The Local Structures of Human Mobility in Chicago

JAMES SAXON
Harris School of Public Policy and Center for Spatial Data Science, University of Chicago

ABSTRACT. A large literature establishes the role of mobility in the maintenance of neighborhood social structures. Jane Jacobs famously argued that social capital is maintained through “cross-use of space,” and James Coleman formalized it as the “closure” of human interactions. Since many of these interactions entail human movement, neighborhoods with higher social capital should be distinguishable by more cohesive mobility networks. I observe the mobility of Chicago residents through a large dataset of smartphone users. I construct a neighborhood-level mobility network for the city and characterize neighborhoods according to their local graph structure. Neighborhoods that are well integrated with their surroundings have higher income and educational attainment. Consistent with social capital theory and routine activity theory in criminology, higher local network integration independently predicts lower levels of violent and property crime. The methodologies presented provide a meaningful, replicable, and inexpensive approach to the structural measurement of neighborhood networks and social structure.

1. Introduction

Why measure movement? Social ties in neighborhoods require and enforce routines and behaviors in physical space. From a theoretical perspective, one might start from Coleman’s seminal work on social capital, which he encoded as the ‘closure’ of social ties – in network terms, the clustering coefficient. (1988) Modern telecommunications notwithstanding, maintenance of formal and informal ties require individuals to traverse space. The clustering coefficient of mobility networks should then reflect neighborhood social capital. Forrest and Kearns (2001) develop this argument in the context of neighborhoods. They stress that neighborhoods are not simply territorial, but are instead a “series of overlapping social networks,” and suggest that differences among neighborhoods might be analyzed as differences in social networks. They write that these residence-based networks govern everyday routines that are “arguably the basic building-blocks of social cohesion.” This argument jibes with Small’s (2009) work on the organizational embeddedness of social capital in the sense that while individuals’ social ties may be embedded in organizations, those organizations are themselves often embedded in space.

In the half-century since its publication, Jane Jacobs’ Death and Life of Great American Cities (1961) has had a deep and pervasive impact on urban planning, sociology, and economics. The ideas of mixed use development, neighborhood social capital, and eyes on the street have all accumulated rigorous theoretical expression and broad empirical support. But it is only with modern data sources that the “intricate ballet” she describes of human movement across neighborhoods has become measurable in practice at large scale and fine granularity. This ballet is an imprint of the social and economic ties of a neighborhood. Neighborhood social ties both drive and depend on movement.

This interplay and tension between physical space and social relationships lies at the core of urban sociology. Human interactions between spatial neighbors are at once a normative goal for the fabric of healthy neighborhoods per Jacobs, and both the focus and central tenet of the ecological tradition of neighborhood work, advanced originally by Park, Burgess, and McKenzie. (Park and Burgess, 1925) Likewise for geographic analysis, Tobler’s “first law” (1970) articulates a primacy of relationships between “near things,” that has served as both a credo and a methodological foundation for the field. Yet despite the central role of physical space in social interactions, there has been scant measurement of levels of interactions across urban spaces: data have not previously allowed it.

In this paper, I leverage a large dataset of smartphone users to measure quotidian movement between Chicago neighborhoods. Drawing from both social theory and geographic analysis, I construct new variables to characterize the local mobility behaviors of neighborhoods. I show that these constructions yield “comprehensible” results: strong, expected correlations with established neighborhood observables. I then apply these new variables to the prediction of property and violent crime. Consistent with social disorganization and routine activity theories of criminality, they offer independent explanatory power, beyond traditional controls.

Still, criminality is of only incidental interest for this project. Rather, the aim is to develop inexpensive, extensible methods for characterizing the behavioral cohesion of neighborhoods and to demonstrate that, consistent with social theory, these properties are linked with other social outcomes. The fundamental, personal experience of physical movement through one’s own space varies predictably with wealth and education. This paper thus suggests that the structure of local movement participates in the constellation of factors that define the experience of urban poverty and privilege.
On the other hand, this perspective of dyadic interpersonal ties is neither necessary nor sufficient. Forrest and Kearns also describe the quieter processes of “repair work” and “normalization” that follow implicitly as fellow urbanites attend to adjacent routines. Anderson (2011) develops this position in his work on city-wide “cosmopolitan canopies”: urban spaces where people from diverse backgrounds share positive experiences and observe others at ease in a pleasant habitat. In a sense, successfully-shared spaces provide a “ballast” against prejudices and inevitable social slights that would otherwise shear the social fabric. Do denizens’ neighborhoods serve this role at a local level? In order to do so, the neighborhood spaces must be shared. Are they? In short, insofar as movement reflects social structures, it offers a behavioral vantage point for observing those structures.

Of course, mobility does not directly map to social interaction. Participation in economic, social, and civic organizations may all require movement while entailing different social behaviors. Putnam’s (1995) fabled bowling leagues illustrate that a single physical location can have ambiguous social content. One can bowl alone or with friends, and one can visit a park to walk a dog or to partake in a game of chess. On their own, physical locations betray only the aggregate forces that motivate individuals across space. Nevertheless, even devoid of explicitly social interactions, physical trajectories reflect the environments that humans experience. A solitary, aimless stroll expresses confidence in the security of an environment or appreciation of its aesthetic flavor.

Concern over a non-correspondence between spatial and social life is further moderated by empirical work that attests to the strong social content of spatial behaviors. For example, cell phone call data records have been used repeatedly to show that users who interact socially (exchange calls) are more likely to frequent similar locations (use the same cell towers). (Bagrow and Lin, 2012; Toole et al., 2015) These analyses have focused on users instead of neighborhoods, and have not assessed the social meaning of variation across neighborhoods in spatial integration. To that end, Browning and colleagues have, in a series of papers, found that Los Angeles neighborhoods where residents shared more locations had higher reported trust and collective efficacy. (Browning et al., 2017a,b) Using simulated mobility data for Columbus, they characterize the closure of the “econetwork” of activity spaces: the degree to which households share multiple spaces, conditional on sharing one. They show that neighborhoods with better closure have lower crime rates. (Browning and Soller, 2014; Browning et al., 2017c) The analysis that follows mirrors and reproduces this result using measured behaviors in Chicago.

The project also engages the geographic literature on “activity spaces.” This concept was first articulated in the early 1970s by Horton and Reynolds (1970; 1971) as the set of locations with which individuals come into “direct contact as the result of day-to-day activities.” At about the same time, Hägerstrand (1970) urged regional scientists to develop models for human trajectories in physical space. He advocated most vigorously a “time geographic” approach that privileged the time constraints incumbent on individuals in their daily lives. Still, like Horton and Reynolds, Hägerstrand also considered the impact of physical and social barriers. In the past half-century, the study of used activity spaces has thus intermingled with the constraints described by Hägerstrand, often to the advantage of temporal mechanisms at the expense of social ones. In particular, the dominant operationalizations of activity spaces have been as standard-deviational ellipses, minimum convex polygons (convex hulls), or buffered paths. (Dijkstra, 1999a,b; Järv et al., 2014; Jones and Pebley, 2014; Kwan and Hong, 1998; Manaugh and El-Geneidy, 2012; Patterson and Farber, 2015; Schönfelder and Axhausen, 2003) Each of these strategies treats space as socially “flat.” They assume that points between or physically adjacent to used locations must be accessible to an individual since, temporally, they are.

These techniques can be justified by appeal to Tobler’s Law, that “near things are more related than distant things,” but doing so ignores or assumes away much of the rationale for studying activity spaces in the first place. What is the social subtext of locations that are consistently bypassed or avoided? In their initial salvo, Horton and Reynolds suggested testing the “useful hypothesis” that action spaces are “to a large extent, shared by groups of people in close geographical propinquity.” Similarly, Hägerstrand scoffed at anyone who might believe that an individual’s “path could be a true time-space random walk.” While his work emphasized the temporal constraints, he recognized from the get-go the impact of those imposed by society; he advised in his conclusion that individuals consider space not simply as distance, but instead organize it in “sharply bounded territories.” Are residents, in fact, cut off from adjacent neighborhoods? Do their trajectories differ from random walks? What are the social constraints on movement, and what is the impact on residents whose environments are closely bounded? These questions parallel those of urban sociology. In effect, they probe the social consequences of contraventions of (a purely spatial application of) Tobler’s Law: what happens when nearer spaces are less related?

The measurement for the present work is based on a large, month-long sample of smartphone GPS locations in the Chicago region. For each user, I construct a vector of visits to neighborhoods and impute residence based on night-time location. I aggregate over residents of Census tracts to construct a network dataset between residences and visited locations in the city. This “neighborhood network” builds on the approach advocated by Sampson.
and Graif in the context of the Project on Human Development in Chicago Neighborhoods (PHDCN) and the associated Key Informant study. (Graif et al., 2014; Sampson, 2012) Neighborhoods are not just bounds within which to aggregate a more fundamental population; they are inter-related observations in and of themselves.

Using this network, I characterize neighborhoods in two ways, one inspired from sociology and the other from geography. The clustering coefficient is motivated directly from Coleman’s treatment of social closure and capital. It encodes whether individuals have destinations in common with the residents of the neighborhoods that they visit. Alternatively, the “local out-degree” measures the time spent by residents in the vicinity of (but outside of) their homes. It is constructed by reference to \( k \) nearest neighbor weights, to quantify how “related” nearby observations really are. In addition, I evaluate neighborhoods’ “ambient” populations: the numbers of devices regularly present within their bounds. This may be contrasted with the residential or work-time populations measured by the Census. I find that these features are strongly correlated with socioeconomic status.

But this begs the question: do these new variables reveal anything new? Or are they simply rococo representations of wealth and education, identifying walkable neighborhoods affordable only to elites?

Implications for crime. To investigate this question, I test these variables in simple models of criminality, with standard controls from theories of social disorganization and routine activities. Social disorganization theory weighs a community’s criminal tendencies against its capacity for effective social control. In the routine activities framework, crime occurs when a motivated offender coincides in space and time with a target without an effective guardian present. (Cohen and Felson, 1979) This perspective underscores the need for an accurate accounting of the population “at risk” for criminal behavior or victimization in a location, as first suggested by Boggs in 1965. For example, the residential population is an inappropriate normalization for property crime rates in the core business district, where total human activity (targets) far exceeds that of the residents.

The two theories hence privilege different controls, but make the consistent predictions that enhanced social control and guardianship depress crime. The preceding subsection describes the role of mobility in the maintenance of social cohesion and as a necessary, physical imprint of social closure. According to the social disorganization view, local mobility patterns should then correlate with lower crime through mechanisms of social cohesion and control. In the routine activities perspective, one would argue analogously that locals might be better attuned to their surroundings and more motivated to protect them; they might be better guardians. This behavior is quantified by the characteristics described above and derived more formally in what follows.

The practical hypothesis is thus that – with full controls – neighborhoods with greater local activity (out-degree) or that share more location activity with their neighbors (clustering) should see lower crime. This is, in fact, what the data show.

2. Data and Initial Processing

This project aims first to meaningfully characterize the mobility behaviors of urban neighborhoods and then to present the relationship between these behaviors and other social observables. This entails two classes of data: (1) a record of locations from individuals’ cell phones and (2) a raft of common socioeconomic controls.

Cell Phone Locations. The primary dataset consists of over 600 million GPS locations from smart phone users in Chicago and its western suburbs. The data were recorded in May 2017 by active applications on users’ phones, aggregated by LiveRamp and provided by Carto. These and other similar sources record consistent data for the entire United States, over years. This means that the strategies developed are in practice repeatable across time and space. The project thus responds to persistent calls for new methods for “big data” in the social sciences. (Arribas-Bel and Tranos, 2017; Bettencourt, 2014; Glaeser et al., 2018; Kitchin, 2016; Lazer et al., 2009)

Location data like those used here are increasingly available to researchers and support a burgeoning ecosystem of applications. They have been used in measurements of the consistency between social and mobility networks (Toole et al., 2015) and for determining the scaling properties of these networks (Schläffer et al., 2014). They have been applied to derive behavioral partitions of cities or countries (Calabrese et al., 2011; Poorthuis, 2017; Ratti et al., 2010). They are central to the development of increasingly accurate and affordable models of aggregate mobility. (Jiang et al., 2016; Song et al., 2010a,b) Criminologists have long called for better measurements of the populations “at risk” for criminal behavior and victimization (Boggs, 1965; Cohen and Felson, 1979), and big data sources allow these measurements at ever finer spatial and temporal granularity. (Andresen, 2006; Malleson and Andresen, 2016; Ristea et al., 2018; Song et al., 2018) This is their first application, however, to the measurement of the structure of local mobility and the characterization of networks of neighborhoods.

Substantial processing is required to construct this network. Some applications have access only to the user’s approximate location; I discard these imprecise data. Using OpenStreetMap data (OpenStreetMap contributors, 2018), I also suppress data where individuals
are in transit: points within 10 meters of motorways, trunk, primary or secondary roads, as well as railways or subways. After data cleaning, over 268 million user locations remain.

I associate these points to 2015 Census tracts with a simple point-in-polygon merge. For each user (device) in the dataset, I define the “home” Census tract as the modal location between midnight and 6am. This location can be defined for 279 thousand users, amounting to a roughly 3% sample of the 10 million-person data area. These users collectively account for 323 million points; the median user records 424 locations over the month.

Unlike most surveys, the location data is obviously a convenience sample. Nevertheless, the data appear to be fairly representative. As of the study period in 2017, over three quarters of Americans had smartphones. (Pew Research Center, 2018) In Appendix A, I compare distributions of Census tracts, weighted by either the “device” population or the official estimates from the American Community Survey (ACS). The device-weighted tracts are slightly wealthier, whiter, and more educated than the ACS population; this may also hold within tracts, but it is not measurable. The greater weakness of the nocturnal assignment method is that places with large tourist populations or 24-hour operations are over-represented, presumably with workers mis-assigned as residents. I drop “residents” of airports as well as the two devices assigned to the now-vacant site of Robert Taylor Homes.

Socioeconomic Covariates. I use four data sources from the US Census and the City of Chicago to evaluate the relationship between mobility and more-traditional socioeconomic observables. Most of these observables – education, income, racial composition, etc. – are drawn from the 5-year estimates of the Census’s 2017 American Community Survey (ACS).

For the routine activities model of criminality, I assemble three population counts or “risk sets.” The first of these, residential population, is also from the ACS. The second population, workers, comes from the Census’s LEHD Origin Destination Employment Statistics (LODES). The LODES data are derived from administrative records of unemployment insurance for wage and salary jobs. They cover approximately 95 percent of these jobs. (Graham et al., 2014) Last, the “ambient” population is built from the cell phone data as already mentioned. It represents the total population density of a space. It includes people in all of their functions: work and home, but also leisure and transit for instance. Its derivation is presented below.

I draw crime counts and zoning data from the City of Chicago. The geocoded crime data are for the 5-year period from 2013 to 2017. This time window significantly exceeds the single-month observation from the cell phone data, but allows for logged violent and property crimes at the Census tract level. Violent and property offenses are categorized by FBI Code following the definitions of the Chicago Police Department. From the zoning shapefile, I derive a single control variable: the fraction of each Census tract in the city that is zoned commercial. I define commercial zoning loosely, as anything that is neither residential nor parkland. This includes most downtown spaces, planned developments like airports, ballparks, and universities, as well as traditional business, commerce, manufacturing, and transportation designations.

3. Methods

Characterizing Neighborhoods. Using the cell phone location data, I characterize the local structure of mobility between Chicago neighborhoods. To do so, I first construct a network of neighborhoods: an interaction matrix representing the (normalized) rates at which residents of each neighborhood (Census tract) visit every other neighborhood in the city. For a user device $u$, I call the fraction of locations generated in Census tract $\ell$, $A_u^\ell$. For home neighborhood $h$, the set of devices based there is $R_h$, and the number of such devices is $n_h$. The neighborhood-averaged mobility of users based at location $h$ is thus $\bar{A}_h \ell = \sum_{u \in R_h} A_u^\ell / n_h$. The matrix $\bar{A}_h \ell$ represents a network of neighborhoods in the city. Cells in the matrix (edges in the network) represent the average rates at which residents of $h$ visit location $\ell$.

I next construct properties of urban neighborhoods (nodes) and evaluate them on this network. Census tract populations in Chicago vary between 2,500 and 15,000 people, but are on average lower than in the suburbs. The intensive character of neighborhoods does not depend on the Census Bureau’s decisions to split or merge Census tracts. The measures proposed below are therefore designed to be independent of tract population; they are “node-split invariant.” (Heitzig et al., 2012)

Clustering Coefficient. The clustering coefficient has long been recognized as a critical property of social networks. Informally, it measures whether a node’s neighbors are also neighbors with each other. More formally, it quantifies the number of closed triangles out of all triplets on the graph (three nodes connected by at least two edges). This definition suffices for simple graphs but is inadequate for graphs with edge and node weights (interactions and populations). The definition has been extended to graphs with weighted (Saramäki et al., 2007) and directed edges (Fagiolo, 2007), but not to graphs with weighted nodes as in the present case.

For any three nodes, there are eight directed triangles. I focus on the two triangles with two edges outgoing from the reference node, so the question is whether the residents of a node’s neighbors themselves

\begin{itemize}
  \item The triangles are: $\Delta \Delta \Delta \Delta \Delta \Delta \Delta$.
\end{itemize}
interact. If I visit locations \(A\) and \(B\), do residents of \(A\) visit \(B\)? The interactions \(A_{ht}\) are an intensive characteristic of \(h\); they are averaged over the population of the node. But person-to-person interactions are diluted by the population at the destination. Dividing the interaction from one cell to its neighbor by the neighbor’s population, I define weights as \(w_{ij} = \hat{A}_{ij}/n_j\). This is effectively “doubly-normalized” over the number of potential pairs of individuals over whom it is spread. It may be thought of as the likelihood of a person in \(i\) interacting with a specific person in \(j\). A larger tract is likely to attract more visitors, but is less affected by each individual visit. A nearly node-split invariant measure may then be constructed as

\[
c_h = \sum_{i \neq h} \sum_{j \neq h} \hat{A}_{hi} \hat{A}_{hj} w_{ij} / \sum_{i \neq h} \sum_{j \neq h} \hat{A}_{hi} \hat{A}_{hj},
\]

though it remains in principle sensitive to the population of the home tract. (See Appendix B.) The clustering coefficient mirrors the “econetwork intensity” of Browning et al. (2014; 2017a; 2017b; 2017c) in the sense that it quantifies the closure of movement behaviors. Their measure differs in its construction from a bigraph of households and (simulated) locations. In other words, actors and destinations are treated as distinct categories. Browning et al aggregate the closure over households in a neighborhood, whereas I aggregate the movement of devices in the neighborhood.

**Local Out-Degree.** The “local out-degree” of a neighborhood \(h\) is the sum of its residents’ interaction fractions that take place within its immediate vicinity. People interact with people, not Census tracts, so for a node-split invariant measure the vicinity must be defined with constant population. I set this population to \(n\) people. I call the cumulative population of the nearest \(k\) tracts \(N_k\) and the population of the next nearest tract \(n_{k+1}\). I define \(k\) as the smallest value for which \(N_k + n_{k+1} > N\), and call the corresponding set of \(k\) tracts the vicinity of \(h\), \(\mathcal{V}_h\). I take the notational liberty of indexing \(\hat{A}_{ht}\) by distances from \(h\), so that \(\hat{A}_{h,k+1}\) is the interaction of \(k+1\)’th neighbor. The local out-degree can then be written as

\[
\text{out-deg}_{\text{local}}(h) = \left[ \left( \sum_{\ell \in \mathcal{V}_h} \hat{A}_{h\ell} \right) + \hat{A}_{h,k+1} \frac{N - N_k}{n_{k+1}} \right] \times \left( 1 - \hat{A}_{hh} \right)^{-1}.
\]

This represents the sum of the \(k\) interaction fractions in \(\mathcal{V}_h\), corrected with the interaction with tract \(k + 1\) scaled by the remaining population difference \((N - N_k)/n_{k+1}\), and normalized to the total out-of-home activity, \(1 - \hat{A}_{hh}\).

Because individuals do tend to spend time in the vicinity of their own residence, the local out-degree is correlated with the clustering coefficient. In this sense, it codifies a similar characteristic of neighborhoods in a simpler way and offers a check of the robustness of results. But the local degree differs because it explicitly incorporates the spatial distribution of interactions, which the clustering coefficient does not do.

**Ambient Population.** Neighborhoods’ populations vary by hour and day; at a given moment, the population is neither the count of residents nor workers. Averaging over the observation window, I estimate the actual population present – the ambient population – by summing over all users the fraction of time spent (by user) in each tract. I normalize this population by the number of users \(U\) and tracts \(T\):

\[
\text{ambient } (\ell) = \frac{T}{U} \sum_u A_{\ell u}^U.
\]

This is obviously not node-split invariant and it is not meant to be: it is an extensive measure of the size and usage of the tract. Its logarithm is therefore included as a covariate in the logged crime model of the following section, to establish the “base rate.”

**Modeling Criminality.** I use a spatial error model of the form

\[
y = \beta x + \lambda W u + \epsilon,
\]

where the first term on the right-hand side represents standard covariates, and the second two the autoregressive and idiosyncratic parts of the error. (Anselin, 1988)

The weights matrix \(W\) is constructed as \(k = 8\) nearest neighbors. The outcome \(y\) is the logarithm of the crime count. The SAR error model is chosen over the spatial lag model based on the results of the LM test, and evaluated through maximum likelihood estimation using pySAL. (Anselin and Rey, 2014)

Criminologists and sociologists commonly convert crime counts to rates by dividing by residential populations, despite an awareness that alternative risk sets might be more appropriate. (Boggs, 1965; Cohen and Felson, 1979) As viable alternative normalizations have become available, Andresen (2006) and others have shown that this choice of denominator is consequential in practice. I reproduce this effect with a preliminary analysis, by regressing logged crime counts on “standard” covariates from social disorganization theory, with different population normalizations available. The three normalizations are the populations of residents and workers from the Census, and the ambient population described above. These fits yield unstable estimates of social disorganization parameters, in particular social disadvantage. This motivates the use of a “hybrid” approach for the main results, with both social disorganization and routine activities controls.

The second set of regressions confronts the main question: it tests the relationship between crime and clustering or local out-degree. These regressions include all three population controls as well as the logarithm of the tract area. In conjunction with the logged populations,
the log of the tract area implicitly includes population density. From the routine activities perspective, this affects the likelihood of interactions between offender and target, though one could also argue that there is safety (potential guardianship) in numbers.

For both sets of models, “social disorganization” covariates are derived from the ACS data. To facilitate comparisons and freeze parameters in the model, the variables are constructed to align with Browning et al. (2017c) and Morenoff et al. (2001) to the extent practicable. Social disadvantage is calculated as the first component from a principal component analysis combining employment, education, and income levels, along with fraction of workers in managerial positions, and the fraction of households headed by women. Residential instability similarly combines the share of residents who were living in the same house one year ago, with the shares of units occupied by renters and vacant. Racial heterogeneity is the sum of the squares of the shares of each race or ethnicity. (See Appendix D for details.)

4. Results

Figure 1 displays choropleths of the constructed variables: (a) the logarithm of the clustering (scaled by a factor of $10^8$, for legibility), (b) the local out-degree, and finally (c) the logarithm of the normalized ambient population. Each of these shows significant variation across the city. Higher values in (a) and (b) denote places with higher integration of movement behaviors. The local out-degree (b) can be interpreted as the strength of the relation to physical neighbors. It shows heterogeneity in the degree of conformity to Tobler’s Law, that nearer places are more related than distant places. In the City of Chicago, the intensity of relations with neighbors depends on where you are. The ambient population (c) highlights the active regions of the city.

Figure 2 presents the relationship of the constructed mobility variables with logged crime rates and adult bachelor’s degree attainment, for Census tracts in the city. The crime rates combine violent and property crime, and are normalized by residential population. The data are divided into three bins of roughly equal populations, by median annual household income: tracts with income less than $40k, between $40k and $60k, and above $60k. There is a strong, non-linear relationship between the variables. The population-weighted correlation between the bachelor’s attainment of the population aged 25 years and older, and the log clustering is 84% and that to the local out-degree is 45%. In the lower-income bin, crime rates are significantly higher at moderate clustering and out-degree, but converge towards the trends for higher-income tracts at high clustering or out-degree. The strong relationships suggest that the constructed variables are not just noise. The question then becomes whether the mobility data provide any independent explanatory power for predicting other social outcomes.

The results for the two models of crime are shown in Tables 1 and 2. The first three columns of Table 1 present regressions with three different, individual population normalizations. As might be expected, parameter coefficients and qualitative conclusions are not robust to these changes in the controls available for the risk set. In particular, the table shows significant changes in estimates of the coefficients for commercial zoning and social disadvantage. The changes in the commercial zoning variables are to be expected since they provide a crude proxy for the “missing” populations, which change from column to column. But the changes in the social disadvantage estimates are, while understandable, unexpected. A failure to account for all populations results in misleading estimates. In short, the ambient (or work-time) population is a straightforward and necessary control in establishing relevant risk set for crime rates.

Table 2 shows results for the regression of violent and property crime in Chicago on the population covariates and clustering coefficient. The first two columns present results for property crime, first without controls for social disorganization and then including them. The third column repeats the second, replacing clustering with local out-degree. The final column replaces violent with property crime. These regressions show that with full controls, property and violent crime are consistently and significantly lower, in neighborhoods with higher clustering or local out-degree.

The results are insensitive to the choice of weights: queen weights and $k = 6$ or $k = 10$ nearest neighbors yield equivalent results. Results are also equivalent with estimation through the Generalized Method of Moments (GMM). See Appendix C.

5. Discussion

Observers familiar with the geography of Chicago will find in Figures 1(a) and (b) a confirmation of their priors for the city. The wealthier North Side and the suburbs have “better” outcomes while the West and South Sides are more depressed. Major highways (I-90 and I-55), the Chicago Shipping and Sanitation Canal, and the North Branch of the Chicago River all show up as cleavages in the local out-degree: physical borders unsurprisingly affect mobility. Proceeding south along Lake Michigan, the notable outlier is the neighborhood of Hyde Park. This is no surprise: anchored by the University of Chicago, Hyde Park is economically and institutionally isolated from its neighbors. (Boyer, 2015; Jacobs, 1961; Levi, 1961; Sampson, 2012; Winling, 2017)

This point – that the socioeconomic pattern of the city is reflected in the daily trajectories of individuals in their communities – is worth emphasizing. The strong relationship with established outcomes lends credence
to the clustering coefficient and local out-degree as measures of local-scale neighborhood integration. Physical trajectories characterized through the clustering coefficient predict 84% of the variance among Chicago census tracts in adult educational attainment. It may be expected that privileged households will choose safe, walkable neighborhoods; the data show that residents of privileged and poorer neighborhoods engage their local environments in measurably different ways. In other words, denizens of wealthy neighborhoods exhibit more activity in the vicinity of their place of residence than their counterparts from poorer or less-educated neighborhoods. There is substantial variation in the “relatedness” of adjacent environments, which tracks with socioeconomic status. The purely spatial implication of Tobler’s Law – high relatedness with physical neighbors – is more true for the rich and educated than for the poor. This perspective is possible only with modern datasets.

Consistent with their theoretical motivations, the constructed variables also significantly and independently predict crime rates in the city. This conclusion holds with a rich array of controls for social organization, with three alternative definitions of the population or risk set, and for both violent and property crime. It is stable with respect to the estimation method and with alternative spatial weights matrices. The clustering and out-degree provide distinct alternatives on the measurement but yield consistent qualitative results.

These findings are consistent with work on “ecoworks” by Browning et al. (2017c) already discussed. They support both the social disorganization and routine activity theories of criminality. In the former, greater social capital and better organization improve social
This work was motivated by the idea that local mobility behaviors provide a window into levels of neighborhood social capital. This capital is expected to suppress crime, but the results presented show only that these behaviors correlate negatively with crime levels. Causality need not flow from social capital to crime rates, nor need social capital be even involved. A more prosaic, alternative explanation is that residents are aware of the relative safety of their environment, and respond by using or avoiding it. The mobility data would then capture this local knowledge. In that case, to the extent that physical mobility is required for the maintenance of social bonds, criminality would reduce social capital instead of the other way around.

The routine activities perspective emphasizes the need for an appropriate measure of the risk set. The ambient population as constructed is serviceable in this regard – either as the base risk or as a correction to the residential population. Removing it wholesale from the model, moreover, results in a substantial drop in the pseudo-\( R^2 \); a measure of the “ambient” (or work) population improves the accuracy of crime prediction (see Table 1). This is true, even though the illustrative models include measures of the fraction of space zoned for commerce. Those variables respond predictably to the presence or absence of components of the effective population. Lacking a direct measure of local employment, commercial spaces stand in as a proxy for the working population. Conversely, absent a control for the residential population, the fraction commercial indicates the space remaining for residences. In the model, the coefficients for fraction of tracts’ area zoned commercial and its square are positive and significant with residential population alone, negative and significant for work population alone, and insignificant with the ambient population alone.

But the zoning is ultimately an unsatisfactory proxy. As already noted and observed by Andresen in particular (2006; 2016; 2018), incomplete controls for the risk set impact conclusions about the impact of social disorganization on crime. The coefficients of social disadvantage and residential instability change significantly with alternative normalizations. It is worth acknowledging that the model’s sensitivity to the risk normalization is stronger for property than violent crime; in the latter case, residential population is a better measure of the risk. Still, it is striking that the coefficients for social disadvantage on property crime are negative in both tables, sometimes significantly. This represents a tension with the social disorganization approach to criminality: one might have expected disadvantage to be associated with higher crime. Perhaps high disadvantage denotes fewer appealing targets, rather than a greater appeal for criminal behavior.

This section has elaborated the relationship between crime and the constructed mobility variables: with a
dle of controls, communities with more-cohesive mobility have lower crime. But the broader conclusion is different: the clustering coefficient, out-degree, and ambient population as measured in cell phone data have strong relationships with other sociological observables, but they are distinct concepts related to previously unobserved phenomena.

It is worth noting that these findings are all based on measured activities at the scale of a Census tract. Other scales are likely to reveal and veil different processes. Time spent in one’s neighborhood might evidence an attractive environment, while excessive time within one’s own residence signals a poor instead of a positive outcome: unemployment. Within a single neighborhood or Census tract, some places may foster social interactions (cafés or sidewalks, for instance) while others (offices or traffic) might not. Within establishments, individuals’ interactions will be colored by their role: patron to server, worker to supervisor, regular and tourist. Turning outwards, communities with strong local ties may be poorly-integrated in the broader city and its economy. Whyte (1943) famously described the rich street life but limited outward connections of Boston’s North End in the 1930s. Observing the same local character decades later, Jacobs (1961) praised the neighborhood’s healthy cross-use and social control. Understanding the relationship between spatial and social networks depends on continued refinement in the measured granularity of spatial behaviors and improved knowledge of their social content in each space.

Developing this understanding has clear relevance beyond spatial networks, social interactions, movement, or criminality. From a methodological perspective, the specification of interactions or spatial weights matrices has been a source of “some arbitrariness” for spatial econometric models. (Anselin, 2002) A small industry has developed statistical strategies for determining whether these interactions are needed, as well as which set of weights to use. (Anselin, 1984, 1986; Anselin et al., 1996; Kelejian, 2008; Kelejian and Piras, 2011; LeSage, 2015; Stakhovych and Bijmolt, 2009) But a better likelihood or more explained variance does not necessarily mean that the model is correct: if there are strong theoretical reasons to use a certain form of interaction, then that is the parameter that should be measured. Directly adopting the empirical interactions matrices might not be advisable, since these weights are apt to be endogenous to the situation. Nevertheless, they provide new insight on the choice of weights and might motivate unexpected alternatives. For example, Figure 1 shows that that the level of interactions with nearest neighbors is – broadly – lower in Chicago than its suburbs. This might motivate the use of separate regimes when modeling information diffusion for instance, in a metropolitan region. Simply “testing weights,” one might have only compared distance decay functions, contiguity rules, or numbers of nearest neighbors.

Social interactions are also of central importance to a number of active topics in urban economics, and measurement of these interactions is sorely needed. For example, a robust literature suggests that residential networks serve as an information pathway for job searches. (Bayer et al., 2008; Beaman and Magruder, 2012; Durlauf, 2004; Hellerstein et al., 2011, 2014; Ioannides and Loufy, 2004; Topa, 2001) Is the strength of residential social networks measurable? If so, then it is natural to ask whether neighborhoods with stronger ties see greater use of neighborhood networks for job search. Next, with some notable exceptions (Echenique and Fryer, 2007; Jones and Pebley, 2014; Wang et al., 2018), work on segregation in cities has focused on place of residence rather than actual interactions; a shift towards the full environment offers a richer paradigm. Finally, influential models of skill development in cities describe a “follower-the-leader” mechanism, where individuals learn from contact with more-skilled peers or teachers. (Glaeser, 1999; Glaeser and Maré, 2001; Lucas, 2004) Who in the city actually interacts with high-skills communities? Twenty years ago, Mankiw (2000) argued that “throughout the modern development of economics, empirical analysis of social interactions has lagged far behind theory.” Despite explosive growth in work on social and economic networks (see Jackson et al. 2017, for a review), this remains true today.

6. Conclusions

This project has used novel data sources to characterize the local structure of human mobility in Chicago neighborhoods. The variables defined – the clustering and local out-degree – are motivated from close antecedents in sociological theory and geographic analysis to encode neighborhood social capital and spatial weights. As measured in the data, these variables are strongly correlated with adult educational attainment, but also provide independent explanatory power for the prediction of crime levels. Outliers are comprehensible in the social and economic context of the city. In short: the variables as constructed have good face validity and represent new and distinct concepts. They quantify consequential variation in the local structure of Chicago neighborhoods, that has not before been observable.

This window on neighborhoods is possible thanks to new, large data sources. These data sources and the methods used to analyze them allow replication over time and space. Detailed work is needed to ground-truth the relationship between mobility and social interactions across socioeconomic classes and physical spaces. But these data ultimately hold promise for continuous and consistent measurement of empirical mobility behaviors and local networks at national scope.
References


DIST, M., “Action space as planning concept in spatial planning,” *Netherlands journal of housing and the built environment*, 14:2, 163-182 (June 1, 1999a). DOI: 10.1007/BF02496820.


MANAUGH, K. and EL-GENEIDY, A., “What makes
REFERENCES


Appendix A. Population Comparisons

GPS devices are assigned “home” locations based on their night-time location, as discussed in the text. The device to population ratio in the study area is shown Figure A.1. This ratio averages around 3% but exhibits significant variation. In particular, places with 24-hour workers – hospitals, airports, large production facilities, and transportation depots are over-counted. The reader can orient his or herself with Figure A.2, which shows major physical features of the city along with several of these sites. The Ford plant in the city is notably “red.” So too, is a large train depot just south of Midway Airport (itself excluded). There are a number of similar facilities along I-55, exiting the city.

GPS devices (cell phones) are used in place of people, and it is natural to ask whether this biases results. Figure A.3 compares several key demographic characteristics from the American Community Survey, for Census tracts.

Figure A.1. Device to population ratio for the cell trace dataset. The effective sampling rate is around 3%. The downtown core is a clear outlier, presumably due to tourists. The other outliers appear to arise from the use of a night-time location for imputing residence. This includes the regions around O’Hare and Midway airports (themselves excluded), major hospitals, the Ford assembly plant, and several multimodal transportation hubs. These are all spaces with night shifts.

Appendix B. Clustering Coefficient

We require a node split invariant clustering coefficient, for graphs with weighted edges and nodes. It is claimed that

$$c_h = \frac{\sum_{i \neq h} \sum_{j \neq h} \hat{A}_{hi} \hat{A}_{hj} w_{ij}}{\sum_{i \neq h} \sum_{j \neq h} \hat{A}_{hi} \hat{A}_{hj}}$$

satisfies this.

Splitting the destinations would divide the interactions $A_{ij}$, but this is compensated by the additional terms. Consider a node $k$ of population $n_k$ split into two nodes of populations $n_k'$ and $n_k''$. A node $i$ that interacted with $k$ at a level $A_{ik}$, would interact with the two new nodes at rates proportional to their shares of the original population: $A_{ik'} = A_{ik} n_{k'}/n_k$ and $A_{ik''} = A_{ik} n_{k''}/n_k$. Note that $A_{ik} = A_{ik'} + A_{ik''}$. Before splitting $k$, the clustering contains a term

$$\hat{A}_{ij} \hat{A}_{ik} w_{jk} = \hat{A}_{ij} \hat{A}_{ik'} \hat{A}_{jk}/n_k.$$

Afterwards, this is

$$\hat{A}_{ij} \left( \hat{A}_{ik'} \hat{A}_{jk'} \hat{A}_{ijk'} n_{k'}/n_k + \hat{A}_{ik''} \hat{A}_{jk''} n_{k''}/n_k \right).$$

Substituting $\hat{A}_{jk'} = \hat{A}_{jk} n_{k'}/n_k$ and $\hat{A}_{jk''} = \hat{A}_{jk} n_{k''}/n_k$ shows the two expressions to be equivalent:

$$\hat{A}_{ij} \left( \hat{A}_{ik'} \hat{A}_{jk} n_{k'}/n_k + \hat{A}_{ik''} \hat{A}_{jk} n_{k''}/n_k \right) = \hat{A}_{ij} \left( \hat{A}_{ik'} + \hat{A}_{ik''} \right) \hat{A}_{jk}/n_k = \hat{A}_{ij} \hat{A}_{ik} \hat{A}_{jk}/n_k.$$
Splitting an origin changes nothing for interactions outside the home: the “intensive” preferences are the same on either side of the split.

But the “home location” itself is special – it is excluded from the denominator and the numerator. The clustering and local out-degree are both constructed in this way, because the social implications of staying “closed-in” at home are very different from being outside. In fact, the fraction of time-at-home is higher in Chicago’s more depressed neighborhoods. However, though a large piece of a user’s home interactions are actually at the place of residence, some of them are in the immediate outside environment. Because the immediate outside location is the most-likely out-of-home space, the clustering coefficient – and the local out-degree – will be biased low for large home Census tracts where a higher fraction of immediately-local, non-residence locations are suppressed. To this extent, the node-split invariance is imperfect for the home location.

This weakness could be addressed by identifying home locations at finer granularity, but doing this begins to come up against the resolution of the data.

Appendix C. Alternative Estimation

The main results on mobility variables are completely insensitive to re-estimation with crime rates instead of counts (Table C.1) or with GMM estimation or alternative weights definitions (Table C.2).

Appendix D. Variable Definitions

To avoid tuning any results, the definitions of variables of social organization are defined to parallel the work Browning et al in Browning et al. (2017c). All input variables listed here are drawn from the American Community Survey 2017 5-year estimates. Disadvantage and residential instability are both the first principal component of the following normalized inputs:

- **Disadvantage**: fraction not working (unemployed or out of the labor force), log median household income, fraction of the adult population with a bachelor’s degree, fraction of workers in managerial positions, and fraction of households with a female head.

- **Residential instability**: fraction of residents who were living in the same house one year ago, fraction of units occupied by renters, and fraction of units vacant.

Each of these PCAs is computed for all tracts in Cook County, Illinois – including but not limited to Chicago. The sign of disadvantage is set to be anti-correlated with bachelor’s degree attainment. Residential instability is defined to be positively correlated with fraction of units occupied by renters.

**Fraction young** is slightly more expansive than Browning et al: I use the entire population from 15-24, which was available in the Census profile API. **Racial heterogeneity** is the sum of the squares of all fractional races: White, Black, Native American, Asian, Hawaiian/Pacific Islander (all non-Hispanic), or Hispanic. Similarly, the variable **Fraction Black** does not include Hispanics.

The “fraction commercial” is defined as spaces that were zoned neither for parks, nor for residential (single, twin, multiple, or downtown).

Crime data are from the City of Chicago. Offenses were categorized by FBI Code according to the classification of the Chicago Police Department.

- **Violent crime includes**: homicide of both first and second degree (code 1A), criminal sexual assault (2), robbery (3), aggravated assault (4A), and aggravated battery (4B).

- **Property crime includes**: burglary (5), larceny (6), motor vehicle theft (7), and arson (9).

### Table C.1. Regressions with logged crime rates per residential population, instead of counts yield consistent parameter estimates for clustering. This consistency is mathematical: because the denominator of the logged rate is included on the right-hand side, the fit simply adjusts to its removal from the outcome.

<table>
<thead>
<tr>
<th>Dependent</th>
<th>Violent</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.58</td>
<td>2.31***</td>
</tr>
<tr>
<td>Log Clustering</td>
<td>-0.24***</td>
<td>-0.22***</td>
</tr>
<tr>
<td>Log Population</td>
<td>-0.60***</td>
<td>-0.57***</td>
</tr>
<tr>
<td>Density</td>
<td>-0.03*</td>
<td>0.03</td>
</tr>
<tr>
<td>Log Worker</td>
<td>-0.29***</td>
<td>-0.28***</td>
</tr>
<tr>
<td>Log Area</td>
<td>-0.43***</td>
<td>-0.42***</td>
</tr>
<tr>
<td>Disadvantage</td>
<td>0.08***</td>
<td>0.06**</td>
</tr>
<tr>
<td>Married</td>
<td>-0.04</td>
<td>-0.16</td>
</tr>
<tr>
<td>Households</td>
<td>0.12***</td>
<td>0.12***</td>
</tr>
<tr>
<td>Residential Instability</td>
<td>0.94***</td>
<td>0.91***</td>
</tr>
<tr>
<td>Immigrants</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Fraction Black</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Fraction Young</td>
<td>-0.38</td>
<td>-0.41*</td>
</tr>
<tr>
<td>Recent</td>
<td>-1.35*</td>
<td>-1.34*</td>
</tr>
<tr>
<td>Fraction Black</td>
<td>0.11***</td>
<td>0.11***</td>
</tr>
</tbody>
</table>

N: 777 777 777 777  
Pseudo-$R^2$: 0.74 0.76 0.62 0.65  
AIC: 675 654 340 310  
Weights: KNN8 KNN8 KNN8 KNN8  
Routine: ML Error ML Error ML Error ML Error  
Standard errors in parentheses; *p < 0.05, **p < 0.01, ***p < 0.001.
Table C.2. Estimates of are consistent with respect to choice of weights, and with GMM estimation.

<table>
<thead>
<tr>
<th>Dependent</th>
<th>Log Violent Crime per Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.00*** 1.21*** 0.97*** 1.93***</td>
</tr>
<tr>
<td></td>
<td>(0.81) (0.31) (0.32) (0.31)</td>
</tr>
<tr>
<td>Log Clustering</td>
<td>-0.20*** -0.23*** -0.22*** -0.22***</td>
</tr>
<tr>
<td></td>
<td>(0.05) (0.05) (0.05) (0.04)</td>
</tr>
<tr>
<td>Log Res.</td>
<td>0.43*** 0.42*** 0.44*** 0.37***</td>
</tr>
<tr>
<td></td>
<td>(0.04) (0.04) (0.04) (0.04)</td>
</tr>
<tr>
<td>Log Ambient</td>
<td>0.28*** 0.29*** 0.28*** 0.25***</td>
</tr>
<tr>
<td></td>
<td>(0.03) (0.04) (0.03) (0.04)</td>
</tr>
<tr>
<td>Log Work</td>
<td>0.03* 0.02 0.03* 0.01</td>
</tr>
<tr>
<td></td>
<td>(0.01) (0.01) (0.01) (0.01)</td>
</tr>
<tr>
<td>Log Area</td>
<td>0.13*** 0.12*** 0.11*** 0.18***</td>
</tr>
<tr>
<td></td>
<td>(0.05) (0.03) (0.03) (0.03)</td>
</tr>
<tr>
<td>Disadvantage</td>
<td>0.05* 0.06** 0.07*** 0.09***</td>
</tr>
<tr>
<td></td>
<td>(0.02) (0.02) (0.02) (0.02)</td>
</tr>
<tr>
<td>Married</td>
<td>-0.13 -0.23 -0.09 -0.29</td>
</tr>
<tr>
<td>Households</td>
<td>(0.19) (0.19) (0.19) (0.20)</td>
</tr>
<tr>
<td>Residential</td>
<td>0.12*** 0.12*** 0.13*** 0.12***</td>
</tr>
<tr>
<td></td>
<td>(0.02) (0.02) (0.02) (0.02)</td>
</tr>
<tr>
<td>Instability</td>
<td>-0.46* -0.48* -0.48* -0.16</td>
</tr>
<tr>
<td></td>
<td>(0.20) (0.20) (0.21) (0.21)</td>
</tr>
<tr>
<td>Fraction Young</td>
<td>-1.28* -1.44** -1.45** -1.55**</td>
</tr>
<tr>
<td></td>
<td>(0.52) (0.52) (0.54) (0.54)</td>
</tr>
<tr>
<td>Recent</td>
<td>0.81*** 0.90*** 0.92*** 0.92***</td>
</tr>
<tr>
<td>Immigrants</td>
<td>(0.12) (0.11) (0.11) (0.10)</td>
</tr>
<tr>
<td>Residential</td>
<td>0.01 -0.01 -0.04 0.11</td>
</tr>
<tr>
<td></td>
<td>(0.11) (0.11) (0.11) (0.11)</td>
</tr>
<tr>
<td>Log Fraction</td>
<td>-0.29*** -0.30*** -0.30*** -0.32***</td>
</tr>
<tr>
<td>Commercial</td>
<td>(0.07) (0.07) (0.07) (0.07)</td>
</tr>
<tr>
<td>Log Fraction</td>
<td>-0.09*** -0.09*** -0.09*** -0.11***</td>
</tr>
<tr>
<td>Commercial Sq.</td>
<td>(0.02) (0.02) (0.02) (0.02)</td>
</tr>
<tr>
<td>λ (Error)</td>
<td>0.13 0.14*** 0.09*** 0.22***</td>
</tr>
<tr>
<td></td>
<td>(0.00) (0.00) (0.00) (0.00)</td>
</tr>
</tbody>
</table>

N 777 777 777 777
Pseudo-\(R^2\) 0.66 0.69 0.69 0.71
AIC 664 667 991
Weights KNN8 KNN6 KNN10 Queen
Routine GMM Error ML Error ML Error ML Error

Standard errors in parentheses; *\(p < 0.05\), **\(p < 0.01\), ***\(p < 0.001\).