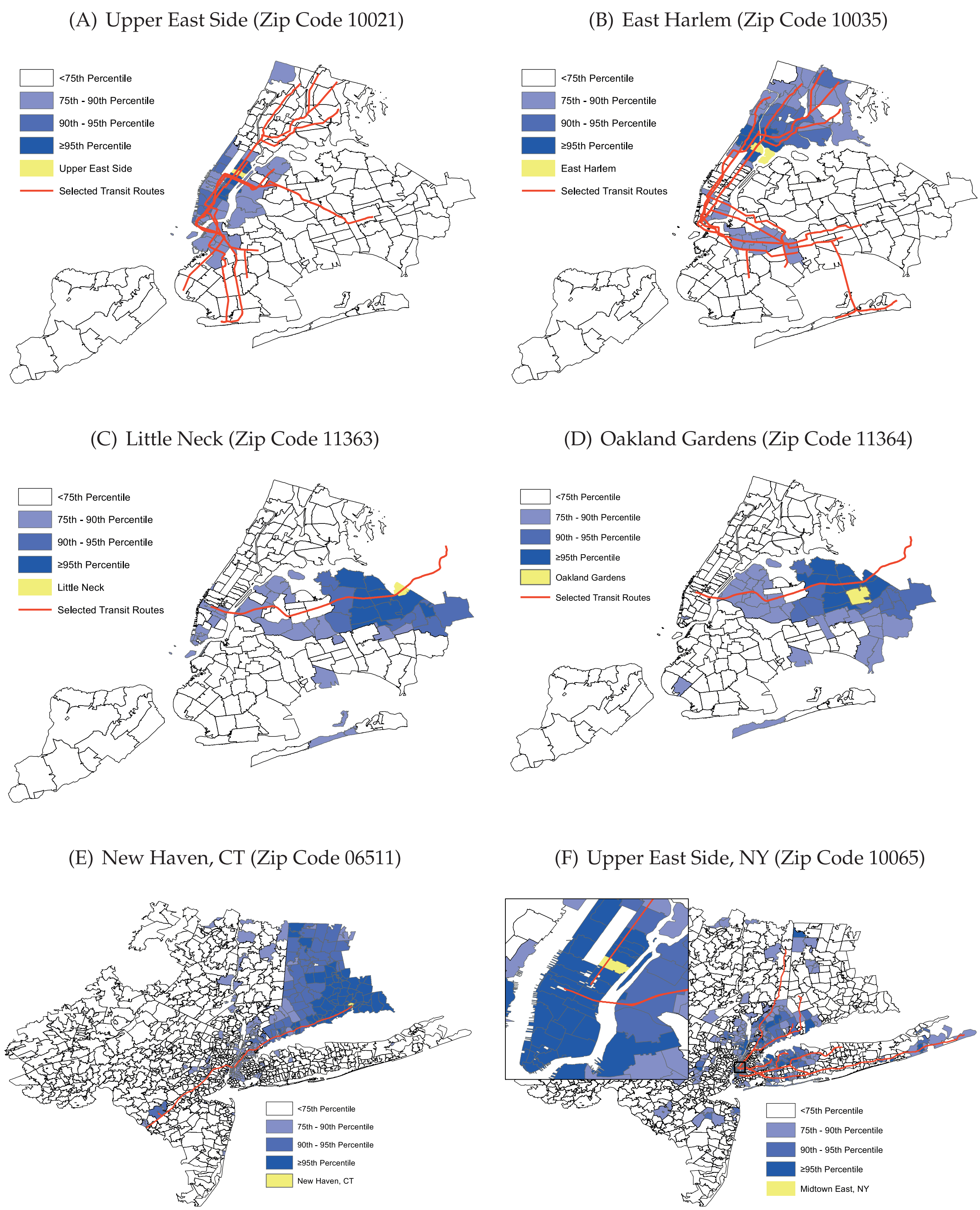

Fig. 1. Social Network Distributions.

Note: Figure shows social networks distributions along transit routes. Panels A, B, C, and D show the percentile rank of the relative probability of connection, as measured by $SocialConnectedness_{i,j}$, of all zip codes *j* in NYC to four zip codes *i* in the Upper East Side (Panel A), East Harlem (Panel B), Little Neck (Panel C), and Oakland Gardens (Panel D). Panels E and F show the percentile rank of the relative probability of connection, as measured by $SocialConnectedness_{i,j}$, of all zip codes *j* in the New York CSA to two zip codes *i* in New Haven, CT (Padnel E) and the Upper East Side, NY (Panel F). Darker zip codes have a greater probability of connection to a given zip code *i*.

is likely driven by students and researchers at Yale University (located in New Haven) and Princeton University; these connections are likely strengthened by the ease of train travel between New Haven and Princeton Junction. Panel F of Fig. 1 shows the social network of zip code 10065 in the Upper East Side, which is home to some of the wealthiest residential areas of NYC. This zip code exhibits strong social connectedness to the wealthy northern suburbs as well as to the Hamptons, a popular vacation destination for the well-heeled, roughly 100 miles away at the eastern tip of Long Island. Notably, there are stronger connections to parts of the Hamptons than to zip codes in Long Island City and Astoria, directly across the East River in Queens.

2.2. Geographic distance, social distance, and social connectedness

The previous section presented a number of case studies that suggest a relationship between social connectedness, geographic distance, transit availability, and demographic similarity. We next estimate the elasticity of social connectedness with respect to these objects more formally.

To systematically measure the ease of travel between two zip codes, we use the Google Maps API to collect travel times on public transit between the geographic centers of all zip codes on a weekday morning,[4] and measure cab cost in dollars using data from the New York Taxi and Limousine Commission (TLC).[5] There is substantial variation in transit travel time and cab costs over similar geographic distances in NYC. For example, Fig. 2 shows that the 95th percentile transit trip time between zip codes that are (roughly) 2.5 miles apart is only 4 minutes less than the 5th percentile trip time between zip codes that are 10 miles apart. Much of this variation is driven by a combination of public transit infrastructure and geography. For example, a transit trip between the East Village and Greenpoint, two neighborhoods facing one another across the East River that lack a connection via a tunnel or bridge, is at the 90th percentile of trip time compared to trips between other zip code-pairs separated by a similar geographic distance. Fig. 2 also shows there is substantial variation in the cost of cab trips for similar distances. As an alternative way of presenting the same information, Panel A shows a scatter plot demonstrating the variation of transit travel times between zip code-pairs that are a given distance apart from each other; Panel B shows the variation in cab costs for zip code-pairs that are a given distance apart.

To obtain a more systematic understanding of the effect of transportation links on social networks, we next use Eq. 2 to explore the pairwise friendship links between zip codes:

$$\log(SocialConnectedness_{i,j}) = \beta_0 + \beta_1 \log(d_{i,j}) + X_{i,j} + \psi_i + \xi_j + \epsilon_{i,j}. \quad (2)$$

The dependent variable is the log of social connectedness (defined in Eq. 1), and $\log(d_{i,j})$ denotes the log of the "distance" between i and j. Here, "distance" will be variously defined as the geographic distance between the central points of zip codes i and j, the public transit time between the central points of zip codes i and j, and the average cost of cab trips between zip codes i and j. Control variables $X_{i,j}$ include measures of the dissimilarity of the two zip codes along demographic and socioeconomic factors. These factors are income (the difference in average income across the zip code-pair), education (the difference in the shares of residents without a high school degree across the zip code-pair), and race (the difference in the non-Hispanic white shares of the populations across the zip code-pair). All specifications include fixed effects ψ_i and ξ_j for zip codes i and j, respectively.

Table 2 shows the results of the regression 2 with $\log(d_{i,j})$ representing geographic distance in columns 1 and 2, public transit time in columns 3 and 4, and cab cost in columns 5 and 6. Columns 1, 3, and 5 are the baseline specifications as shown in regression 2. Columns 2, 4, and 6 include interaction terms for pairs of zip codes that are both in the top third of the income distribution and pairs that are both in the bottom third of the income distribution with $\log(d_{ij})$, to test if the social connectedness of zip codes responds differently to transit times or cab costs based on differences in zip code incomes. When we compare columns 1 and 3, we find that the coefficient for transit time is over 60% greater in magnitude than that for geographic distance. The estimates imply that a 10% greater geographic distance between zip codes is associated with 8.7% lower social connectedness, while a 10% increase in public transit time is associated with 14.2% lower social connectedness. Likewise, column 5 indicates that a 10% increase in cab cost is associated with a 10.6% decline in social connectedness. These results suggest that public transportation infrastructure plays a more important role in the formation of social networks in urban settings than simple geographic distance does.

Table 2 also documents that, beyond the various measures of distance, zip codes that are more similar in terms of their education levels and their racial composition are more likely to be socially connected, providing evidence for homophily within New York City. For example, conditional on the geographic distance and differences in income and education levels, a 10 percentage point increase in the difference in the share of the population that is white is associated with a 11% to 12% decline in social connectedness. Similarly, a 10 percentage point increase in the difference of the population shares with no high school is associated with a 7%-10% decline in social connectedness. While differences in income do not imply differences in social connectedness (once we condition for differences in racial composition and educational attainment), we do find that the elasticity of social connectedness to the various measures of distance is larger when zip codes have very different income measures. In particular, columns 2, 4, and 6 show that the effect of increasing distance on social connectedness is smaller across zip code-pairs with similar incomes (i.e., zip code-pairs where both zip codes are in the top tercile or those where zip codes are in the bottom tercile of the income distribution). Said differently, reducing travel times appears to have a disproportionate effect on fostering social connectedness across regions with different incomes.

In order to examine how distance affects social connectedness at the CSA level, Table 3 shows the result from performing regression 2 for zip codes across the New York CSA. Since many of these zip codes are not well connected via public transport, we only use the log of geographic distance as the measure of $\log(d_{i,j})$. Column 1 excludes zip code fixed effects and socioeconomic dissimilarity variables $X_{i,j}$, and column 2 includes zip code fixed effects but excludes socioeconomic dissimilarity variables $X_{i,j}$. Column 3 includes an additional variable indicating whether both zip codes are within the same state. Column 4 includes differences in demographic variables, and column 5 adds interaction terms

[4] Our data on public transit time is collected from the Google Distance Matrix API. For each pair of zip codes i and j we collected the transit time of a trip from i to j and from j to i. Trips originate and end at the geographic centers of each zip code. The transit time measure between two zip codes ($TravelTime_{i,j}$) is the average time of trip i to j and trip j to i. We queried Google for travel times on a weekday, March 15th, 2017. We queried travel times more than two weeks in advance, so that contemporaneous delays or construction work would not influence trip times. We pulled the travel time on a weekday morning for a traveler that has to arrive at the other zip code by 9AM, to estimate travel time on a work day.

[5] The TLC reports the data for each cab trip taken in the first six months of 2016. The latitude and longitude of the origin and destination of 19.7 million trips, composed of 11.2 million yellow cab trips and 8.5 million green cab ("borough cab") trips, were matched to their origin and destination zip codes. For green cabs, which primarily serve the outer boroughs, all trips taken in the first six months of 2016 were matched to zip codes. For yellow cabs, which provide a greater share of trips but are more concentrated in Manhattan, only trips in March 2016 were matched to zip codes. The cost of a trip from zip code i to zip code j is calculated as the average of the costs of all trips originating in zip code i and ending in zip code j. We only consider zip codes that have at least one trip in each direction, and calculate the cost of travel between a zip code-pair composed of zip codes i and j as the average of the cost of trips from i to j and trips from j to i.

	Transit Travel Times Between Zips That Are:			Cab Trip Costs Between Zips That Are:		
	2.5 Miles Apart	5 Miles Apart	10 Miles Apart	2.5 Miles Apart	5 Miles Apart	10 Miles Apart
Mean	0:33:51	0:51:16	1:18:45	$15.61	$26.23	$47.40
P5	0:18:45	0:32:46	0:53:44	$12.39	$20.83	$39.12
P10	0:20:51	0:34:06	0:56:45	$13.10	$22.08	$41.52
P25	0:25:53	0:41:23	1:07:49	$13.71	$23.96	$44.48
Median	0:33:44	0:49:51	1:17:51	$14.82	$25.81	$47.23
P75	0:39:43	0:59:12	1:28:22	$16.42	$28.44	$50.63
P90	0:47:02	1:08:43	1:40:17	$18.59	$31.03	$53.56
P95	0:49:59	1:17:49	1:47:04	$23.50	$32.16	$54.73
N	138	224	225	131	194	94

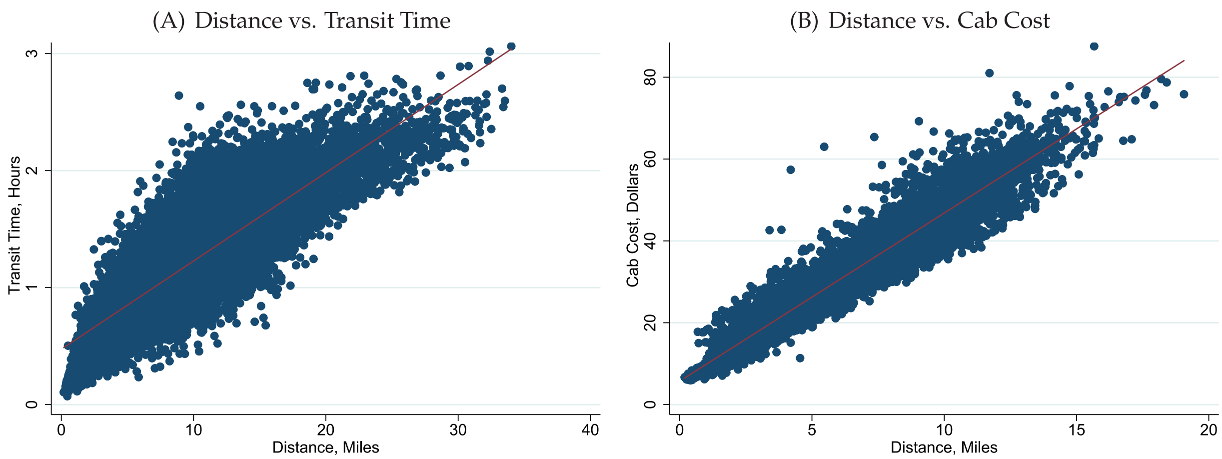

Fig. 2. Travel Time and Costs Variation by Distance, NYC.
Note: Table shows across-zip-code-pair summary statistics for transit time and cab trip cost between zip codes in a zip code-pair at various distances. All travel times and cab costs for zip code-pairs that are between ± .1 miles of the indicated distance are included in each column. Not all zip code-pairs were traveled between via cab during our sample period, allowing us to only calculated cab trip costs for a subset of zip code-pairs. We also show scatter plots at the zip code-pair level of transit time (Panel A) and cab trip cost (Panel B) on the vertical axes. The horizontal axes for both panels show the geographic distance between the centers of each zip code pair.

Table 2
The Effect of Distance and Transportation on Social Connectedness, NYC

	(1)	(2)	(3)	(4)	(5)	(6)
Log(Distance in Miles)	-0.872***	-0.951***				
	(0.044)	(0.044)				
Log(Avg. Time on Transit)			-1.418***	-1.498***		
			(0.067)	(0.071)		
Log(Avg. Cab Cost)					-1.059***	-1.113***
					(0.044)	(0.048)
Δ Share Pop White (%)	-0.012***	-0.011***	-0.013***	-0.012***	-0.011***	-0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Δ Share Pop No High School (%)	-0.010***	-0.007***	-0.009***	-0.007***	-0.008***	-0.006**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
Δ Avg. Income (k$)	-0.000	-0.001	-0.001	-0.000	-0.001	-0.001**
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Interaction: Rich Zip-Pair		0.234***		0.150*		0.444***
(X Dist., Transit Time, or Cab Cost)		(0.064)		(0.083)		(0.072)
Interaction: Poor Zip-Pair		0.199***		0.274***		0.020
(X Dist., Transit Time, or Cab Cost)		(0.050)		(0.080)		(0.064)
Dummy for Zip-Pair Type		Y		Y		Y
Zip Code Fixed Effects	Y	Y	Y	Y	Y	Y
Number of Observations	16,283	16,283	16,283	16,283	7,873	7,873
R-Squared	0.759	0.765	0.759	0.763	0.836	0.841

Note: Table shows results from regression 2. The unit of observation is a zip code-pair. The dependent variable in all columns is the log of *Social-Connectedness* as defined in Eq. 1. All specifications include zip code fixed effects and measures of the similarity of zip codes within the pair along socioeconomic and demographic dimensions. The measure of "distance" in regression 2 is variously defined as geographic distance (columns 1-2), transit time (columns 3-4), and cab cost (columns 5-6). Columns 2, 4, and 6 include interaction terms for rich zip code-pairs and poor zip code-pairs with "distance." Coefficients for the dummy variables are excluded for brevity. Standard errors are double clustered by each zip code i and zip code j in a zip code-pair. Significance levels: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

Table 3
The Effect of Geographic Distance on Social Connectedness, New York CSA

	(1)	(2)	(3)	(4)	(5)
Log(Distance in Miles)	-1.229***	-1.582***	-1.383***	-1.268***	-1.329***
	(0.026)	(0.023)	(0.024)	(0.025)	(0.024)
Same State			1.081***	1.158***	1.155***
			(0.040)	(0.039)	(0.039)
Δ Share Pop White (%)				-0.010***	-0.010***
				(0.001)	(0.001)
Δ Share Pop No High School (%)				-0.021***	-0.017***
				(0.002)	(0.002)
Δ Avg. Income (k$)				-0.006***	-0.005***
				(0.000)	(0.000)
Interaction: Rich Zip-Pair					0.183***
(X Dist. Miles or Transit Time)					(0.030)
Interaction: Poor Zip-Pair					0.196***
(X Dist. Miles or Transit Time)					(0.032)
Dummy for Zip-Pair Type					Y
Zip Code Fixed Effects		Y	Y	Y	Y
Number of Observations	625,743	625,741	625,741	625,741	625,741
R-Squared	0.389	0.714	0.738	0.784	0.788

Note: Table shows results from regression 2 for zip code-pairs in the New York CSA. The unit of observation is a zip code pair. The dependent variable in all columns is the log of *SocialConnectedness* as defined in Eq. 1. The measure of "distance" is geographic distance in all specifications. Column 1 does not include zip code fixed effects and controls. Column 2 includes zip code fixed effects. Column 3 incorporates a control variable for zip codes that are in the same state. Column 4 adds measures of the similarity of zip codes along socioeconomic and demographic dimensions. Column 5 additionally includes interaction terms for rich zip codes and poor zip codes. Coefficients for the dummy variables for the various zip pair types are excluded for brevity. Standard errors are double clustered by each zip code *i* and zip code *j* in a zip code-pair. Significance levels: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

for rich zip code-pairs and poor zip code-pairs, defined as above. The effect of geographic distance on social connectedness is greater for zip codes across the CSA than it is for the subset of zip codes within NYC. This is consistent with prior research demonstrating that urban social networks are less geographically determined than those over larger areas (Herrera-Yague et al., 2015). The coefficients on distance in these regressions are generally smaller in magnitude than the ones for regressions in earlier research by Bailey et al. (2018b) at the county level for counties within 200 miles of one another (this is the relevant comparison, as there are very few zip code-pairs more than 200 miles apart in the New York CSA). This difference may be due to differences in the properties of social networks measured at this finer level of aggregation, or due to our sample of zip codes centered on a large urban area where the effect of distance is weaker.

Column 3 of Table 3 shows that the social connectedness between two zip codes in the same state is about twice as large as the connectedness between equidistant zip codes across different states; this same-state effect was already visible in Panel E of Fig. 1. This could, for example, be the result of school districts that do not cross state lines; other possible explanations include the role of occupational licensing in restricting cross-state moves, and thereby cross-state friendship formation. The negative coefficients on all of the socioeconomic dissimilarity measures in column 4 are suggestive of homophily in the New York CSA; homophily based on income and educational attainment seems stronger in the New York CSA relative to NYC, while homophily based on race appears similarly large. Finally, the estimates in column 5 support the findings from the within-NYC analysis: social connectedness across zip codes with populations of different income drops off faster in distance than the social connectedness across zip codes with more similar incomes.

2.3. Connected communities in New York

We next provide an alternative description of the geographic structure of social networks across the New York metro area. In particular, we use a hierarchical agglomerative linkage clustering algorithm to construct hypothetical "communities" of zip codes that maximize within-group social connectedness. This procedure allows us to determine which groups of zip codes groups are maximally connected to one another, and to compare the resulting connected communities to existing administrative boundaries, such as NYC boroughs or states.

The algorithm starts by considering each of the *N* zip codes in a region (either NYC or the New York CSA) as separate communities of size one. We define the "distance" between two zip codes as the inverse of $SocialConnectedness_{i,j}$ in Eq. 1. The two "closest" zip codes, based on their relationships with all other zip codes, are then merged into one larger community, thus producing $N-1$ total communities. The "distance" between the newly formed community *i* and each other zip code *j* is then calculated as the average of the "distances" for both of the constituent zip codes in the community to each zip code *j*. The two most connected communities are then again merged, producing $N-2$ total communities. This process continues until all zip codes are merged into a given number of "connected communities."

Panel A of Fig. 3 shows the result of grouping NYC zip codes into five connected communities. A large band of Brooklyn is clustered together with Harlem and the Bronx; interestingly, this connected community thus consists of two non-contiguous elements that are more connected with each other than they are with Manhattan, which lies between them. This finding again suggests that geographic distance might not be as relevant a measure of "distance" within dense urban areas as it is at other levels of aggregation. Manhattan below Harlem and Morningside Heights joins with a handful of neighborhoods across the East River in Brooklyn and Queens; Brooklyn south of Prospect Park to Coney Island is grouped with Staten Island; and the rest of Queens is split into a small northern community adjacent to LaGuardia Airport and a large eastern community.

We also repeat the hierarchical agglomerative clustering for all zip codes in the New York CSA. Panel B of Fig. 3 shows the resultant connected communities. At the CSA level, the algorithm groups the majority of Long Island with NYC. New Jersey's border with Pennsylvania and New York is mostly preserved, with the exception of a patch of New York that is grouped with northern New Jersey (in this area, several New Jersey Transit and MTA Metro North lines in New Jersey extend north into New York). Connecticut's border is also largely preserved. Both upstate New York and Pennsylvania are broken into numerous smaller communities. Unlike the connected communities constructed within NYC, all connected communities at the CSA level are contiguous.

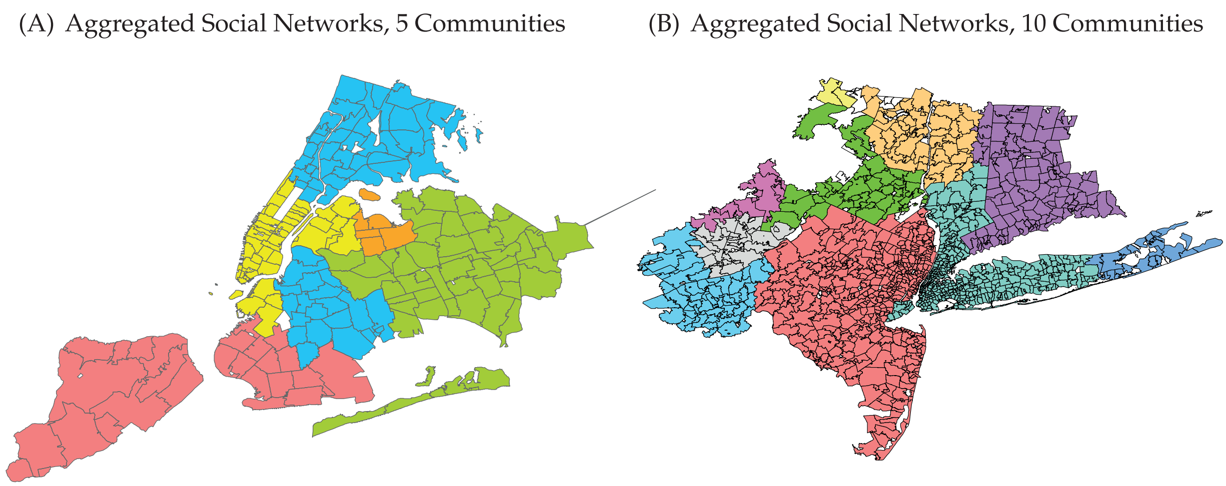

Fig. 3. Agglomerative Linkage Clustering of Communities.
Note: Figure shows the results of the hierarchical agglomerative linkage clustering algorithm. Panel A shows NYC zip codes grouped together to create 5 connected "communities." Panel B shows New York CSA zip codes grouped together to create 10 "communities."

3. Social connectedness and across-zip code flows

Social networks may be related to important flows of economic interaction. For example, Barwick et al. (2019) document that phone call volumes correlate strongly with worker flows, suggesting that information provided by social contacts are critical for labor market performance. Similarly, Schmutte (2015) finds that job referral networks are important for matching high-ability workers to high-paying jobs. Prior research also documents relationships between social connectedness and migration patterns between US counties (Bailey et al., 2018b), travel between European regions (Bailey et al., 2020b), and trade between nations and subnational regions (Bailey et al., 2020a). In this section, we explore the relationship of social connectedness between zip codes in New York City with two measures of flows of economic interactions: the number of home to work commutes and the number of cab trips.[6] While we cannot prove the causal effect of social networks, there are many mechanisms through which they could affect flows. Job opportunities in a given geographic area, for instance, may be shared by those who live there with individuals in their social network. To study the relationship between social networks and economic interactions we use equation 3:

$$\log(EconomicFlow_{i,j}) = \beta_0 + \beta_1 \log(SocialConnectedness_{i,j}) + X_{i,j} + \psi_i + \xi_j + \epsilon_{i,j}. \quad (3)$$

The dependent variable is the log of an economic flow — either commutes or cab flows from zip code i to zip code j. Control variables X_{ij} again include demographic and socioeconomic measures of dissimilarity. All specifications include fixed effects ψ_i and ξ_j for zip codes i and j, respectively.

This regression framework, however, does not allow us to rule out stories of reverse causality in which economic interactions cause social networks to form. For example, it is possible that individuals from zip code i that work in zip code j meet individuals living in zip code j while traveling to and from work, driving a positive relationship between social connectedness and commuting flows. To also explore the relationship between social networks and economic flows in the absence of this reverse causality, we use an Instrumental Variables (IV) approach. Specifically, we instrument social connectedness on similar measures constructed from only family connections and connections between individuals who attended the same high school. These high school and family connection measures are highly correlated with our overall connection measure, as shown in Section 1. Furthermore, individuals from the same high school or family are likely to be connected independently of things such as commuting flows. Note, though, that while this IV strategy allows us to address concerns about the role of reverse causality in explaining our results, it does not completely address all stories of omitted variables that could be correlated also with the structure of highschool friendship networks. We therefore refrain from a causal interpretation of the IV estimates.

Letting $\log(\widehat{SocialConnectedness}_{i,j})$ be the prediction given by the regression in Eq. 4, the first and second stages of the IV regressions are given by Eqs. 4 and 5, respectively:

$$\log(SocialConnectedness_{i,j}) = \rho_0 + \rho_1 \log(SocialConnectednessRestricted_{i,j}) + X_{i,j} + \psi_i + \xi_j + \epsilon_{i,j} \quad (4)$$

$$\log(EconomicFlow_{i,j}) = \beta_0 + \beta_1 \log(\widehat{SocialConnectedness}_{i,j}) + X_{i,j} + \psi_i + \xi_j + \epsilon_{i,j} \quad (5)$$

Table 4 shows results from regressions 3 and 5. Panel A shows results using the log of the number of individuals who live in zip code i and commute to work in zip code j. A 10 percent increase in connectedness between a pair of zip codes is associated with more than a 3 percent increase in the number of commuting flows between the two, even after controlling for socioeconomic factors as well as geographic and transportation distance. Columns 4 and 5 of Panel A present the IV estimates using social connectedness constructed between individuals that attended the same high school and have a family connection, respectively. The coefficient on social connectedness does not change with the family instrument and increases with the same high school instrument. This finding is inconsistent with a story that argues that the relationship between social networks and commuting patterns is due to commuting driving social connections. Instead, while we cannot rule out all stories of omitted variables, our findings are suggestive of social networks contributing to the location of individuals' commutes in accordance with Barwick et al. (2019).

Panel B shows the results similar regressions, where the dependent variable is the the log of the number of cab trips that start in zip code i and end in j as the economic flow of interest. Adjusting for socioeconomic factors and measures of distance, a 10 percent increase in connectedness between zip codes is associated with a 6 percent increase in

[6] Data on commuting patterns are available at the block level from the Census Bureau's LEHD Origin-Destination Employment Statistics (LODES).

Table 4
Social Connectedness and Across-Zip Code Flows

(A) log(Across-Zip Code Commuting Flows)					
	(1)	(2)	(3)	(4)	(5)
Log(Distance in Miles)	-0.240***	-0.147***	-0.160***	-0.161***	-0.160***
	(0.058)	(0.053)	(0.050)	(0.050)	(0.050)
Log(Avg. Time on Transit)	-0.204***	-0.095***	-0.084***	-0.049*	-0.087***
	(0.034)	(0.030)	(0.028)	(0.027)	(0.030)
Log(Avg. Cab Cost)	-0.877***	-0.678***	-0.627***	-0.561***	-0.632***
	(0.065)	(0.058)	(0.055)	(0.055)	(0.065)
Log(SCI)		0.306***	0.341***	0.422***	0.334***
		(0.024)	(0.024)	(0.031)	(0.044)
Δ Share Pop White (%)			0.002***	0.003***	0.002***
			(0.001)	(0.001)	(0.001)
Δ Share Pop No High School (%)			0.002	0.003	0.002
			(0.002)	(0.002)	(0.002)
Δ Avg. Income (k$)			-0.001**	-0.001**	-0.001**
			(0.000)	(0.000)	(0.000)
IV on Log(SCI)				Log(Same High School SCI)	Log(Same Family SCI)
Zip Code Fixed Effects	Y	Y	Y	Y	Y
Number of Observations	15,541	15,541	15,541	15,541	15,541
R-Squared	0.928	0.938	0.938	0.742	0.744
(B) log(Across-Zip Code Cab Flows)					
	(1)	(2)	(3)	(4)	(5)
Log(Distance in Miles)	0.061	0.252***	0.270***	0.268***	0.269***
	(0.112)	(0.095)	(0.095)	(0.093)	(0.094)
Log(Avg. Time on Transit)	-0.394***	-0.176**	-0.193***	-0.135*	-0.166**
	(0.081)	(0.068)	(0.069)	(0.067)	(0.063)
Log(Avg. Cab Cost)	-3.034***	-2.636***	-2.638***	-2.529***	-2.588***
	(0.158)	(0.132)	(0.134)	(0.135)	(0.135)
Log(SCI)		0.613***	0.600***	0.734***	0.662***
		(0.037)	(0.039)	(0.051)	(0.050)
Δ Share Pop White (%)			0.001	0.002**	0.001
			(0.001)	(0.001)	(0.001)
Δ Share Pop No High School (%)			0.001	0.002	0.001
			(0.003)	(0.003)	(0.002)
Δ Avg. Income (k$)			-0.001***	-0.001***	-0.001***
			(0.000)	(0.000)	(0.000)
IV on Log(SCI)				Log(Same High School SCI)	Log(Same Family SCI)
Zip Code Fixed Effects	Y	Y	Y	Y	Y
Number of Observations	15,755	15,755	15,755	15,755	15,755
R-Squared	0.912	0.930	0.930	0.873	0.874

Note: Table shows results from regressions 3 and 5 for zip code-pairs in New York City. The unit of observation is a single-direction zip code pair. The dependent variable is the number of commutes from zip code i to j in Panel A and the number of cab trips from zip code i to j in Panel B. In both panels, columns 1 and 2 exclude then include the log of *SocialConnectedness* as defined in Eq. 1. Column 3 includes socioeconomic and demographic controls. Columns 4 and 5 display results from the IV specification, using the log of social connectedness constructed from individuals from the same high school and family connections, respectively, as instrumental variables. Observations are limited to pairs with cab flows in both directions in Panels A and B, and pairs with commuting flows in Panel A. Standard errors are double clustered by each zip code i and zip code j in a zip code-pair. Significance levels: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

the number of cab trips. A decrease in transit time is associated with an increase in cab trips in all regression, likely reflecting similar commuting patterns across both modes of transportation. Interestingly, unlike in panel A, the coefficient for distance in miles is always positive. This suggests that, as the distance between two zip codes increases cab trips become more frequent, holding all else constant. This result is in line with expectations as cab trips are often quicker for longer trips. In the following section we will show evidence that zip codes with higher average incomes tend to have more dispersed social networks, which could also contribute to this result as cab trips are more expensive. Columns 4 and 5 present the IV estimates. Again, the coefficient on the log of social connectedness increases, allowing us to rule out reverse causality.

4. The geographic concentration of social networks

We next document heterogeneity in the geographic concentration of social networks across zip codes. We also explore which factors are associated with the geographic dispersion of these networks, and we investigate the relationship between the geographic dispersion of social networks and socioeconomic outcomes such as income, education, and entrepreneurship.

4.1. Measurement of social network concentration

We consider two measures of the geographic concentration of social networks: the share of friends that lives within a certain geographic radius (e.g., 1 mile or 5 miles), and the share of friends that lives within a certain number of people (e.g., within the nearest 1 million or 5 million people).

To construct our concentration measures for small distances such as one mile, we have to determine which friends are included within this range, even though we only observe the locations of individuals and their friends at the zip code level. We therefore construct our measures by weighting friendships to individuals in each region j by the population-weighted share of census blocks in region j that are within that distance of the population-weighted center of zip code i. Specifically, we use the following equation to construct our measure of the

Table 5
Summary Statistics of Geographic Concentration of Social Networks

(A) NYC

	Share of Friends Living Within:			Share of Friends Among Nearest:		
1 Mile	5 Miles	10 Miles	250K	People	1 Mil. People	10 Mil. People
Mean	6.3%	29.3%	44.0%	10.2%	21.9%	55.1%
P5	3.3%	19.5%	34.1%	4.9%	13.1%	38.8%
P10	4.0%	22.2%	36.0%	5.9%	14.2%	40.2%
P25	4.8%	26.2%	39.5%	7.3%	17.4%	49.3%
Median	6.2%	29.0%	44.0%	9.5%	22.0%	58.2%
P75	7.4%	32.4%	48.4%	11.6%	25.4%	61.2%
P90	8.8%	37.2%	52.7%	14.4%	29.5%	64.2%
P95	10.0%	39.6%	53.8%	19.6%	32.7%	66.0%

(B) New York CSA

	Share of Friends Living Within:			Share of Friends Among Nearest:		
	5 Miles	10 Miles	50 Miles	1 Mil. People	10 Mil. People	50 Mil. People
Mean	25.8%	38.1%	64.1%	32.3%	57.9%	75.4%
P5	12.2%	22.4%	48.6%	14.8%	40.6%	61.5%
P10	14.8%	25.1%	53.0%	17.5%	47.3%	65.8%
P25	20.2%	31.1%	60.0%	22.9%	54.6%	73.8%
Median	25.8%	38.4%	65.9%	31.4%	59.1%	76.8%
P75	31.0%	45.7%	69.1%	41.4%	62.4%	79.2%
P90	37.1%	51.1%	72.0%	49.2%	66.1%	81.1%
P95	40.4%	53.7%	73.4%	52.3%	68.5%	81.9%

Note: Table shows summary statistics of the geographic concentration of social networks. Panel A shows across-zip code summary statistics of the share of domestic friends of a zip code's population that live within 1, 5, and 10 miles of a zip code, and the share of domestic friends of a zip code's population that are among the nearest 250 thousand, 1 million, and 10 million people in and surrounding a zip code for NYC. Panel B shows across-zip code summary statistics of the share of domestic friends of a zip code's population that live within 5, 10, and 50 miles of a zip code, and the share of domestic friends of a zip code's population that are among the nearest 1 million, 10 million, and 50 million people in and surrounding a zip code for the New York CSA. Zip codes are weighted by their populations.

geographic concentration of zip code i's friendship network:

$$ShareWithinDMiles_i = \sum_j ShareFriends_{i,j} * \frac{\sum_{j_b} Pop_{j_b} * \mathbf{1}_{d_{i,j_b} \le D}}{TotalPop_j} \quad (6)$$

Here, d_{i,j_b} indicates the distance from the population-weighted center of zip code i to the population-weighted center of each census block j_b in region j. We find the population of each region j that is within a given distance D from zip code i by summing the population of all census blocks j_b for which d_{i,j_b} is less than D, and divide this by the total population of region j. We then weight the share of friends of zip code i living in region j, given by $ShareFriends_{i,j}$, by the share of the population of zip code j that lives within D miles of the center of zip code i, before summing over all regions j. We will use the following two objects as our measures of the geographic concentration of social networks. For our first measure, the share of friends living within a certain radius D representing one, five, ten, or fifty miles. For our second measure, the share of friends living within a certain number of people, we define D as the radius from the center of each zip code i that contains a given number of people, and then construct the statistics as above based on that distance.

Panel A of Table 5 provides summary statistics at the zip code level of the geographic concentration of social networks in NYC, based on the distribution of U.S. Facebook friends of users residing in each zip code. For the residents of the median zip code in NYC, 6.2% of U.S. friends live within one mile, 29.0% of U.S. friends live within five miles, and 44.0% of U.S. friends live within ten miles. There is significant heterogeneity in the geographic concentration of friendship links: across zip codes, the 5-95 percentile range of U.S. friends living within one mile is 3.3% to 10.0%. Similarly, Panel B of Table 5 provides summary statistics on the geographic concentration of U.S. friendship networks by zip code for the New York CSA. For the residents of the median zip code in the New York CSA, 25.8% of U.S. friends live within five miles, 38.4% of U.S. friends live within ten miles, and 65.9% of U.S. friends live within fifty miles. Once again, there is significant heterogeneity in the concentration of friendship connections: the 5-95 percentile range of friends living within ten miles is 22.4% to 53.7%.

In Table 6, we also explore the concentration of various components of individuals' social networks. The table presents, for zip codes in the New York CSA, the median zip code concentration of certain friendships. It also presents the correlation these concentrations across zip codes with the concentration of all connections. Overall, the median concentration is fairly similar across these different components and the across-zip code distributions are highly correlated. However, networks of individuals who attended the same college and are currently 30 or older (i.e. generally no longer in college) are less geographically concentrated than all other connections. This result is consistent with existing research on college experiences playing an important role in shaping individuals' social networks and, as a result, their access to opportunity and eventual labor market outcomes (see Zimmerman, 2019; Shue, 2013; Chetty et al., 2020). By contrast, networks between individuals who simply attended the same high school are *more* concentrated than other connections. This suggests that some of the heterogeneity in the concentration of friendship connections at the zip code level is driven by the share of individuals in that zip code that attended college.

In Panel A of Fig. 4 we map the spatial distribution of the share of U.S. friends living within five miles for each zip code in NYC. Zip codes with the most geographically dispersed friendship networks are primarily in the western area of Brooklyn and the eastern portion of Queens, as well as the Downtown and Midtown West neighborhoods of Manhattan. Panel B of Fig. 4 shows the spatial distribution of the share of friends within ten miles for zip codes within the New York CSA, revealing that networks are generally more geographically concentrated in the urban areas within the CSA, with high concentrations most evident in the area in and surrounding NYC but also present in New Haven, CT, Allentown, PA, and Seaside Heights, NJ.

We find similar heterogeneity in the share of friends living within a certain number of people: Panel A of Table 5 indicates that for median zip code, 22.0% of friendship links are to the one million closest individ-

Table 6
Geographic Concentration of Components of Social Networks - New York CSA

	Share Within 5 Miles		Share Within 10 Miles		Share Within 50 Miles	
	Median	Corr With All	Median	Corr With All	Median	Corr With All
(1) All	25.8%	1.00	38.4%	1.00	65.9%	1.00
(2) Added <1 Year Ago	26.2%	0.95	39.8%	0.94	67.8%	0.89
(3) Added 1-5 Years Ago	26.0%	0.98	39.5%	0.98	67.4%	0.98
(4) Added <5 Years Ago	26.1%	0.97	39.7%	0.97	67.6%	0.97
(5) Added >5 Years Ago	23.8%	0.97	35.4%	0.98	65.3%	0.99
(6) Same High School	30.7%	0.85	43.9%	0.84	71.2%	0.91
(7) Same College, Age 30+	14.1%	0.85	25.9%	0.85	53.5%	0.80
(8) Both Female	25.9%	0.99	38.2%	0.99	67.8%	0.98
(9) Both Male	25.0%	0.99	37.7%	0.99	65.3%	0.98
(10) Both age <30	30.1%	0.95	42.2%	0.95	70.0%	0.95
(11) Both age 30-54	21.3%	0.97	35.4%	0.98	64.7%	0.96
(12) Both age >54	20.2%	0.92	32.6%	0.92	59.8%	0.88
(13) Ages Within 5 Years	25.6%	1.00	37.6%	1.00	66.8%	0.99
(14) Family	21.2%	0.93	30.7%	0.95	56.5%	0.95
(15) Interaction in Last Month	26.7%	0.98	41.1%	0.98	69.5%	0.97
(16) Top Half Friendship	26.3%	0.99	40.5%	0.99	68.6%	0.98
(17) Top Decile Friendship	27.7%	0.98	42.2%	0.98	69.7%	0.96

Note: Table shows summary statistics of the geographic concentration of various components of social networks. Columns 1, 3, and 5 show the across-zip code median of the share of certain domestic friendships from a zip code's population to others that live within 5, 10, and 50 miles of a zip code, respectively, for the New York CSA. Columns 2, 4, and 6 show the correlation across zip codes between the shares for each friendship type and the shares for all friendships. A description of each friendship restriction is provided in the note of Table 1.

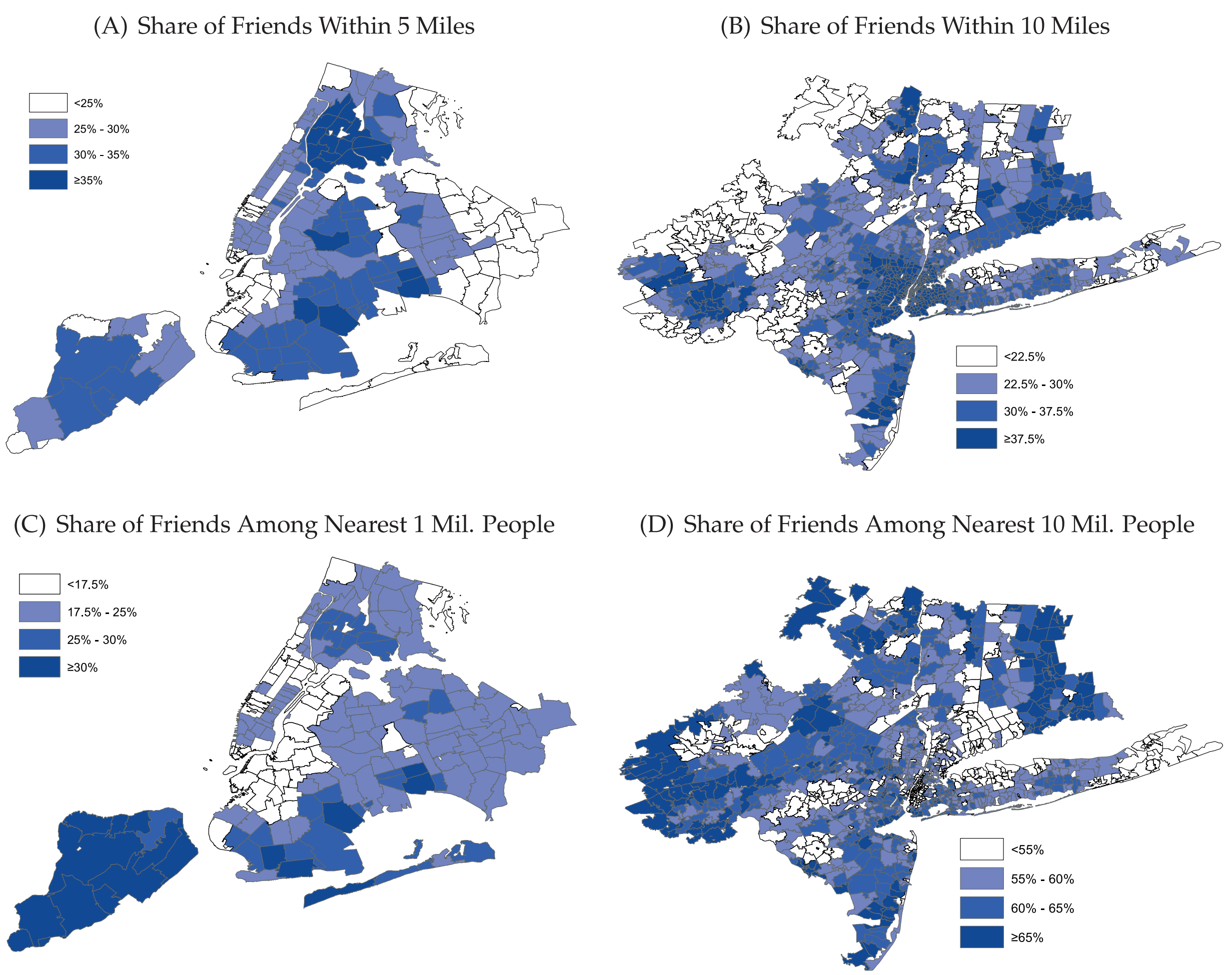

Fig. 4. Geographic Concentration of Social Networks.
Note: Figure shows the geographic concentration of social networks for New York zip codes. Panel A shows a map at the zip code level of the share of all U.S. friends that live within 5 miles for each NYC zip code. Panel B shows a map of the share of all U.S. friends that live within 10 miles for each zip code in the New York CSA. Panel C shows a map at the zip code level of the share of all U.S. friends that are among the nearest 1 million people for each NYC zip code. Panel D shows a map of the share of all U.S. friends that are among the nearest 10 million people for each zip code in the New York CSA.

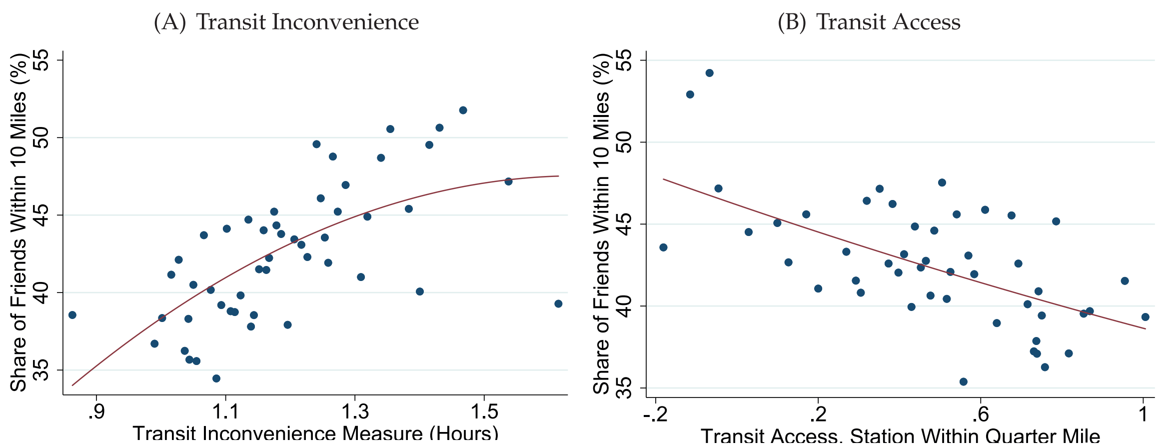

Fig. 5. Ease of Transit and Social Network Concentration, NYC.
Note: Figure shows binned scatter plots at the zip code level of the share of friends within 10 miles on the vertical axis. The horizontal axis in Panel A shows the average transit time from a zip code to all other zip codes within New York City as defined in Eq. 7. The horizontal axis in Panel B shows the share of the population of each zip code that is within a quarter mile of a subway or LIRR stop. Both panels include non-linear controls for population density (the population within 10 miles divided into 10 quantiles).

Table 7
Transit and the Geographic Dispersion of Social Networks, NYC

	(1)	(2)	(3)	(4)	(5)	(6)
Transit Inconvenience (Hours)		16.197***		12.682***		5.958*
		(3.133)		(2.572)		(3.025)
Share Pop 1/4 Mile from Transit (%)			-0.062***		-0.075***	-0.058***
			(0.016)		(0.012)	(0.015)
Share Pop White (%)	-0.004			-0.022	0.007	-0.004
	(0.021)			(0.020)	(0.019)	(0.020)
Share Pop No High School (%)	0.249***			0.257***	0.340***	0.323***
	(0.056)			(0.053)	(0.053)	(0.053)
Avg. Income (k$)	-0.017***			-0.010**	-0.011***	-0.009**
	(0.004)			(0.004)	(0.004)	(0.004)
Pop. Within 10 Miles (Millions)	0.868***	2.629***	1.107***	2.351***	1.440***	2.007***
	(0.245)	(0.494)	(0.326)	(0.379)	(0.243)	(0.375)
Number of Observations	182	182	182	182	182	182
R-Squared	0.446	0.144	0.089	0.513	0.542	0.552

Note: Table shows the results from regression 8. The unit of observation is a NYC zip code. The dependent variable is the share of friends that live within 10 miles. All columns include a control for population density. Column 1 includes demographic and socioeconomic characteristics. Column 2 includes only transit inconvenience as defined in Eq. 7. Column 3 includes only the share of the population of each zip code that is within a quarter mile of a subway or LIRR stop. Column 4 includes the variables from Columns 1 and 2. Column 5 includes the variables from Columns 1 and 3. Column 6 includes all variables. Significance levels: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

uals, but this number ranges from 13.1% to 32.7% between the 5th and the 95th percentiles of the zip code distribution. Panel B of Table 5 also highlights that for the residents of the median zip code in the New York CSA, 31.4% of U.S. friends are among the nearest one million people, 59.1% of U.S. friends are among the nearest ten million people, and 76.8% of U.S. friends are among the nearest fifty million people. There is also a high degree of heterogeneity for these measures: the 5th-95th percentile range of friends living among the nearest one million people is 14.8% to 52.3%.

Panel C of Fig. 4 maps the share of friends of individuals who live within the nearest one million people for NYC zip codes. Notably, the aggregate social networks of users in population-dense regions, such as those in north Brooklyn, are comparatively more dispersed in terms of the share of friends within a certain number of people than in terms of the share of friends living with in a certain geographic distance. The opposite pattern characterizes the social networks of Staten Island, the NYC borough with the lowest population density. Panel D of Fig. 4 shows the spatial distribution of social network density in the New York CSA, using as the measure the share of friends among the nearest 10 million people. The distribution is different from that in Panel B, as the urban cores and inner suburbs display less social network density using this measure. The differences are primarily driven by variation in population densities across urban and non-urban areas within the New York CSA.

4.2. Ease of transit and the concentration of social networks

Having established that there is substantial heterogeneity in the geographic concentration of social networks, we next explore whether differences in the public transit infrastructure across zip codes can explain part of this heterogeneity. We construct two measures of the ease of public transit at the zip code level, which we call “transit inconvenience” and “transit access.” Transit access is measured as the share of the zip code's population that lives within a quarter mile of a rail transit station.[7] Transit inconvenience is based on the travel times computed in

[7] A rail transit station is defined as either an MTA subway stop or a Long Island Railroad (LIRR) stop, as these are the two most important rail transit options within NYC. This transit access measure is intended to capture access to physical rapid transit infrastructure. Of course, zip codes may have access to other forms of public transit, and the measure of public transit time that we collect from Google allows for transit via any vehicle (trains as well as buses, ferries, trams, etc.), but rail transit provides the majority of public transit trips within the city (MTA, 2016a; 2016b).

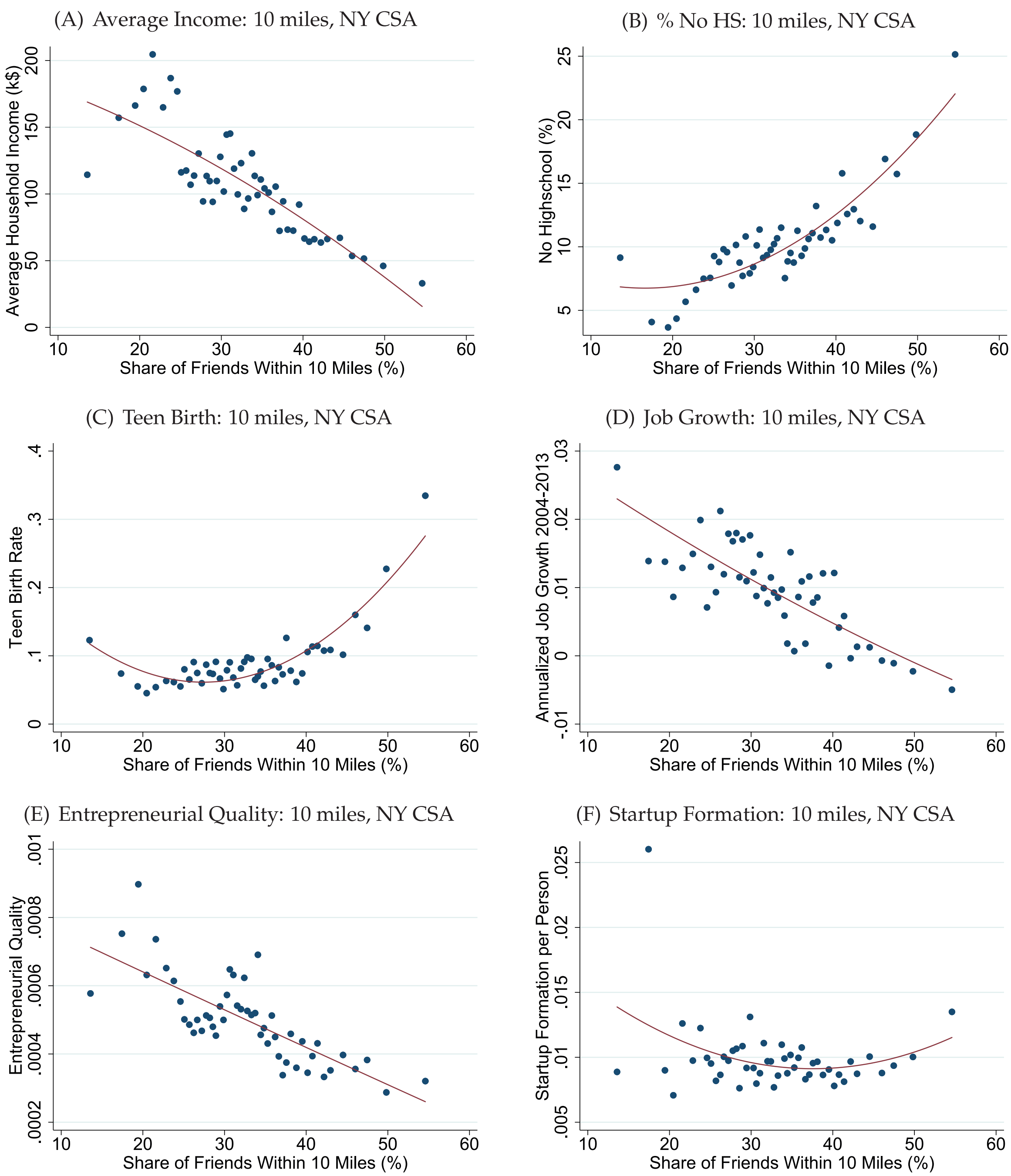

Fig. 6. Demographics and Geographic Concentration of Social Networks.
Note: Figure shows binned scatter plots with zip codes as the unit of observation. The horizontal axis plots the share of the U.S.-based friends of a zip code's population that live within 10 miles of a zip code. Each panel controls for the total population within 10 miles. Panels A and B plot the median income for residents of a zip code and the share of the zip code's population without a high school degree on the vertical axis, respectively. Panels C and D plot the teen birth rate for individuals born between 1978 and 1983 and annualized job growth from 2004-13 from Chetty et al. (2016). Panels E and F plot the "policy model" Entrepreneurial Quality Index and Startup Formation Rate per person from Andrews et al. (2019). The R-Squared values corresponding to the quadratic line of fit are: 9.3% (Panel A), 19.1% (Panel B), 26.1%(Panel C), 3.5% (Panel D), 7.0% (Panel E), 1.4% (Panel F).

Section 2, and constructed as the average of $TravelTime_{i,j}$ for each zip code i with all zip codes j over the number of zip code observations n_j:[8]

$$TransitInconvenience_i = \frac{\sum_j TravelTime_{i,j}}{n_j}. \tag{7}$$

Panel A of Fig. 5 shows a binned scatter plot of the relationship between the inconvenience of transit in NYC measured in hours (see Eq. 7) and the geographic concentration of social networks measured by the share of friends who live within the nearest 10 miles, controlling non-linearly for the number of individuals within 10 miles. The social net-

[8] All results in this section are similar for measures of transit inconvenience that weight each zip code j in Eq. 7 by the population in zip code j, so that having a longer travel time to a high-population zip code counts more towards transit inconvenience than high travel time to a low-population zip code does.

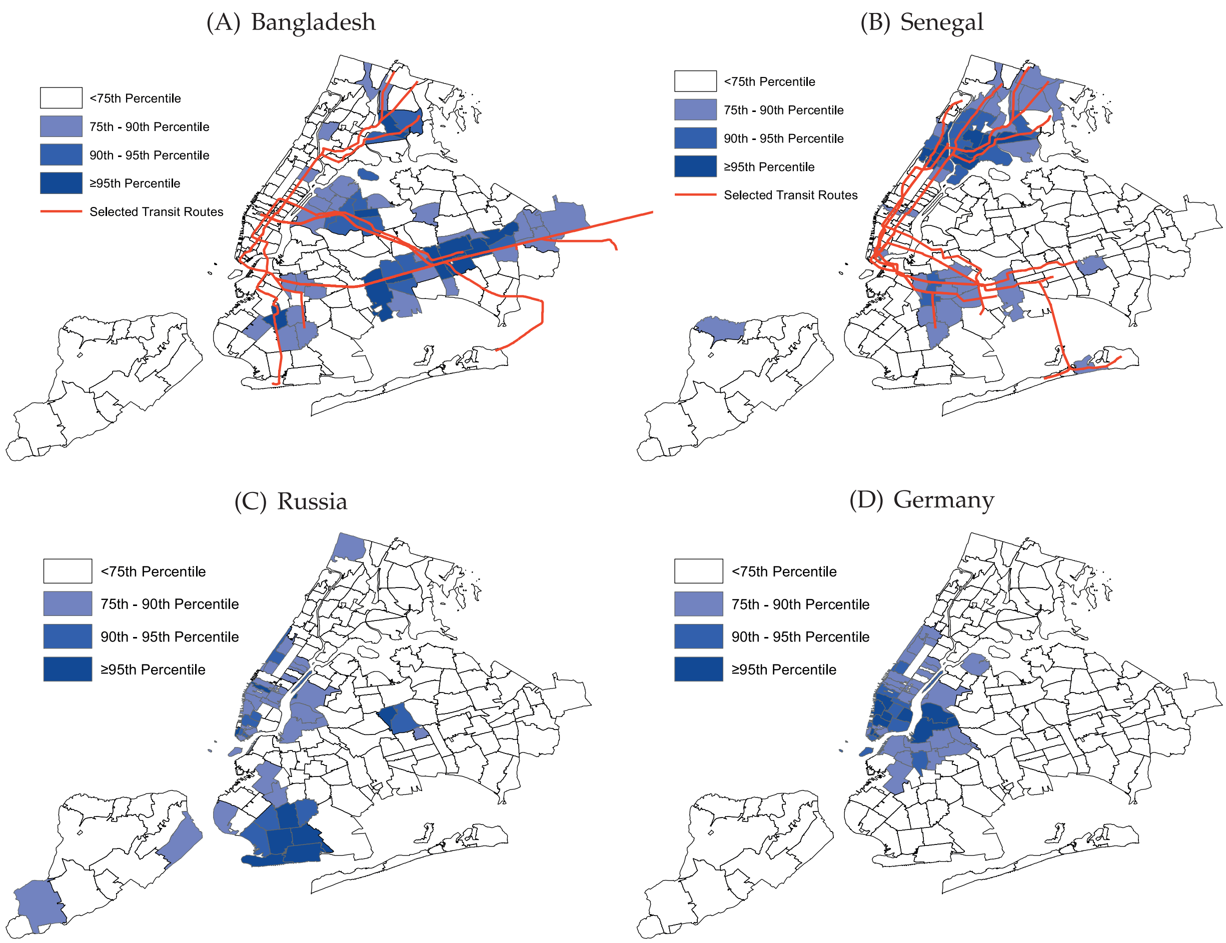

Fig. 7. International Connectivity, NYC.
Note: Figure shows the percentile rank of the probability of a friendship link, as measured by $SocialConnectedness_{i,j}$, of all zip codes j in NYC to four countries i: Bangladesh (Panel A), Senegal (Panel B), Russia (Panel C), and Germany (Panel D).

works of those zip codes with more convenient transit are less geographically concentrated compared to those with less convenient transit.

Panel B of Fig. 5 shows a binned scatter plot relating the transit access of zip codes to the share of friends living within the nearest 10 miles, again controlling non-linearly for population density. In this case, zip codes with greater access to public transit have more geographically dispersed social networks. Overall, the findings from these plots are consistent with the notion that ease of transportation is associated with a wider geographic dispersion of social networks.

There are many potentially confounding factors that could influence the relationship between the share of friends within certain distance or number of people and the ease of travel via public transit. For instance, due to the radial design of New York's subway system, all but one train service runs through the relatively wealthy areas of midtown or downtown Manhattan. To separately explore the role of transit infrastructure beyond the demographic measures that it is correlated with, we next estimate regression 8:

$$ShareWithin10mi_i = \beta_0 + \beta_1 Transit_i + \beta_2 X_i + \epsilon_i \tag{8}$$

The dependent variable is the geographic concentration of social networks, measured as the share of friends that live within the nearest 10 miles, though our conclusions are similar when using our other measures of the geographic concentration of social networks. Depending on the specification, $Transit_i$ will represent the transit inconvenience measure (Eq. 7), or the share of a zip codes' population within a quarter mile of a transit stop. X_i includes controls for socioeconomic characteristics of each zip code and a control for the share of the population within 10 miles.[9]

The estimates from regression 8 are presented in Table 7. Column 1 shows a relationship between demographic and socioeconomic characteristics and the geographic concentration of social networks. Columns 2 and 3 confirm that there is a statistically significant positive relationship between the transit inconvenience measure and the geographic concentration of social networks and a statistically significant negative relationship between transit access and the geographic concentration of the networks, respectively. Quantitatively, a 15 minute increase in the average travel time to all zip codes is associated with a 4.0 percentage point increase in the share of friends living within 10 miles, controlling for population density. Similarly, a ten percentage point increase in the zip code population that lives within a quarter mile of a transit stop is associated with a 0.6 percentage point decline in the share of friends living within 10 miles. Controlling for transit inconvenience and transit access substantially increases the R^2 of the regressions over and above the demographic controls, highlighting that each explains a sizable part of the across-zip-code heterogeneity in the concentration of social networks. Columns 4 and 5 show that the estimated relationship between social network concentration and the ease of public transit is largely unaffected by the addition of the demographic and socioeconomic con-

[9] While we present results controlling linearly for population density, our results are robust to non-linear population density controls.

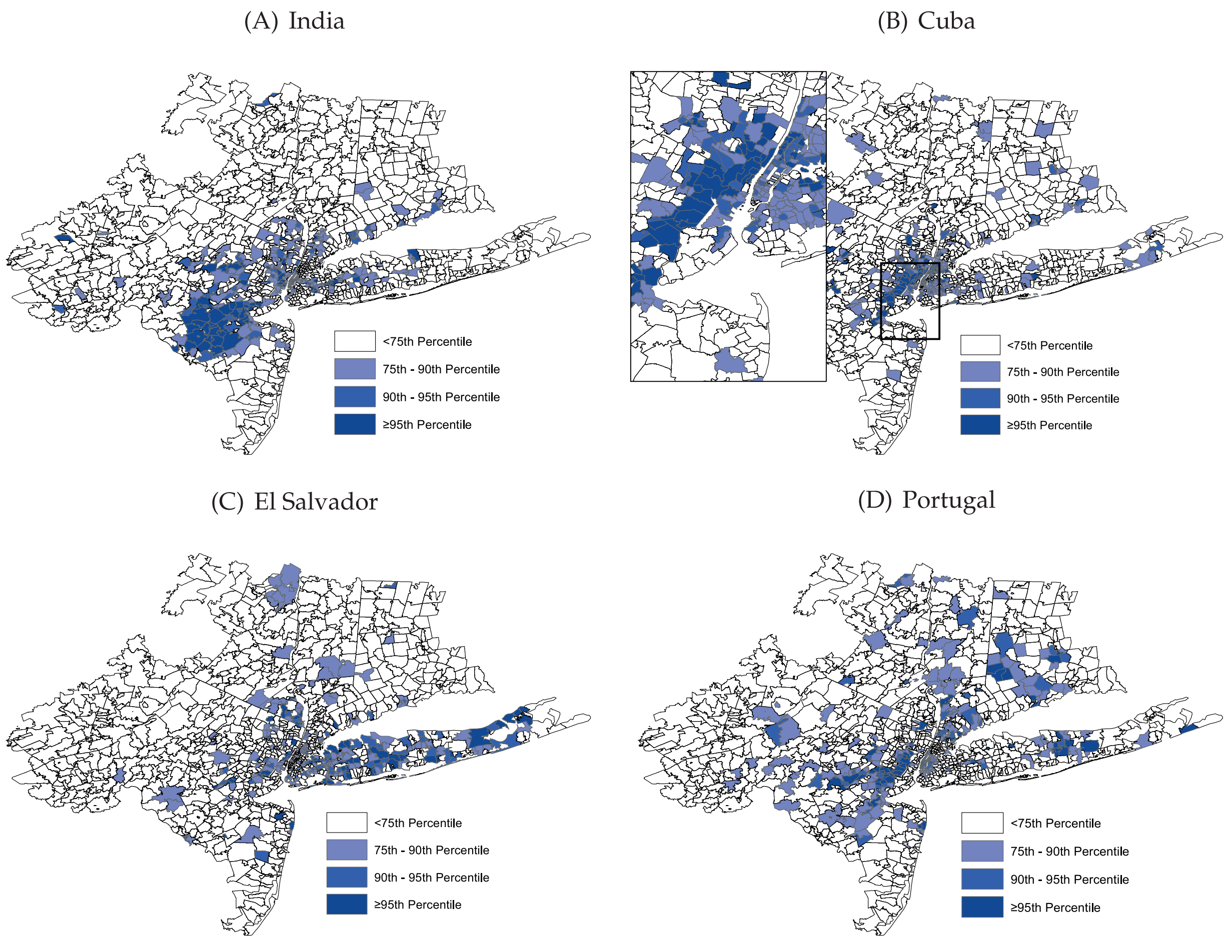

Fig. 8. International Connectivity, New York CSA.
Note: Figure shows show the percentile rank of the relative probability of connection, as measured by $SocialConnectedness_{i,j}$, of all zip codes j in the New York CSA to four countries i: India (Panel A), Cuba (Panel B), El Salvador (Panel C), and Portugal (Panel D).

trols. Finally, in column 5 we include both of our measures of ease of transit, which are highly correlated with each other. The magnitude of the coefficient on the transit access variable stays similar to earlier regressions, while the magnitude on the transit inconvenience variable falls.

4.3. Socioeconomic outcomes and the concentration of social networks

We next explore correlations between the geographic concentration of social networks and observable individual characteristics at the zip code level. Figure 6 shows zip code-level binned scatter plots of the share of friends living within 10 miles against socioeconomic measures in the New York CSA, controlling for the population within 10 miles; similar patterns arise when we measure the concentration of social networks at other distances or as the share of friends within a certain number of people. Panels A and B illustrate that zip codes with more widely dispersed social networks generally have higher incomes and education levels. Panel C shows that network concentration is higher in zip codes with a higher rate of teen birth, as measured by the share of individuals who were born between 1978 and 1983 that grew up in this zip code and had children before age 20. Panel D shows that the annualized growth in jobs from 2004 to 2013 in zip codes is also negatively correlated with social network geographic concentration. Panels E and F show that zip codes with more geographically dispersed social networks have higher measures of entrepreneurial quality (a predictive measure of their growth potential), but do not have higher levels of startup formation.[10] While the relationships in Fig. 6 are not necessarily causal, the literature has proposed many causal mechanisms for the observed patterns: indeed, access to diverse information through broad social networks is central to many theories of innovation, social mobility, and economic growth (Jackson, 2014; Granovetter, 2005; Greve and Salaff, 2003).

Overall, the results in this section can be summarized as follows. First, there is substantial heterogeneity across zip codes in various measures of the geographic concentrations of their social networks. Second, much of this variation in social network concentration is explained, at least statistically, by variation in the ease of travel via public transit to the rest of NYC. Third, zip codes with more concentrated social networks generally perform worse on socioeconomic indicators such as income and education levels. While we do not have a research design that allows us to give causal interpretations to these relationships, our results are highly consistent with stories in which public transportation infrastructure contributes to more geographically dispersed networks, thereby contributing to agglomeration externalities.

[10] The entrepreneurship data are described in Andrews et al. (2019). Here we use data from the "policy model", which relies only on institutional features (such as the registration of patents or trademarks) to determine quality. The economic importance of startup quality, as opposed to simple startup quantity, is highlighted by Guzman and Stern (2016). Census tract-level estimates are crosswalked to ZCTA-level using population weighting.

5. International dimension of urban social networks

In addition to exploring the domestic social connectedness of the New York metro area, we next look at the international dimension of social networks of New York residents. Figure 7 shows the percentile rank of the probability that a user in a given zip code within NYC has of being connected on Facebook to a user in a given country. Panel A shows connections to Bangladesh. Those areas with a high degree of social connectedness to Bangladesh correspond to areas within NYC with large Bangladeshi populations (see NYU School of Medicine, NYU Center for the Study of Asian American Health, 2019). Notably, regions with strong connectivity to Bangladesh are almost entirely concentrated along the LIRR and the 4-5-6 train service, consistent with a desire among recent immigrant communities to live in areas with easy travel to existing ethnic enclaves. Panel B shows connections to Senegal and reveals a distribution of connections concentrated in Harlem and portions of Brooklyn north of Prospect Park. Both of the locations contain a substantial number of residents with Senegalese backgrounds (see Duthiers et al., 2013; All Peoples Initiative, 2009). There are similarities to the distribution of the social network of East Harlem zip code 10035 shown in Panel B of Fig. 1. This suggests that areas that have strong connections to the same foreign countries are also more likely to be connected with each other. Consistent with this interpretation, the two broad areas with strong connections to Senegal were also grouped together into a joint "connected community" in Fig. 3. Panels C and D of Fig. 7 present the social connectedness of NYC zip codes to two European countries. As shown in Panel C, connections to Russia are concentrated in south Brooklyn, particularly around Coney Island and Brighton Beach; these areas correspond to parts of the city that welcomed large numbers of Russian-speaking immigrants since the 1970s, with increasing numbers arriving after the breakup of the Soviet Union (Ortiz and Untapped Cities, 2014). The pattern of connections to Russia is mirrored by the distribution of connections to most other Eastern European and former Soviet countries. In comparison, the distribution of connections to Germany, which is shown in Panel D, is typical of most Western European countries. Connections to Germany are primarily concentrated in the Midtown and Downtown regions of Manhattan and in neighboring areas in Brooklyn.

Figure 8 shows the percentile rank of the probability that a user in a given zip code within the New York CSA has of being connected on Facebook to a user in a given country. Panel A shows connections to India, highlighting the large Indian community in New Jersey (Berger and New York Times, 2008; Batalova et al., 2015). Panel B shows connections to Cuba, highlighting the stretch of Cuban communities in New Jersey nicknamed "Havana on the Hudson" (ShareAmerica, 2015). Panel C shows connections to El Salvador, which are primarily concentrated on Long Island. Indeed, El Salvador is the only country with a consulate on Long Island, located in the town of Brentwood (de Relaciones Exteriores de El Salvador, 2019). Panel D shows connections to Portugal revealing several cities and towns referred to as "little Portugal": Newark, NJ, has the highest concentration of connections to Portugal (Levy and New York Times, 1995), and in New York there are two longstanding immigrant communities on Long Island, Mineola and Farmingville, that display high degrees of social connectedness to Portugal. Portugal also exhibits high levels of social connectedness to the wealthy northern suburbs, potentially related to vacation travel (Rosenblum and New York Times, 1989; Fishler and New York Times, 2001).

Overall, these findings highlight that the degree of social connectedness of different NYC or New York CSA zip codes is to a substantial degree determined by the presence of migrants from these countries in the respective zip codes.

6. Conclusion

We use de-identified and aggregated data from Facebook to better understand social networks in the New York metro area, both at the city level and at the CSA level. Even in an era of increasing reliance on communications technology, we find that physical distance remains an important determinant of social connections. Yet, we also provide evidence of the important role of transportation infrastructure in forming and maintaining urban social connections by showing that social networks are distributed along public transportation routes and that social connectedness between locations declines more in travel time than it does in physical distance. We then show that commuting flows are stronger between socially connected areas, suggesting that social networks play an important role in shaping economic interactions. We document substantial heterogeneity in the geographic concentration of social networks, and highlight that locations with better public transit access have less geographically concentrated social networks, even after controlling for demographic characteristics of the neighborhoods; for example, we find that areas with more geographically dispersed social networks have higher-quality entrepreneurial activities. We also document how similarity along socioeconomic characteristics and the legacy of past migration movements are important drivers of social connectedness of the New York metro area.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jue.2020.103264

References

All Peoples Initiative, 2009. Senegalese in the New York metro area.

Andrews, R.J., Fazio, C., Guzman, J., Liu, Y., Stern, S., 2019. The startup cartography project: Measuring and mapping entrepreneurial ecosystems. Working Paper.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., 2018. The economic effects of social networks: Evidence from the housing market. J. Polit. Econ. 126 (6), 2224–2276.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., Wong, A., 2018. Social connectedness: measurement, determinants, and effects. J. Econ. Perspect. 32 (3), 259–280.

Bailey, M., Dávila, E., Kuchler, T., Stroebel, J., 2019. House price beliefs and mortgage leverage choice. Review of Economic Studies 86 (6), 2403–2453. doi:10.1093/restud/rdy068.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R., Stroebel, J., 2020. International trade and social connectedness. Working Paper.

Bailey, M., Johnston, D., Kuchler, T., Stroebel, J., Wong, A., 2019. Peer effects in product adoption. Natl Bureau Econ. Res. Working Paper No. 25843.

Bailey, M., Kuchler, T., Russel, D., State, B., Stroebel, J., 2020. Social connectedness in europe. Working Paper.

Bairoch, P., 1991. Cities and Economic Development: From the Dawn of History to the Present. The University of Chicago Press.

Barwick, P.J., Liu, Y., Patacchini, E., Wu, Q., 2019. Information, mobile communication, and referral effects. Natl. Bureau Econ. Res. Working Paper No. 25873.

Batalova, J., Zong, J., New York Times, 2015. Indian Immigrants in the United States.

Baum-Snow, N., 2013. Urban transport expansions, employment decentralization, and the spatial scope of agglomeration economies. Working Paper.

Berger, J., New York Times, 2008. A Place Where Indians, Now New Jerseyans, Thrive.

Bramoulle, Y., Galeotti, A., Rogers, B., 2016. The Oxford Handbook of the Economics of Networks. Oxford University Press.

Brooks, L., Lutz, B., 2014. Vestiges of transit: Urban persistence at a micro scale. Working Paper.

Brueckner, J.K., Largey, A.G., 2008. Social interaction and urban sprawl. J. Urban Econ. 64, 18–34.

Büchel, K., von Ehrlich, M., 2016. Cities and the structure of social interactions: Evidence from mobile phone data. Working Paper.

Cairncross, F., 2001. The Death of Distance: How Communications Revolution Is Changing our Lives. Harvard Business Review Press.

Chetty, R., Friedman, J.N., Hendren, N., Jones, M.R., Porter, S.R., 2016. The opportunity atlas: Mapping the childhood roots of social mobility. Natl. Bureau Econ. Research. Working Paper No. 25147.

Chetty, R., Saez, E., Turner, N., Yagan, D., 2020. Income segregation and intergenerational mobility across colleges in the united states. Q. J. Econ. Forthcoming.

Davis, D. R., Dingel, J. I., Monras, J., Morales, E., 2017. How segregated is urban consumption?.

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., Madden, M., 2015. Social media update 2014. pew research center.

Duggan, M., Greenwood, S., Perrin, A., 2016. Social media update 2016. Pew Research Center.

Duranton, G., Puga, D., 2015. Handbook Region. Urban Econ. 5, 467–560.

Duthiers, V., Chen, A., CNN, 2013. Little Senegal in the Big Apple: Harlem's West African heart.

de Relaciones Exteriores de El Salvador, M., 2019. Consulado General de El Salvador en Long Island, New York.

Farazmand, A., 1999. Globalization and public administration. Publ. Admin. Rev. 59, 509–522.

Fishler, M. S., New York Times, 2001. Long Island Journal: Long Islanders, Yes, But Not Quite Home.
Fujita, M., Thisse, J.-F., 2013. Economics of Agglomeration: Cities, Industrial Location, and Globalization. Cambridge University Press.
Gaspar, J., Glaeser, E.L., 1998. Information technology and the future of cities. J. Urban Econ. 43, 136–156.
Gilder, G., 1995. Forbes ASAP.
Glaeser, E., 2005. Urban colossus: Why is new york America's largest city? Working Paper.
Glaeser, E., 2011. Triumph of the City. Pan.
Glaeser, E., Gottlieb, J., 2009. The wealth of cities: agglomeration economies and spatial equilibrium in the united states. J. Econ. Lit. 47 (4), 983–1028.
Glaeser, E., Kim, H., Luca, M., 2017. Nowcasting the local economy: Using yelp data to measure economic activity. Working Paper.
Glaeser, E., Shapiro, J., 2001. Is there a new urbanism? The growth of u.s. cities in the 1990s.
Glaeser, E., Steinberg, B.M., 2016. Transforming cities: does urbanization promote democratic change? Working Paper.
Glaeser, E.L., Kahn, M.E., 2004. Sprawl and urban growth. In: Handbook of regional and urban economics, 4. Elsevier, pp. 2481–2527.
Glaeser, E.L., Kallal, H.D., Scheinkman, J.A., Shleifer, A., 1992. Growth in cities. J. Polit. Econ. 100 (6), 1126–1152.
Goldenberg, J., Levy, M., 2009. Distance is not dead: Social interaction and geographical distance in the internet era. arXiv preprint.
Granovetter, M., 2005. The impact of social structure on economic outcomes. J. Econ. Perspect. 19 (1), 33–50.
Granovetter, M.S., 1977. The strength of weak ties. In: Social networks. Elsevier, pp. 347–367.
Green, N., 2002. On the move: Technology, mobility, and the mediation of social time and space. Inf. Soc. 18, 281–292.
Greve, A., Salaff, J.W., 2003. Social networks and entrepreneurship. Entrepreneurship Theory Pract. 28(1), 1–22.
Guzman, J., Stern, S., 2016. The state of american entrepreneurship: New estimates of the quality and quantity of entrepreneurship for 32 us states, 1988-2014. Natl. Bureau Econ. Res. Working Paper No. 22095.
Helsley, R.W., Strange, W.C., 2014. Coagglomeration, clusters, and the scale and composition of cities. J. Polit. Econ. 122 (5), 1064–1093.
Helsley, R.W., Zenou, Y., 2014. Social networks and interactions in cities. J. Econ. Theory 150, 426–466.
Herrera-Yague, C., Scneider, C., Couronne, T., Smoreda, Z., Benito, R., Zurfiria, P., Gonzalez, M., 2015. The anatomy of urban social networks and its implications in the searchability problem. Sci. Rep. 5.
Ioannides, Y.M., 2013. From neighborhoods to nations: the economics of social interactions. Princeton University Press.
Ioannides, Y.M., 2015. Neighborhoods to nations via social interactions. Econ. Modell. 48, 5–15.
Jackson, M.O., 2014. Networks in the understanding of economic behaviors. J. Econ. Perspect. 28 (4), 3–22.
Jacobs, J., 1969. The Economy of Cities. Vintage.
Kim, J.S., Patacchini, E., Picard, P.M., Zenou, Y., 2017. Urban interactions. Working Paper.
Kowald, M., van den Derg, P., Frei, A., Carrasco, J.-A., Arentze, T., Axhausen, K., Mok, D., Timmermans, H., Wellman, B., 2013. Distance patterns of personal networks in four countries: a comparative study. J. Transp. Geograph. 31, 236–248.
Kuchler, T., Peng, L., Stroebel, J., Li, Y., Zhou, D., 2020. Social Proximity to Capital: Implications for Investors and Firms. Technical Report. Working paper
Kuchler, T., Russel, D., Stroebel, J., 2020. The geographic spread of COVID-19 correlates with structure of social networks as measured by facebook. Working Paper. National Bureau of Economic Research.
Lazarsfeld, P., Merton, R.K., 1954. Friendship as a social process: a substantive and methodological analysis. In: Berger, M., Abel, T., Page, C.H. (Eds.), Freedom and Control in Modern Society. Van Nostrand, New York, pp. 18–66.
Levy, C. J., New York Times, 1995. A Portuguese Village in Newark.
Marmaros, D., Sacerdote, B., 2006. How do friendships form? Q. J. Econ. 79–119.
Mok, D., Wellman, B., Carrasco, J., 2010. Does distance matter in the age of the internet? Urban Stud. 47, 2747–2783.
Mossay, P., Picard, P., 2011. On spatial equilibria in a social interaction model. J. Econ. Theory 146, 2455–2477.
MTA, 2016a. Annual bus ridership by route.
MTA, 2016b. Annual subway ridership.
NYU School of Medicine, NYU Center for the Study of Asian American Health, 2019. The Bangladeshi Community in the United States and New York City.
Ortiz, B., Untapped Cities, 2014. NYC's Micro Neighborhoods: Little Odessa in Brighton Beach, Brooklyn.
Perlman, E.R., 2016. Dense enough to be brilliant: patents, urbanization, and transportation in nineteenth century america. Working Paper.
Picard, P.M., Zenou, Y., 2018. Urban spatial structure, employment and social ties. J. Urban Econ. 104, 77–93.
Rosenblum, K., New York Times, 1989. Portuguese Voyagers Reach New Shores.
Sato, Y., Zenou, Y., 2015. How urbanization affects employment and social interactions. Eur. Econ. Rev. 75, 131–155.
Scellato, S., Mascolo, C., Musolesi, M., Latora, V., 2010. Distance matters: geo-social metrics for online social networks. Proc. WOSN'10.
Schläpfer, M., Bettencourt, L.M.A., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G.B., Ratti, C., 2014. The scaling of human interactions with city size. J. R. Soc. Interf. 11 (98).
Schmutte, I.M., 2015. Job referral networks and the determination of earnings in local labor markets. J. Labor Econ. 33, 1–32.
ShareAmerica, 2015. Havana on the Hudson: How Cuban Americans remade Union City, New Jersey.
Shue, K., 2013. Executive networks and firm policies: evidence from the random assignment of mba peers. Rev. Financ. Stud. 26(6), 1401–1442.
Takhteyev, Y., Gruzd, A., Wellman, B., 2012. Geography of twitter networks. Soc. Netw..
Verbrugge, L.M., 1983. A research note on adult friendship contact: a dyadic perspective. Soc. Forces 62 (1), 78–83.
Zimmerman, S.D., 2019. Elite colleges and upward mobility to top jobs and top incomes. Am. Econ. Rev. 109, 1–47.
Zipf, G.K., 1949. Human behavior and the principle of least effort.. addison-wesley press.