

Unfolding community homophily in U.S. metropolitans via human mobility

Xiao Huang ^{a,*}, Yuhui Zhao ^b, Siqin Wang ^c, Xiao Li ^d, Di Yang ^e, Yu Feng ^f, Yang Xu ^g, Liao Zhu ^h, Biyu Chen ^b

^a Department of Geosciences, University of Arkansas, Fayetteville, AR 72701, USA

^b State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

^c School of Earth and Environmental Sciences, The University of Queensland, Brisbane, QLD 4067, Australia

^d Texas A&M Transportation Institute, Bryan, TX 77807, USA

^e Wyoming Geographic Information Science Center, University of Wyoming, Laramie, WY 82071, USA

^f Institute of Cartography and Geoinformatics, Leibniz University Hannover, Hannover 30167, Germany

^g Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China

^h Department of Statistics and Data Science, Cornell University, Ithaca, New York 14853, USA

ARTICLE INFO

Keywords:

Homophily
Community entropy
Segregation
Mobile phone data
Human mobility

ABSTRACT

As described in the proverb “birds of a feather flock together”, the term *homophily* narrates the principle that stronger spatial interactions tend to be formed among locations with similar characteristics. Taking advantage of mobility networks derived from around 45 million mobile devices in the U.S. and targeting the top twenty most-populated U.S. Metropolitan Statistical Areas (MSAs), we extract human mobility structures by detecting communities formed by strong spatial links and unravel the *homophily* effect at the community level using information entropy that measures the chaoticness of societal settings within communities. The results suggest that the power-law still, to a large extent, governs the travel patterns in MSAs. However, communities featured by strong human interactions can sometimes transcend geographic proximity in modern metropolitans. The entropy varies across communities, and a community can exhibit variation of *homophily* levels when different socio-demographic settings are investigated. Our study proves the ubiquity of the *homophily* phenomenon in modern metropolitans and documents its variation from different social perspectives from a mobility-oriented setting. The conceptual and analytical knowledge, as well as the results of this study, are expected to facilitate better policymaking to promote social integration in metropolitan areas.

1. Introduction

As described in the proverb “birds of a feather flock together” (Ferguson, 2017), the term *homophily* (the counterpart of *heterophily*) depicts the tendency that individuals tend to establish strong bonds with similar others. From a spatial perspective, *homophily* also narrates the principle that stronger spatial interactions tend to be formed among locations with similar characteristics (e.g., demographic and socioeconomic status). As one of the most important regularities that govern human spatial and social interactions, the *homophily* effect has been a hot research topic in various fields and investigated by many scholars (e.g., McPherson et al., 2001; McCrea, 2009; Kwan, 1999; Limtanakool et al., 2006; Reardon & Firebaugh, 2002; Reardon & Bischoff, 2011).

The investigations of *homophily* in the early days were mostly survey-

driven, in reliance on surveys (e.g., interviews and questionnaires) as the major data source to obtain people’s travel patterns. Further, the collected travel behaviors were analyzed across residential zones, social classes, and demographic tiers (Curranini et al., 2009; Kwan, 1999; Limtanakool et al., 2006; Mollica et al., 2003). Although detailed background information of respondents can be collected from surveys, the issues of restricted spatiotemporal coverage and high labor/time cost largely limited the scopes of these studies. In today’s era of big data, the rapid development of location-aware technologies allows researchers to have access to various emerging mobility datasets that detail people’s whereabouts with fine spatiotemporal granularity. Datasets that include social media posts, mobile phone records, WIFI, and smart cards, to list a few, enable the investigation of the *homophily* effect from a data-driven perspective. Given privacy concerns, the

* Corresponding author.

E-mail addresses: xh010@uark.edu (X. Huang), zhaoyuhui@whu.edu.cn (Y. Zhao), s.wang6@uq.edu.au (S. Wang), x-li@tti.tamu.edu (X. Li), dyang1@uwyo.edu (D. Yang), yu.feng@ikg.uni-hannover.de (Y. Feng), [yang.ls.xu@polyu.edu.hk](mailto.yang.ls.xu@polyu.edu.hk) (Y. Xu), lz384@cornell.edu (L. Zhu), chen.biyu@whu.edu.cn (B. Chen).

demographic and socioeconomic backgrounds of users are less likely to be released or associated with their moving patterns. However, certain strategies can be employed to infer users' backgrounds by referring to the neighborhood profile of where they live, and such inference may base on certain well-established statistical units.

The investigation of the *homophily* effect among places relies on the definition of place connectivity, which is usually derived from the spatial closeness of places, e.g., the closeness of residential zones (Massey & Denton, 1987; Reardon & Firebaugh, 2002; Reardon & O'Sullivan, 2004), or people's travel patterns (Galiana et al., 2018; Shelton et al., 2015; Xu et al., 2018; Xu et al., 2021). The former reveals the static residential segregation featured by the uneven distribution of residents' demographic and socioeconomic status, while the latter reveals the dynamic strengths of spatial interactions between humans and places. The limitation of *homophily* investigation via spatial closeness is obvious, as it relies solely on fixed spatial settings that may not well-govern the actual human-space interactions. Regardless of various dictations of the "distance decay" effect in the spatial domain, geographic units that are spatially close do not necessarily present strong human interaction in our modern urban fabrics. On the other hand, existing efforts of *homophily* investigation using travel patterns also present shortcomings, as most studies tend to build connectivity networks using travel patterns distributed with great skewness in terms of travel distances and the number of trips. Such networks usually fail to reflect the skeleton of spatial interactions due to the fact that essential spatial structures can be influenced by long-tailed links.

Different from existing efforts, in this study, we extract human mobility structures by detecting communities formed by strong spatial links and unravel the *homophily* effect at the community level using information entropy that measures the chaoticness of local settings within communities. We take advantage of mobility networks derived from around 45 million mobile devices in the U.S. and select the top twenty most-populated U.S. Metropolitan Statistical Areas (MSAs) as our study areas. The objectives of this study are to

- 1) reveal the similarity and dissimilarity in mobility profiles of these MSAs and explore how detailedly documented travel patterns follow the "distance decay" effect, one of the fundamental laws in geography.
- 2) detect communities characterized by strong spatial interactions using a network community detection algorithm via aggregated mobility networks.
- 3) introduce the concept of "community entropy" that reflects the chaoticness of local settings among community members within a community formed by strong spatial ties.
- 4) perform community entropy calculation of selected sociodemographic variables as well as different aspects of social vulnerability, and further explore the distribution patterns of the calculated community entropy among these investigated MSAs.

2. Background

2.1. Definition of homophily and related studies

In a broad sense, *homophily* describes a principle that contacts between similar individuals or spatial entities occur at a higher rate than those between dissimilar individuals or spatial entities, leading to localized and targeted flows of cultural, behavioral, genetic, or material information (McPherson et al., 2001). The phenomenon of *homophily* has been noted in many domains and has gradually become one of the fundamental principles that drive the operation of social networks and human spatial interactions (McCrea, 2009).

From a spatial perspective, *homophily* can be explained by the hypothesis that great satisfaction tends to be obtained from interactions between people with a similar background. As one's social background is greatly associated with the underlying local settings characteristics of

where he/she resides (e.g., income level, racial composition, and community structure), the *homophily* principle also implies stronger social connections among locations with similar characteristics (Xu et al., 2021). In general, the investigations on *homophily* can be grouped into two categories, given the units with which investigations are conducted, 1) individual-level, i.e., the investigation that targets person connectivity and 2) place-level, i.e., the investigation that targets location connectivity.

Before the proliferation of information and communication technologies, travel surveys had been used as the most reliable data source to assess and compare human travel behaviors across social classes and demographic tiers. Notable efforts include Kwan (1999), who investigated the disparity of travel links due to gender differences in Columbus, Ohio, U.S., and Limtanakool et al. (2006), who explored the impact of the spatial configuration of land use and transport systems on travel patterns in the Netherlands. Despite that survey-based investigation can retrieve detailed demographic and socioeconomic information from respondents, the difficulties in data collection (e.g., labor-intensive and time-consuming) largely limited the scope of these studies, which were usually limited to a small sampling size of participants during a short period of time (Heine et al., 2021). The advent of location-aware techniques coupled with the emerging concepts of "Web 2.0" (Murugesan, 2007) and "Citizen as Sensors" (Catlin-Groves, 2012) led to the emergence of many new human mobility datasets which detail fine-grained whereabouts of a massive number of users. From these datasets, users' travel patterns, even detailed moving trajectories, can be well documented, thus enabling a better investigation of individual-level *homophily*. Taking advantage of over 75 million geotagged tweets (Twitter posts with locational information), Bora et al. (2014) modeled user-specific travel patterns and explored the racial segregation in three major U.S. cities that include New York, Los Angeles, and Chicago. Their findings revealed that a common trend for all races is to visit areas populated by similar race more often. Similarly, by mining users' travel patterns from a massive amount of geotagged tweets, Wang et al. (2018) expanded the research scope to fifty largest U.S. cities and provided evidence that the interactions across racial and social groups that ultimately contribute to societal integration are, unfortunately, not taking place.

Instead of investigating individual-level *homophily* via person connectivity, the investigation of place-level *homophily* relies on connectivity between physical locations, either summarized from human mobility patterns or based on established spatial settings of the neighborhood. The former is based on the dynamic spatial interaction among places. For example, studies that used call detail records (CDRs) often aggregated mobility patterns to the service areas of cell phone towers (often approximated via Voronoi cells) and further explored the *homophily* effect revealed from these service areas (Blumenstock & Fratamico, 2013; Galiana et al., 2018; Xu et al., 2018; Xu et al., 2021). Other efforts have been made to summarize place connectivity using social media posts and explore how such connectivity follows geographic and socioeconomic contexts. For example, Morales et al. (2019) obtained mobility networks by summarizing individual users' Twitter activity at a neighborhood level and compared the urban segregation situations in Istanbul, Turkey and five major U.S. cities. Similarly, Shelton et al. (2015) used geotagged tweets to explore social-spatial relations and the longstanding problems of socio-spatial inequality in Louisville, Kentucky, U.S. Besides mobility-derived connectivity, established spatial settings of neighborhoods can also reveal place-level *homophily*, as the spatial distribution of neighborhoods can present uneven distributions of sociodemographic backgrounds. Multiple approaches have been established to reveal place-level *homophily* by identifying residential segregation of race, income, and education (Reardon & Bischoff, 2011; Reardon & Firebaugh, 2002; Reardon & O'Sullivan, 2004). Compared to mobility-driven connectivity, however, spatial connectivity reveals residential segregation in a static manner, ignoring the actual interactions among spatial entities.

2.2. Community detection algorithms

The identification of community in this study belongs broadly to a common task in spatial analysis, geography, and network science, termed “regionalization” or “community detection”. The former emphasizes connection strengths among a region’s components, and the latter focuses on the homogeneity of attributes within units of a region (Mu et al., 2015). In this study, the community detection workflow coupled with the proposed entropy measurement represents a unified framework that addresses both spatial connections and the homogeneity of attributes. Mobility patterns, with origin and destination locations serving as vertices (i.e., nodes) and trips (usually characterized by the number of travels) between them serving as edges (i.e., links), assemble as a network. Communities in a network are defined as dense groups of vertices, which are tightly connected to each other inside the community and loosely connected to the rest of the vertices in the network (Khan & Niazi, 2017). A community structure plays an essential role in revealing the structure-function relations of networks. In recent decades, there have been many studies on community detection in the field of complex networks (Barabási, 2009; Zhong et al., 2014), which can be roughly divided into two groups: topological-based methods (e.g., modularity) and flow-based methods (e.g., Infomap). Topological-based methods are suitable for analyzing networks where edges represent pairwise relationships, while flow-based methods are suitable for analyzing networks where edges represent movement patterns between nodes (Rosvall & Bergstrom, 2008). Due to length limitations, their mathematical foundation and applications are not detailed here. For more information on algorithm summary and applications, please refer to review articles by Wang et al. (2015) and Karataş and Şahin (2018), respectively.

Among the above community detection algorithms, Infomap is one of the most widely used community detection methods. Infomap follows the information-theoretic concept proposed by Rosvall and Bergstrom (2008) and has been applied in various domains (Deauna et al., 2021; Hong & Yao, 2019). Infomap is based on the map equation framework (Rosvall et al., 2009), can be applied to the directed weighted graphs, and presents superiority over modularity methods (Bae et al., 2013; Farage et al., 2021). Notable efforts include Hong and Yao (2019), who employed Infomap to detect the hierarchical community structure of urban roads, and Lu et al. (2018), who used the Infomap algorithm to detect traffic-correlated segment clusters for revealing the road traffic correlation. Given the strong capability and high efficiency of Infomap in handling network structures with massive vertices and edges, we adopted the Infomap algorithm to partition our network into communities with strong spatial interactions.

2.3. Information entropy and its usage in homophily investigation

Information entropy was borrowed from the traditional definition of entropy (first recognized in classical thermodynamics) that describes the physical property commonly associated with a state of disorder, randomness, or uncertainty (Lieb & Yngvason, 1998). The concept of information entropy, also called Shannon entropy, was introduced by mathematician Claude Shannon in 1948 (Shannon, 1948). As an essential component in information theory, information entropy describes events’ levels of uncertainty inherent in the events’ possible outcomes. In other words, it determines how much information an event contains: the more deterministic the event is, the less information it has. Since its proposal, information entropy has been widely used in various domains that include Cryptology (Sethi & Sharma, 2012), Linguistics (Bentz et al., 2017), Physics (Beck, 2009), and Computer Sciences (Pinto et al., 2019), to list a few. The applications of entropy in the social sciences, though emerging relatively later than other fields, can be traced back to the pioneering work by Theil and Finizza (1971). A comprehensive review conducted by Proops (1987) detailed early applications of information entropy in social sciences.

In recent years, the notably strengthened human interactions taking place in digital environments have led to a refound fascination with applying information theory in the social sciences, and the emerging big data sources and computational advances allow an increased interest in more sophisticated nonlinear methods and measures (Hilbert, 2021). Numerous efforts have been made to incorporate the concept of information entropy, aiming to reveal, investigate, and explain the homophily effects at various geographical scales. For example, Vanhoof et al. (2018) proposed a mobility entropy indicator that describes the diversity of individual movement patterns captured from mobile phones. Similarly, Lenormand et al. (2020) used mobile phone data and adopted entropy-based measures to quantify the attractiveness of a location in the Rio de Janeiro Metropolitan Area, Brazil, as the diversity of visitors’ location of residence. Taking Philadelphia as a study case, Kramer and Kramer (2019) offered a series of entropic measures to analyze residential segregation (from a racial perspective) at both the local (neighborhood) level and the greater (city/region) area. Zambon et al. (2017) investigated the polycentric development in Europe by proposing an entropy-based indicator of urban centrality built upon local-scale diversity in soil sealing levels. Besides the selected efforts mentioned above, more entropy-based indices that facilitate homophily investigation can be found in a study by Mora and Ruiz-Castillo (2011).

With this regard, we borrowed the concept of information entropy to measure the chaoticity (or disorder) of Social Vulnerability Index (SVI) themes and selected demographic/socioeconomic variables within detected communities formed by strong human spatial interactions.

3. Study area and data

3.1. Study area

We selected the top twenty most population MSAs in the U.S. as our study areas (Fig. 1) according to the MSA population totals in 2020 (U.S. Census Bureau, 2020). The definition of MSA was established by the U.S. Office of Management and Budget, referring to a region that consists of at least one city (with a minimum population of 50,000) and surrounding communities that are strongly linked by social and economic factors. MSAs are generally densely populated urban fabrics, ensuring abundant digital mobility records that facilitate the summarization of travel patterns in these MSAs. The MSAs included in our study are New York–Newark–Jersey City (New York MSA), Los Angeles–Long Beach–Anaheim (Los Angeles MSA), Chicago–Naperville–Elgin (Chicago MSA), Dallas–Fort Worth–Arlington (Dallas MSA), Houston–The Woodlands–Sugar Land (Houston MSA), Washington–Arlington–Alexandria (Washington D.C. MSA), Miami–Fort Lauderdale–Pompano Beach (Miami MSA), Philadelphia–Camden–Wilmington (Philadelphia MSA), Atlanta–Sandy Springs–Alpharetta (Atlanta MSA), Phoenix–Mesa–Chandler (Phoenix MSA), Boston–Cambridge–Newton (Boston MSA), San Francisco–Oakland–Berkeley (San Francisco MSA), Riverside–San Bernardino–Ontario (Riverside MSA), Detroit–Warren–Dearborn (Detroit MSA), Seattle–Tacoma–Bellevue (Seattle MSA), Minneapolis–St. Paul–Bloomington (Minneapolis MSA), San Diego–Chula Vista–Carlsbad (San Diego MSA), Tampa–St. Petersburg–Clearwater (Tampa MSA), Denver–Aurora–Lakewood (Denver MSA), and St. Louis MSA.

SafeGraph, our mobility data source (described in Section 3.2), details the tract-level device count on a daily basis. We calculated the daily sample count of each MSA by summing the number of devices in all Census Tracts within each MSA in 2019 and dividing this number by 365. We further derived the data representativeness for MSA by dividing the daily sample count by the population of the corresponding MSA (Huang et al., 2022). Detailed information on these MSAs can be found in Table 1. These MSAs considerably differ in total population, daily sample size, and data representativeness, with the Dallas MSA having the highest representativeness of 7.97 %, while the San Francisco MSA having the lowest representativeness of 4.89 %. Note that certain MSAs

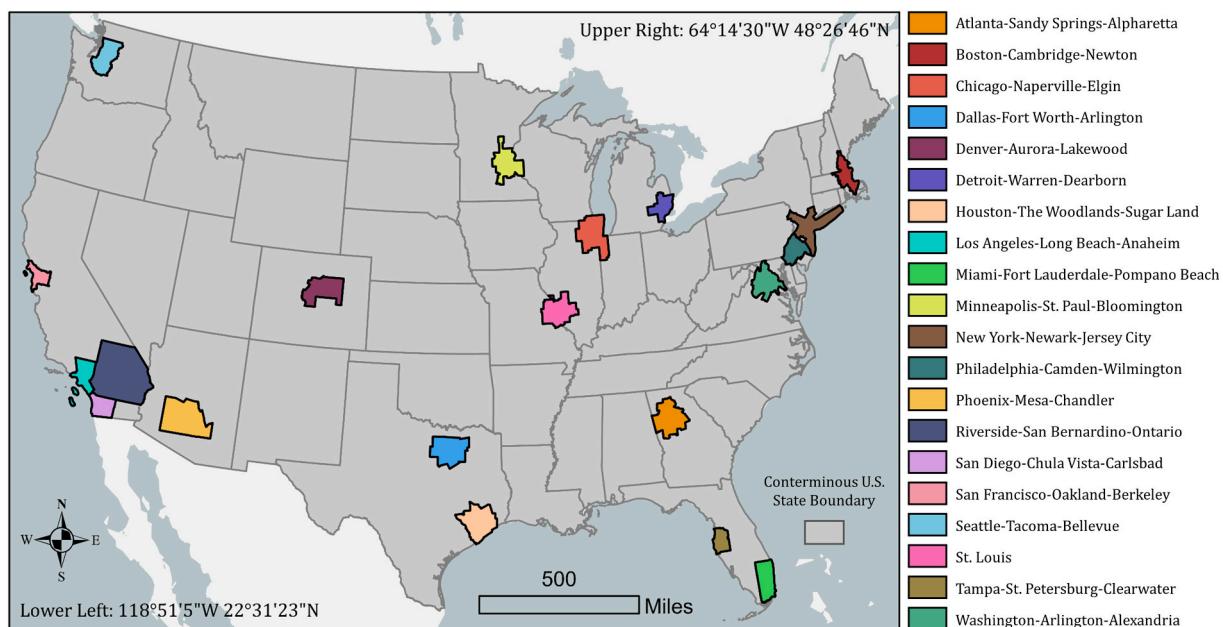


Fig. 1. The geographic locations of the top twenty most population MSAs in the U.S. The boundaries of these MSAs are derived from the 2019 TIGER/Line Shapefile products issued by the U.S. Census Bureau.

Table 1
Statistics of selected twenty MSAs.

MSA full name ^a	MSA short name	Covered U.S. states	# Census Tracts	Population ^b	Daily samples ^c	Representativeness ^d
New York-Newark-Jersey City	New York	NY-NJ-PA	4542	19,124,359	1,213,425	6.34 %
Los Angeles-Long Beach-Anaheim	Los Angeles	CA	2929	13,109,903	723,183	5.52 %
Chicago-Naperville-Elgin	Chicago	IL-IN-WI	2215	9,406,638	624,522	6.64 %
Dallas-Fort Worth-Arlington	Dallas	TX	1312	7,694,138	613,424	7.97 %
Houston-The Woodlands-Sugar Land	Houston	TX	1072	7,154,478	540,093	7.55 %
Washington-Arlington-Alexandria	Washington D.C.	DC-VA-MD-WV	1361	6,324,629	378,540	5.99 %
Miami-Fort Lauderdale-Pompano Beach	Miami	FL	1219	6,173,008	432,612	7.01 %
Philadelphia-Camden-Wilmington	Philadelphia	PA-NJ-DE-MD	1477	6,107,906	386,113	6.32 %
Atlanta-Sandy Springs-Alpharetta	Atlanta	GA	951	6,087,762	472,569	7.76 %
Phoenix-Mesa-Chandler	Phoenix	AZ	991	5,059,909	302,476	5.98 %
Boston-Cambridge-Newton	Boston	MA-NH	1007	4,878,211	249,635	5.12 %
San Francisco-Oakland-Berkeley	San Francisco	CA	980	4,696,902	229,523	4.89 %
Riverside-San Bernardino-Ontario	Riverside	CA	821	4,678,371	280,238	5.99 %
Detroit-Warren-Dearborn	Detroit	MI	1,301	4,304,136	317,091	7.37 %
Seattle-Tacoma-Bellevue	Seattle	WA	720	4,018,598	220,508	5.49 %
Minneapolis-St. Paul-Bloomington	Minneapolis	MN-WI	785	3,657,477	231,654	6.33 %
San Diego-Chula Vista-Carlsbad	San Diego	CA	628	3,332,427	177,114	5.31 %
Tampa-St. Petersburg-Clearwater	Tampa	FL	746	3,243,963	240,534	7.41 %
Denver-Aurora-Lakewood, CO	Denver	CO	620	2,991,231	178,723	5.97 %
St. Louis, MO-IL	St. Louis	MO-IL	615	2,805,473	215,492	7.68 %

^a Full names of U.S. states and their corresponding short forms can be found in Table A in the Appendix.

^b Population statistics are the 2020 population totals in these MSAs, retrieved from U.S. Census Bureau at <https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-totals-metro-and-micro-statistical-areas.html>.

^c Daily samples suggest the total of device counts in all Census Tracts within each MSA in the year 2019 divided by 365.

^d Representativeness denotes the ratio between daily samples and population in each MSA.

(e.g., New York MSA and Boston MSA) cover multiple U.S. states.

3.2. Mobile phone location data

The mobility records used in this study were derived from the Social Distancing Metrics (SafeGraph, 2020) provided by SafeGraph (<https://www.safegraph.com/>), a commercial company that provides insights on the visitation of physical places using locational data from mobile devices. The mobility records are originally summarized at the Census Block Group level, updated on a daily basis, cover the entire Conterminous U.S., and span from January 1, 2019, to April 16, 2021. SafeGraph collects mobility insights using a panel of GPS points from

around 45 million anonymous mobile devices, which comprise around 10 % of mobile devices in the entire U.S.

SafeGraph first determines the home locations of device users by assigning them to a Geohash-7 granularity (153 m × 153m), taking advantage of devices' nighttime locations over a period that spans six weeks (SafeGraph, 2020). Further, device holders' mobility patterns are summarized at the Census Block Group level and reported on a daily basis. Given that SafeGraph identifies the common nighttime location of each device (serving as the "home" location), the movement patterns summarized at various geographical units always start from where device holders' homes are located. To our best knowledge, the mobility dataset released by the SafeGraph company is among the best-quality

mobility datasets that have been open-sourced to the public. Since its release, numerous studies have been conducted that well demonstrate the validity of this dataset (Kang et al., 2020; Li et al., 2021). Its high representativeness makes it an ideal data source to summarize long-term human spatial interactions, facilitating the detection of communities with strong spatial interactions. To protect users' privacy, records are discarded if fewer than five devices visit an establishment in a month from a given Census Block Group (SafeGraph, 2020). However, such a privacy-protecting measure and the anonymization procedure do not affect the quality of this dataset (SafeGraph, 2019).

The COVID-19 pandemic that started in early 2020 led to mobility restrictions with varying strictness, which severely impact human travel patterns at various levels, potentially resulting in uncertainties in detected communities. Thus, we used data that cover the entire 2019 (January 1, 2019, to Dec 31, 2019) in this study. We further re-aggregated the SafeGraph's mobility dataset to a coarser geographical unit, i.e., from Census Block Group to Census Tract, consistent with the geographic units of selected demographic/socioeconomic variables and the social vulnerability index from the Centers for Disease Control and Prevention (2021) (details can be found in the following sessions). Such a re-aggregation to the Census Tract level also mitigates the low sampling issues that occur at the Block Group level noted by Huang et al. (2020). To eliminate the influence of excessively long external travels, often the ones from or to places outside the MSA, we restricted the origins and destinations to be within each MSA.

3.3. Demographic and socioeconomic variables

To investigate the entropy within communities detected via mobile phone records, we retrieved selected demographic and socioeconomic variables from the latest (at the time of writing) five-year American Community Survey (ACS) data, i.e., the 2015–2019 ACS five-year estimates at the U.S. Census Tract level, obtained from United States Census Bureau (<https://www.census.gov/en.html>). We selected five variables: 1) household income; 2) the percentage of Black (% Black); 3) the percentage of education lower than high school (% low education); 4) unemployment rate (% unemployment); 5) the percentage of households without a car (% no car ownership). These five selected variables are widely recognized indicators that respectively correspond to economic status, race, educational attainment, idle labor, and transportation (Huang et al., 2020; Huang et al., 2022). Note that the tract-level household income is the median value among all households within a certain tract. The community-level entropy of these variables denotes the chaoticness (or disorder) of demographic and socioeconomic indicators in communities that are strongly connected by human spatial interactions.

3.4. CDC social vulnerability index (SVI)

Apart from investigating community-level entropy of separated demographic and socioeconomic variables, we also explore community-level entropy of compound indicators. CDC Social Vulnerability Index (SVI), hereafter referred to as SVI for simplicity, is designed and maintained by the Agency for Toxic Substances and Disease Registry (ATSDR) to help public health officials and emergency response planners identify and map the communities that need support before, during, and after hazardous events (CDC SVI Documentation 2018, 2021). Although it was initially designed for mitigating hazardous events, SVI has been widely applied in various domains that include urban planning (Alizadehtazi et al., 2020), social studies (Horse et al., 2020), and public health (An & Xiang, 2015; Jones et al., 2020), given its comprehensiveness in summarizing local demographic and socioeconomic settings.

SVI ranges from 0 to 1, with 0 representing least vulnerable while 1 representing most vulnerable. SVI is calculated from summed percentile rankings at the U.S. Census Tracts from a total of fifteen demographic and socioeconomic variables that cover four major themes: 1)

socioeconomic status (*Theme1*), 2) household composition and disability (*Theme2*), 3) minority status and language (*Theme3*), and 4) housing types and transportation (*Theme4*). An overall tract ranking (*Themes*) is derived by summing all four themes, ordering the tracts, and then calculating overall percentile rankings. Table 2 presents the notation, description, and involved variables of the SVI used in this study.

4. Methodology

4.1. MSA mobility profile investigation

Mobility patterns derived from the SafeGraph dataset can be regarded as a network (G), with centroids of Census Tracts as nodes (N) and the number of travels between them as edges (E), i.e., $G = (N, E)$. As our network is a directed weighted graph, there are two types of nodes: origin node N_o (φ_o, ϕ_o) and destination node N_d (φ_d, ϕ_d), with φ and ϕ denoting their corresponding latitude and longitude, respectively. The edge between N_o and N_d are characterized by the directed travel count as weight, i.e., $E_{o \rightarrow d}^w$, and the distance between them, i.e., $E_{o \rightarrow d}^{dis}$. Given the property of directed weight graphs, $E_{o \rightarrow d}^w$ and $E_{d \rightarrow o}^w$ might differ, while $E_{o \rightarrow d}^{dis}$ and $E_{d \rightarrow o}^{dis}$ have the same value. In this study, the distance between N_o and N_d , i.e., $E_{o \rightarrow d}^{dis}$, is defined by the haversine formula that captures the Great Circle Distance:

$$E_{(o \rightarrow d)}^{dis} = 2r_e \times \sin^{-1} \sqrt{\left(\sin^2\left(\frac{\varphi_d - \varphi_o}{2}\right) + \cos\varphi_o \cos\varphi_d \sin^2\left(\frac{\phi_d - \phi_o}{2}\right) \right)} \quad (1)$$

where r_e denotes the radius of the earth in kilometers (around 6371 km).

We first explored the distribution of edge weights, $E_{(o \rightarrow d)}^w$, in selected twenty MSAs using the complementary cumulative distribution function (CCDF): $CCDF(E_{(o \rightarrow d)}^w) = P(X > E_{(o \rightarrow d)}^w)$. Further, we revealed how travel patterns in these MSAs follow the "distance decay" effect by fitting the edge distance ($E_{(o \rightarrow d)}^{dis}$) to a power-law distribution, one of the most commonly used distributions to model human mobility patterns (Brockmann et al., 2006; Jiang et al., 2021). Such a distribution pattern in a discrete form is also noted as Zipf's law, a fundamental principle that governs human spatial interactions (Zhao et al., 2015). The power-law distribution is controlled by two parameters, i.e., α and β :

$$p(E_{(o \rightarrow d)}^{dis}) = \alpha (E_{(o \rightarrow d)}^{dis})^{-\beta} \quad (2)$$

Table 2

The documentation of CDC SVI used in this study.

Notation	Description	Involved variables
<i>Theme1</i>	Socioeconomic status	% Below Poverty Level % Unemployed Per Capita Income % Age 25 or older with No High School Diploma
<i>Theme2</i>	Household composition and disability	% Age 65 or older % Age 17 or younger % civilian noninstitutionalized population with a disability % Single-parent households with children under 18
<i>Theme3</i>	Minority Status and Language	% minority (all persons except white, non-Hispanic) % Age 5 or older who speak English less than well
<i>Theme4</i>	Housing type and transportation	% Housing in structures with 10 or more units % Mobile houses % Occupied housing units with more people than rooms % Households without a Vehicle % In institutionalized group quarters
<i>Themes</i>	Summation of all the above themes	All above

where $E_{o \rightarrow d}^{dis}$ is the edge distance, $p(E_{o \rightarrow d}^{dis})$ suggests the occurrence probability of this edge distance, α and β are two parameters to be fitted. In this study, the number of iterations to find the optimal settings of α and β is capped at 10,000. When $\beta > 1$, the occurrence probability follows an inverse proportional relationship with edge distance (Jiang et al., 2021). Jiang et al. (2009) further found that, in most cases, β is bounded between 1 and 3.

4.2. Infomap community detection

The Infomap community detection algorithm derives communities using the probability flow of random walks on a network as a proxy for information flows in the real system and further decomposing the network into communities via a compressed description of the probability flow, leading to simplified and highlighted regularities in the structure and their relationships (Rosvall & Bergstrom, 2008). Infomap's random walker moves in a random manner from node to node in the network, resulting in connections with varying weights. To partition network G (with a total of N nodes) into M communities, the lower bound on code length is set to be $L(M)$. The general goal is to minimize $L(M)$ (Rosvall et al., 2009):

$$L(M) = q \curvearrowright H(\Omega) + \sum_{i=1}^M p_i^i \curvearrowright H(P^i) \quad (3)$$

where the first term, i.e., $q \curvearrowright H(\Omega)$, denotes the entropy of the movement between communities, and the second term, i.e., $\sum_{i=1}^M p_i^i \curvearrowright H(P^i)$ denotes the entropy of movements within communities. $H(\Omega)$ denotes the entropy of the community names using Huffman coding, and $H(P^i)$ is the entropy of the within-community movements. Here, the entropy follows the definition of information entropy, i.e., Shannon entropy, which is detailed in Section 4.3. $q \curvearrowright = \sum_{i=1}^M q_i \curvearrowright$ is the probability that the random walk switches communities on any given step while p_i^i is the fraction of intra-community movement that occurs in community i plus the probability of exiting the community k . For the detailed workflow of Infomap, please refer to Rosvall et al. (2009). To ensure statistical robustness in the entropy calculation of community members, we removed communities with a small number of members by setting up a threshold of 20, an empirical value. Eventually, the Infomap algorithm partitions the Census Tracts of an MSA into a total of M satisfactory communities, forming a community set $C = \{C_1, C_2, \dots, C_M\}$.

4.3. Community entropy

The detected communities with strong human spatial interactions are determined jointly by their geographic settings, zonal functionality, and social gravity. Under such strong spatial ties, the chaoticity of societal settings across community members is inversely related to the homophily level of this community. In this study, we investigated the distribution of the societal settings within communities using the Shannon entropy measurement (Shannon, 1948). For a certain MSA and a certain variable to be investigated (either sociodemographic status or SVI), we formed a variable set $X = \{x_1, x_2, \dots, x_N\}$ that includes all Census Tracts of an MSA. We re-ranked X , derived the deciles, and assigned unique labels given different deciles for a total of K Census Tracts partitioned to C_i , thus leading to a community-level label set $\mathcal{L}_i = \{\mathcal{L}_i^1, \mathcal{L}_i^2, \dots, \mathcal{L}_i^K\}$. Shannon entropy with the base of e was further employed to describe the chaoticity of \mathcal{L}_i , which can be explicitly written as:

$$H(\mathcal{L}_i) = - \sum_{j=1}^T P(\mathcal{L}_i^j) \ln P(\mathcal{L}_i^j) \quad (4)$$

where $H(\mathcal{L}_i)$ denotes the community entropy of C_i , $P(\mathcal{L}_i^j)$ denotes the

occurrence probability of label \mathcal{L}_i^j , and T denotes the number of unique labels in \mathcal{L}_i . The community entropy $H(\mathcal{L}_i)$ quantifies the chaoticity of labels \mathcal{L}_i (variables in different deciles), thus reflecting different levels of community *homophily*.

5. Results

5.1. MSA mobility profiles

We first explored the mobility profiles in selected MSAs. The CCDF curves suggest that the edge weight distribution of all MSAs follows a log-normal pattern featured by exponentially decreased occurrence probability with increased edge weights, suggesting a heavily left-skewed distribution (before taking the logarithm) towards a small travel amount (Fig. 2). This unanimous weight distribution pattern, however, is expected and supported by existing studies that rely on other data sources, such as social media data (Wang et al., 2018) and digital device applications (Piorkowski, 2009). Despite the similar weight distribution pattern, MSAs present notable nuances in their CCDF curves, presumably due to the discrepancies in MSAs' land use distribution patterns, road network structures, transportation means, and data representativeness (Table 1). Compared to the CCDF curves of New York MSA and Los Angeles MSA, the ones of Houston MSA and Atlanta MSA are notably right-shifted, suggesting that the extremely large edge weights occur more frequently.

We further explored how MSA's travel patterns follow the distance-decay effect by fitting the travel distance $E_{o \rightarrow d}^{dis}$ to a power-law distribution parametrized by α and β . The Probability Density Function (PDF) curves of $E_{o \rightarrow d}^{dis}$ and the fitted α and β values of selected MSAs are presented in Fig. 3. The results suggest that the widely recognized power-law well governs the travel patterns in MSAs at the Census Tract level, evidenced by the linearly negative relationship between $E_{o \rightarrow d}^{dis}$ and its logarithmic PDF value (Fig. 3). The top three MSAs with the highest β values appear in Boston, San Francisco, and Los Angeles MSAs, while the bottom MSAs with the lowest β values appear in Atlanta, Houston, and St. Louis. As the scaling parameter β represents the law's exponent, a larger β dictates that, with the increase of travel distances, the occurrence of such travel decreases in a more significant manner. We observe that all MSAs (except Atlanta MSA) present β values larger than 1, suggesting that distance decays more severe than inverse proportionality. Although the β of Atlanta MSA is smaller than 1, it does not deviate much from the inverse proportionality: $\beta = 0.995$ [95 % CI, 0.877–1.114]. We also notice that MSAs with lower β values tend to present more sprawling-out patterns with the high ratios of car-dominant travels. Intuitively, β is collectively decided by a variety of determinants that describe the MSAs' contextual settings. Fig. 4 presents the power-law trend of travel distance with fitted α and β in selected twenty MSAs with a distance bin of 3 km. The inconsistency of the first bin (i.e., travel distance from 0 to 3 km) compared to other bins can be explained by the varying spatial constructs of Census Tracts' centroids. Here, we discard the first bin of each MSA during the curve fitting. The fitting results are promising, evidenced by their high R^2 values (all above 0.9) (Fig. 4). San Francisco shows the highest R^2 of 0.987, while Atlanta MSA shows the lowest R^2 of 0.919. Detailed statistics regarding MSAs' α and β values, their associated confidence intervals, the goodness of fit, and total trips involved can be found in Table 3.

5.2. Detected communities

The Infomap algorithm partitions the aggregated mobility network of each MSA into multiple communities featured by strong spatial interactions (Fig. 5). We notice that MSAs contain various numbers of communities, and detected communities greatly differ in terms of their member counts and spatial coverages (Table 4). New York (with the number of detected communities as 27), Chicago (20), Los Angeles (16),

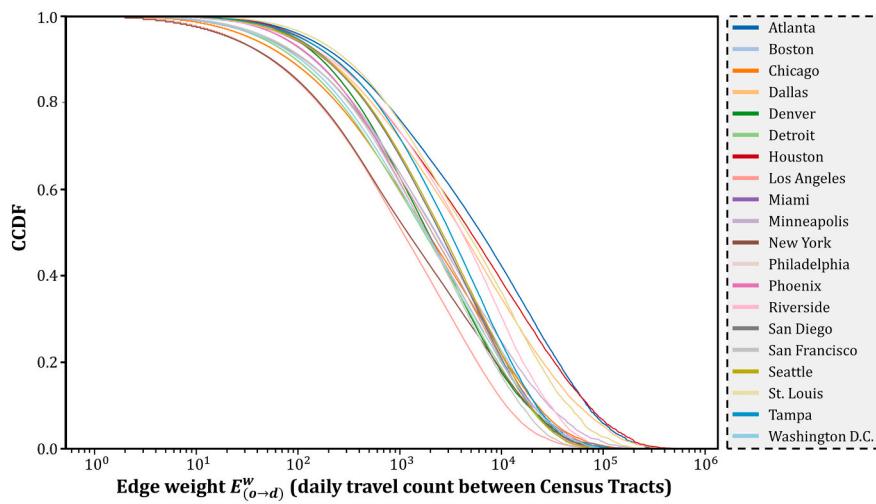


Fig. 2. The complementary cumulative distribution function (CCDF) of edge weight $E_{(o-d)}^w$ in the U.S. top 20 most populated MSAs. Note that $CCDF(E_{(o-d)}^w) = P(X > E_{(o-d)}^w)$. Note that X-axis is on a logarithmic scale.

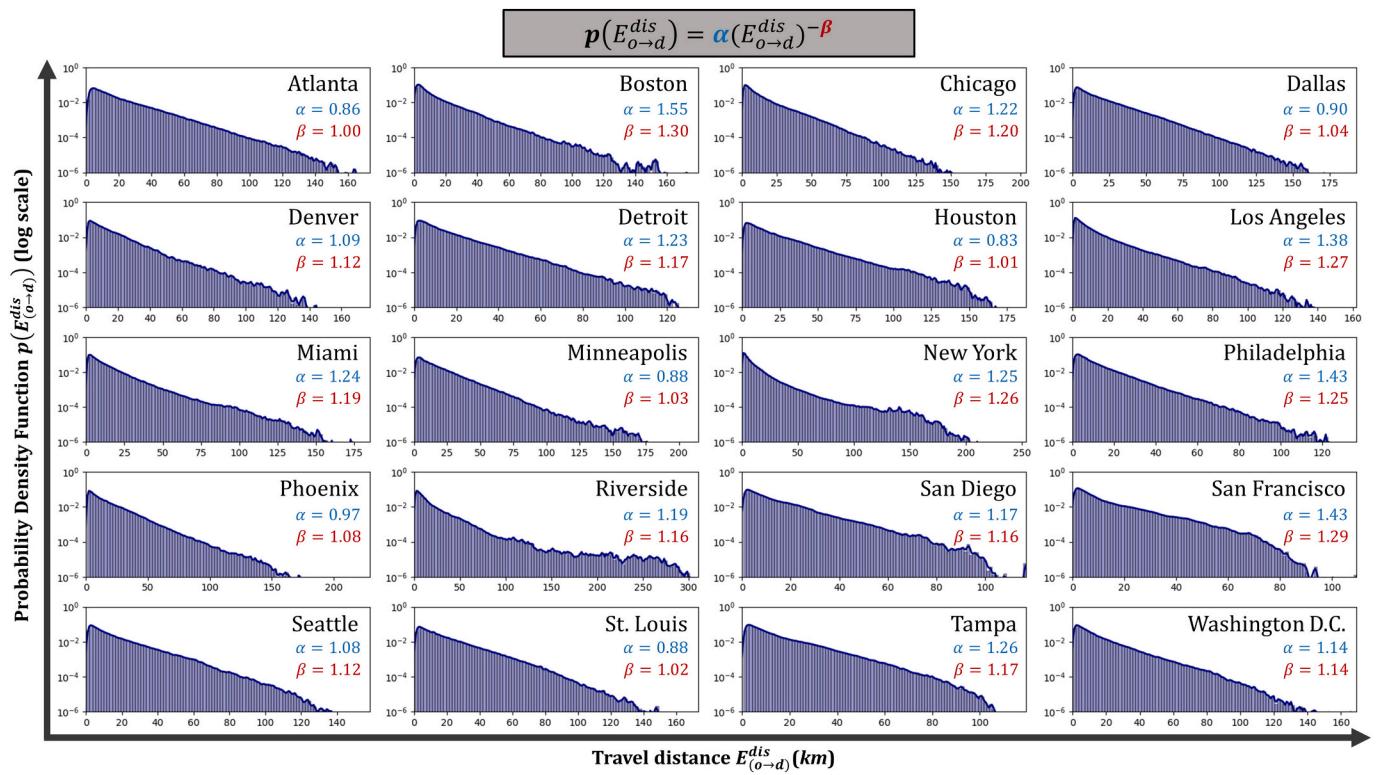


Fig. 3. The Probability Density Function (PDF) of travel distance in the U.S. top 20 most populated MSAs and their corresponding α and β . Note that the Y-axis is on a logarithmic scale.

and Boston (16) are the top four MSAs with the most detected communities, given their large population sizes and extensive spatial coverages. While Seattle (5), Denver (6), and San Diego (6) MSAs are with the lowest number of detected communities. Further, the spatial distribution patterns of detected communities are largely dictated by Tobler's law (Tobler, 1970), i.e., near things are more related than distant things. The formation of most communities is based on the adjacent, spatially closed Census Tracts, reflecting that human moving patterns follow Tobler's law to a great extent. However, exceptions do exist in many MSAs, e.g., New York MSA (Fig. 5k), Detroit MSA (Fig. 5f), and Boston MSA (Fig. 5b), where a number of spatially disconnected Census Tracts exhibit strong spatial interactions. Such strong interactions that

transcend spatial distances are presumably due to the MSA's road network structures, public transportation, and land use patterns. For example, the Census Tracts with strong commercial activities can form a community with other distant ones, given the regular and frequent travels among them. A community containing spatially disconnected Census Tracts is also possible, given the accessibility provided by road networks and public transportation services. The construction of communities from spatially disconnected spatial units has also been revealed in other studies (Poorthuis, 2018; Yildirimoglu & Kim, 2018). We further observe that the detected communities do not usually follow administrative boundaries. The boundaries of detected communities are not aligning with the boundaries of the U.S. County, an upper census

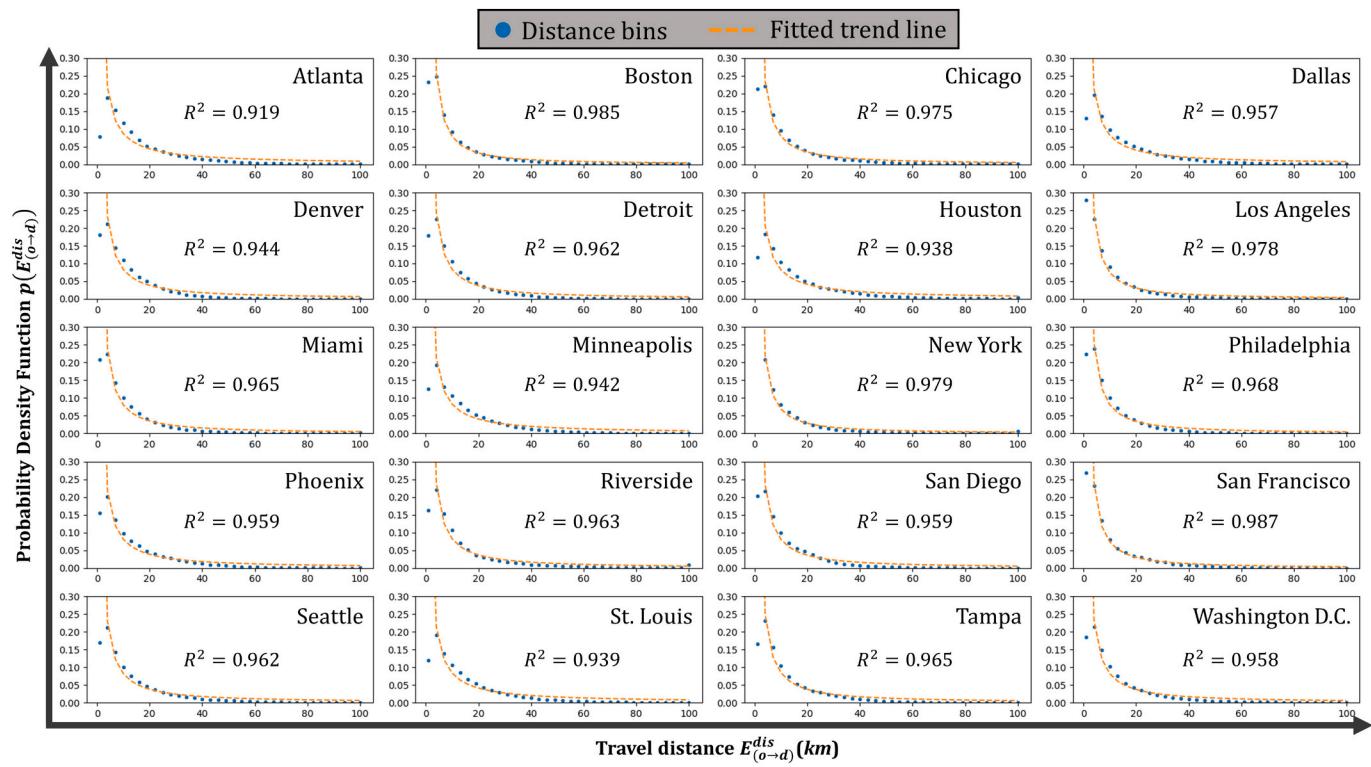


Fig. 4. The power-law trend of travel distance with fitted α and β in the U.S. top 20 most populated MSAs with a distance bin of 3 km.

Table 3
The power-law fitting results for the U.S. top 20 most populated MSAs.

MSAs	α		β		R^2	Total trips
	Fitted value	95 % CI ^a	Fitted value	95 % CI		
Atlanta	0.856	(0.644, 1.068)	0.995	(0.877, 1.114)	0.919	409,028,501
Boston	1.552	(1.349, 1.754)	1.296	(1.223, 1.369)	0.985	166,721,739
Chicago	1.222	(1.030, 1.414)	1.191	(1.107, 1.275)	0.975	498,823,619
Dallas	0.896	(0.733, 1.059)	1.040	(0.950, 1.129)	0.957	570,696,837
Denver	1.091	(0.846, 1.337)	1.120	(1.004, 1.236)	0.944	136,559,660
Detroit	1.226	(0.991, 1.460)	1.167	(1.066, 1.268)	0.962	241,949,321
Houston	0.830	(0.651, 1.009)	1.005	(0.901, 1.109)	0.938	482,372,777
Los Angeles	1.377	(1.162, 1.593)	1.269	(1.182, 1.356)	0.978	564,121,666
Miami	1.241	(1.010, 1.473)	1.188	(1.088, 1.288)	0.965	378,045,011
Minneapolis	0.884	(0.696, 1.071)	1.028	(0.925, 1.132)	0.942	378,045,011
New York	1.247	(1.059, 1.435)	1.258	(1.175, 1.342)	0.979	954,406,839
Philadelphia	1.434	(1.168, 1.700)	1.247	(1.145, 1.349)	0.968	278,902,070
Phoenix	0.971	(0.793, 1.150)	1.077	(0.985, 1.170)	0.959	262,736,237
Riverside	1.190	(0.970, 1.410)	1.159	(1.061, 1.256)	0.963	194,428,871
San Diego	1.168	(0.938, 1.397)	1.160	(1.057, 1.263)	0.959	139,393,591
San Francisco	1.435	(1.261, 1.609)	1.287	(1.220, 1.355)	0.987	155,097,166
Seattle	1.084	(0.887, 1.282)	1.122	(1.028, 1.216)	0.962	162,509,605
St. Louis	0.880	(0.689, 1.072)	1.024	(0.918, 1.130)	0.939	188,770,341
Tampa	1.264	(1.030, 1.497)	1.175	(1.077, 1.273)	0.965	203,897,539
Washington D.C.	1.136	(0.913, 1.358)	1.143	(1.041, 1.246)	0.958	255,153,577

^a CI: Confidence Interval.

unit aggregated by Census Tracts. For MSAs that cover multiple states, we notice that their community boundaries are not always consistent with state borders. Such discrepancies in the observed boundaries of detected communities and administrative boundaries have also been observed in existing studies (e.g., Li et al., 2021; Poorthuis, 2018). The detected communities in this study support the claim that communities with strong spatial interactions are largely spatial connected but also contain areas that are spatially disconnected, and the spatial boundaries of such communities are not fully consistent with the established administrative boundaries.

5.3. Community entropy of selected demographic and socioeconomic variables

After the detection of communities, we explored the chaoticness of selected demographic and socioeconomic variables (described in Section 3.3) across community members and further documented the average and the standard deviation (SD) of the community entropy of the corresponding variable within each MSA. The average statistics reflect the overall homophily levels across communities, while the SD statistics reflect the inconsistency (or dispersion) in homophily levels across communities. Table 5 details the relevant statistics, and Fig. 6 showcases



Fig. 5. Detected communities in the U.S. top 20 most populated MSAs.

a violin plot the describes the distribution of community entropy in each selected MSA.

Several MSAs stand out in this comparison. For example, among all selected MSAs, Atlanta MSA presents the lowest *average* community entropy in terms of household income (1.916), % Black (1.674), and % low education (1.843) (Table 5). As lower entropy denotes less chaoticness of variables, Atlanta MSA presents the strongest *homophily* effect from these perspectives. In other words, compared to other MSAs, Census Tracts in the Atlanta MSA with similar societal constructs in household income, % Black, and % low education are more likely to form communities characterized by strong spatial interactions. Interestingly, the Atlanta MSA also owns the highest *SD* community entropy in household income, indicating its strongest dispersion in income *homophily* across its communities. The strong *homophily* effect in Atlanta MSA revealed in this study coincides with existing studies that document the notable class and racial divisions with widening disparities in the Atlanta MSA (e.g., Bullard et al., 1999; Wyczalkowski et al., 2020). In

sharp contrast to the Atlanta MSA, the Miami MSA shows the highest entropy in household income (2.248) and the third-highest entropy in % Black (2.108). It means that Census Tracts in the Miami MSA with different income levels and the percent of Blacks mingle with each other well and display strong spatial interactions, potentially leading to high-entropy community formation. Another MSA that stands out is the New York MSA, as it presents the lowest community entropy in % no car ownership (1.693). This finding might be explained by its heavily-used transportation system characterized by extensive subway, bus, and taxi networks (Cramer et al., 2009). The transportation system in the New York MSA, with a massive amount of daily users, reshapes transit-oriented urban structures and potentially leads to low-entropy (high-*homophily*) communities in terms of car ownership.

5.4. Community entropy of SVI

After revealing the MSAs' community-level entropy of separated

Table 4

Statistics of detected communities in the U.S. top 20 most populated MSAs.

MSAs	Number of communities	Community members (average)	Community members (SD)	Community spatial coverages in km ² (average)	Community spatial coverages in km ² (SD)
Atlanta	12	73.75	83.30	1519.73	931.71
Boston	16	57.81	65.69	584.17	502.19
Chicago	20	108.20	84.55	752.87	763.52
Dallas	15	85.60	64.24	1260.47	899.29
Denver	6	96.00	94.87	2077.15	2347.05
Detroit	11	114.18	87.27	957.60	633.27
Houston	13	79.85	80.17	1593.90	1104.77
Los Angeles	16	181.88	100.75	755.88	825.61
Miami	8	150.88	84.54	1725.09	1924.68
Minneapolis	10	76.90	65.16	1700.87	1032.70
New York	27	167.00	125.23	811.75	864.78
Philadelphia	15	94.27	78.87	741.94	641.08
Phoenix	7	137.57	109.88	4421.45	3891.42
Riverside	12	63.67	27.06	1335.02	1964.10
San Diego	6	102.83	136.83	1084.45	1317.19
San Francisco	11	88.55	60.19	636.02	519.30
Seattle	5	134.80	22.68	2799.25	2149.34
St. Louis	7	78.14	65.56	1583.69	1159.33
Tampa	9	82.67	59.29	928.54	322.00
Washington	13	101.46	108.46	1240.92	1135.76
D.C.					

Note: SD = Standard deviation.

Table 5

The community entropy of selected demographic and socioeconomic variables in the U.S. top 20 most populated MSAs.

MSAs	Household income		% Black		% Low education		% Unemploy		% no car ownership	
	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD
Atlanta	1.916	0.408	1.674	0.393	1.843	0.392	2.113	0.118	2.015	0.263
Boston	1.936	0.244	1.990	0.237	1.923	0.208	2.073	0.118	1.870	0.173
Chicago	1.985	0.283	1.947	0.339	2.045	0.349	2.041	0.242	1.884	0.256
Dallas	2.022	0.150	2.073	0.332	1.878	0.178	2.179	0.094	2.133	0.113
Denver	2.061	0.222	1.970	0.067	2.008	0.217	2.218	0.057	2.174	0.210
Detroit	1.996	0.203	1.890	0.485	2.024	0.230	2.140	0.171	2.103	0.178
Houston	1.992	0.144	2.057	0.105	1.878	0.182	2.136	0.095	2.124	0.117
Los Angeles	2.042	0.197	2.045	0.339	2.085	0.336	2.229	0.082	2.144	0.182
Miami	2.248	0.098	2.108	0.198	2.142	0.245	2.237	0.103	2.206	0.317
Minneapolis	2.067	0.145	2.041	0.392	2.035	0.241	2.214	0.132	2.099	0.168
New York	1.984	0.235	1.968	0.230	2.102	0.250	2.166	0.128	1.693	0.290
Philadelphia	2.033	0.213	1.980	0.157	2.033	0.345	2.076	0.136	1.978	0.213
Phoenix	2.024	0.132	2.036	0.143	1.916	0.182	2.155	0.236	2.196	0.096
Riverside	2.135	0.163	2.151	0.164	2.004	0.164	2.213	0.176	2.167	0.125
San Diego	2.024	0.356	1.982	0.302	1.947	0.493	2.152	0.075	2.198	0.109
San Francisco	2.034	0.134	1.873	0.173	2.091	0.274	2.196	0.136	2.012	0.070
Seattle	2.151	0.262	2.123	0.146	2.001	0.195	2.258	0.096	2.234	0.121
St. Louis	2.015	0.285	1.912	0.186	2.106	0.330	2.126	0.158	2.018	0.123
Tampa	2.021	0.297	2.095	0.108	1.982	0.294	2.229	0.082	2.167	0.171
Washington D.C.	2.058	0.187	1.758	0.234	2.040	0.215	2.173	0.069	2.010	0.120

Note. average and SD (standard deviation) denote the mean entropy value and the standard deviation of entropy value for all communities within an MSA.

demographic and socioeconomic variables, we explored the chaoticness of compound indicators, i.e., SVI (described in [Section 3.4](#)), across community members. Following the previous session, we documented the *average* and the standard deviation (*SD*) of community entropy in different SVI themes within each MSA ([Table 6](#) and [Fig. 7](#)).

For *Theme1*, the Detroit MSA presents the lowest *average* community entropy (1.873), while the Seattle MSA presents the highest one (2.160) ([Table 6](#)). As *Theme1* is a compound indicator that reflects socioeconomic status, communities in the Detroit MSA show the strongest *homophily* phenomena in terms of the assimilation of socioeconomic status, while the Seattle MSA is the opposite, showing the strongest *heterophily*. The strong socioeconomic *homophily* in the Detroit MSA is supported by [Silverman \(2005\)](#), who documented the strong socioeconomic segregation in the City of Detroit. The Atlanta MSA, with the lowest *average* community entropy in household income (see [Table 4](#)), ranks the third lowest in socioeconomic status. Such a difference in ranking can be explained by the construction of *Theme1* in SVI, which

considers not only income levels but also unemployment and education attainment (see the construction of *Theme1* in [Table 2](#)). For *Theme2* that reveals household composition and disability, the Atlanta MSA presents the lowest *average* community entropy (1.831; the values for other MSAs are all above 2). In comparison, the Miami MSA shows the highest *average* community entropy (2.250), indicating its great mingling of community members in terms of household composition and disability. The Seattle MSA shows the highest *average* community entropy in both *Theme3* (2.219) that reflects minority status and language and *Theme4* (2.253) that reflects household type and transportation. As expected, the New York MSA presents the lowest *average* community entropy in *Theme4* (2.041), given its heavily-used transportation system that may lead to low-entropy communities with similar transportation needs.

For *Themes* as a compound indicator that involves all perspectives, the top five MSAs with the highest *average* community entropy (towards *heterophily*) include Seattle (2.215), Los Angeles (2.165), Washington D.C. (2.162), Miami (2.151), and Minneapolis (2.142), while the top five

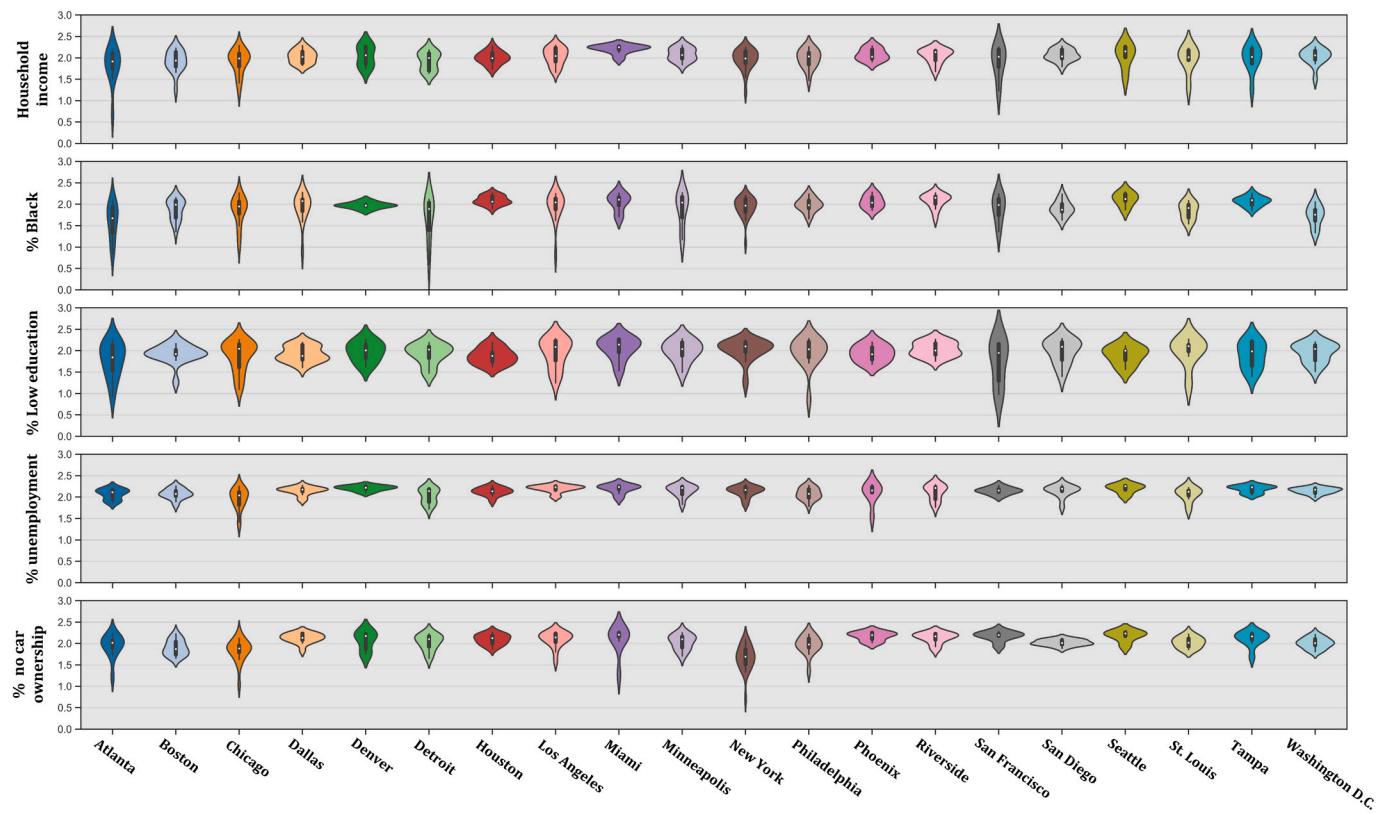


Fig. 6. Community entropy distribution of selected demographic and socioeconomic variables in the U.S. top 20 most populated MSAs.

Table 6
Community entropy of SVI in the U.S. top 20 most populated MSAs.

MSAs	Theme1		Theme2		Theme3		Theme4		Themes	
	Average	SD								
Atlanta	1.931	0.345	1.831	0.217	1.955	0.230	2.119	0.353	2.045	0.354
Boston	1.946	0.193	2.072	0.142	1.804	0.294	2.082	0.126	1.971	0.140
Chicago	2.018	0.298	2.028	0.188	2.130	0.420	2.140	0.203	2.088	0.298
Dallas	1.961	0.170	2.056	0.143	1.888	0.278	2.145	0.085	2.023	0.156
Denver	1.938	0.259	2.197	0.095	1.966	0.243	2.157	0.159	2.046	0.199
Detroit	1.873	0.209	2.074	0.191	2.141	0.340	2.144	0.095	2.046	0.198
Houston	1.927	0.161	2.128	0.150	1.791	0.206	2.152	0.104	2.054	0.149
Los Angeles	2.147	0.264	2.182	0.170	2.060	0.327	2.232	0.130	2.165	0.245
Miami	2.141	0.161	2.250	0.187	1.953	0.356	2.196	0.137	2.151	0.140
Minneapolis	2.110	0.181	2.164	0.155	2.109	0.355	2.122	0.130	2.142	0.175
New York	2.069	0.217	2.151	0.153	2.019	0.432	2.041	0.206	1.987	0.247
Philadelphia	2.015	0.165	2.104	0.137	2.045	0.113	2.215	0.092	2.052	0.149
Phoenix	1.911	0.194	2.161	0.245	1.892	0.140	2.156	0.109	2.041	0.189
Riverside	2.041	0.131	2.129	0.292	2.016	0.216	2.148	0.097	2.138	0.141
San Diego	1.979	0.326	2.065	0.190	2.080	0.231	2.085	0.186	1.969	0.293
San Francisco	2.085	0.168	2.029	0.215	1.964	0.280	2.199	0.107	2.062	0.127
Seattle	2.160	0.198	2.013	0.127	2.219	0.086	2.253	0.032	2.215	0.096
St. Louis	2.014	0.335	2.156	0.117	2.038	0.163	2.182	0.075	2.098	0.188
Tampa	2.116	0.195	2.033	0.133	1.958	0.169	2.202	0.192	2.130	0.209
Washington D.C.	2.145	0.123	2.130	0.109	2.095	0.273	2.132	0.105	2.162	0.093

Note. average and SD denote the mean entropy value and the standard deviation of entropy value for all communities within an MSA.

MSAs with the lowest average community entropy (towards *homophily*) include San Diego (1.969), Boston (1.971), New York (1.987), Dallas (2.023), and Phoenix (2.041).

6. Discussion

6.1. Community entropy: a promising homophily measurement

Diverging from existing efforts, we unravel the *homophily* effect at

the community level by extracting human mobility structures via community detection algorithms and using information entropy measurement to quantify the chaoticness of local settings within communities. Transcending the traditional investigations on residential segregation that focus more on the fixed spatial settings of neighborhoods, we investigate the *homophily* phenomenon from a mobility perspective using fine-grained mobile phone records that reflect the actual human-space interactions. Although the results suggest that the formation of communities largely consists of spatially adjacent Census Tracts in most

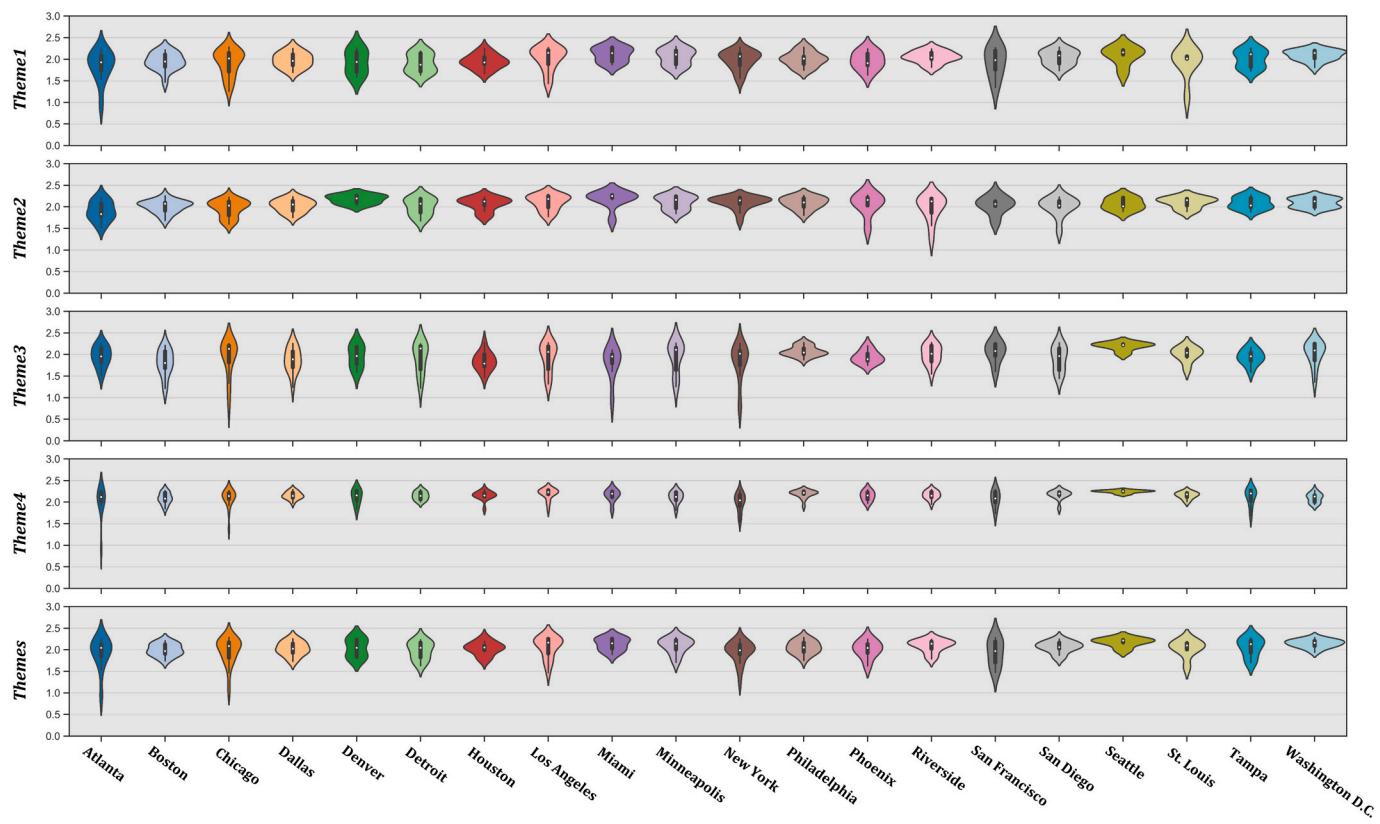


Fig. 7. Community entropy distribution of SVI in the U.S. top 20 most populated MSAs.

MSAs, we notice that communities featured by strong human interactions can sometimes transcend geographic proximity in modern metropolitans presumably due to road network structures, public transportation, and land use patterns. This finding demonstrates the importance of *homophily* investigation from a mobility perspective, as mobility networks are able to capture the *actual* spatial interactions that usually cross beyond the administrative bounds. This needs to be considered in the delineation of communities and can be further applied to broad community studies in social science. The spatially disconnected areas delineated in one community may emerge with gradually improved urban transit infrastructures. Under the strong spatial ties among spatially disconnected areas within a community, we further borrow information entropy (i.e., Shannon entropy) to quantify the chaoticity of social settings across community space. The more chaotic (high entropy) social settings are across community space, the less *homophily* this community exhibits. The concept and utilization of information entropy not only provide a standard measure of how similar/dissimilar communities vary across sociodemographic status but also establish a venue where cross-community comparisons can be conducted. Thus, we argue that community detection coupled with entropic measurement is a promising approach to revealing the *homophily* phenomenon. Our approach can be further applied to delineate other social-spatial concepts (e.g., neighborhoods) that are arbitrary to be defined solely based on administrative boundaries. Our approach by no means provides a definitive delineation of a community but a nuanced perspective to enrich the community studies in sociology by taking account of human-space interactions.

6.2. Policy implications

Homophily continues to characterize U.S. cities. The causes of the establishment of low-entropy communities that reveals strong homophily effect are multifaceted. The Great Migration in the U.S. (from 1916

to 1970) contributed to profound social, economic, demographic, and cultural changes in U.S. cities (Tolnay, 2003). However, during the Great Migration, both private practices and public policies largely constrained the housing options of socially disadvantaged groups, steering them into a few, clearly defined neighborhoods. This potentially leads to severe urban segregation. In the 21st century, although many pieces of evidence suggest that residential segregation has been improved in the US (e.g., Glaeser & Vigdor, 2012), a growing body of literature indicates that *homophily* that transcends geographic proximity still prevails, evident by the formation of strong spatial ties from individuals/places with similar societal settings (Wang et al., 2018). Our study again proves the ubiquity of the *homophily* phenomenon in modern metropolitans by quantifying this phenomenon using community detection algorithms coupled with entropy measurement.

To alleviate urban segregation and promote *heterophily*, we give the following suggestions. First, we need to enhance our knowledge of low-entropy communities, finding the underlying reasons that may lead to their *homophily* in order to appropriately design and implement policies that are necessarily multidimensional. For example, the results from our study suggest that the measures of community entropy greatly vary within an MSA; communities can point to a diverging tendency towards *homophily* or *heterophily* (e.g., the community entropy of household income and % Black in Atlanta MSA). Investigating the hidden determinants that potentially contribute to the disparity in community entropy within an MSA would benefit future policymaking. Second, efforts are needed to maintain community diversity in the face of gentrification by establishing necessary policies to preserve the availability of lower-cost housing, prioritize the increase of housing affordability, and attract investments that help improve services and deliver opportunities for residents within and outside communities. Third, it is highly encouraged to develop policies that provide better transit opportunities with inclusive destinations, given that many high-entropy communities observed in our study transcend geographic proximity. Thus, we believe

that optimized accessibility-based planning is necessary to build efficient and equitable transportation services that break travel barriers and facilitate the integration of all citizens.

6.3. Limitations and future directions

It is important to acknowledge the limitations of this work and provide guidelines for improvement in future studies. First, we used the Infomap algorithm to partition mobility networks into communities with strong spatial interactions. However, there exist other popular community detection approaches, notably the Louvain algorithm (De Meo et al., 2011) and the Leiden algorithm (Traag et al., 2019). Different algorithms may partition the mobility network into varying community structures, thus leading to nuances in community entropy. Future studies can explore how different community structures, due to the usage of different community detection algorithms, influence the general *homophily* levels at MSAs.

Second, our network was constructed using mobility records that cover the entire year of 2019. Although such data with a yearly span is able to summarize MSAs' structure that leads to stable community structures, the intra-year mobility dynamics remain underexploited in this study. The strong seasonality of human mobility patterns potentially leads to inconsistent community structures at different temporal scales within a year, thus resulting in the variation of *homophily* levels over time. Future efforts can be made to explore community *homophily* in a dynamic manner.

Third, before calculating information entropy within a community, we assigned unique labels to Census Tracts using their deciles. Binning (or discretizing) continuous variables before entropy calculation is a preferred step and similar efforts have been made to investigate interactions among categorized social classes (e.g., Xu et al., 2021). Nonetheless, the choice of how to discretize values affects entropy calculation. In this study, we chose to use deciles, i.e., dividing a continuous variable into a total of ten equal groups. However, different discretizing mechanisms should be explored in future studies.

Finally, we selected the top twenty most populated MSAs in the U.S as study areas and constructed mobility networks by summarizing their internal travels in 2019. We expect more efforts to be made by extending the spatial scope to the national wide, incorporating other mobility data sources, and extending the research period in a longitudinal manner.

7. Conclusion

Homophily, one of the most important regularities that govern human

Appendix A

Table A
U.S. state full names and short names that appear in
[Table 1](#).

Full names	Short names
Arizona	AZ
California	CA
Colorado	CO
Delaware	DE
District of Columbia	DC
Florida	FL
Georgia	GA
Illinois	IL
Indiana	IN
Maryland	MD
Massachusetts	MA
Michigan	MI
Minnesota	MN
Missouri	MO
New Hampshire	NH

spatial and social interactions, has been a hot research topic in many spatial and social domains. Taking advantage of human mobility records derived from around 45 million mobile devices in the U.S., we contribute to the current scholarship by providing an analytical framework to detect and delineate communities based on human-space interactions and presenting an empirical study of the top 20 most populated MSAs in the U.S. Our analytical framework that consists of community detection algorithms, the measures of information entropy, and spatial mapping has great potential to enrich the traditional community studies that are deeply rooted in sociology with a spatial perspective and to be applied to broader studies across disciplines. The generality, reproducibility, and applicability of our approach have been proved in the cross-metropolitan investigations in our study and can be augmented to larger and/or different geographic contexts. Our findings provide nuanced evidence of human-space interactions that go beyond the administrative boundaries and further shape the formation and definition of communities. It is beneficial for place-based planning and developing community initiatives to diminish segregation and facilitate connectivity within and across communities for a more integrated and harmonious society.

CRediT authorship contribution statement

Xiao Huang: Conceptualization, Methodology, Formal analysis, Resources, Data curation, Writing – original draft, Supervision, Project administration. **Yuhui Zhao:** Methodology, Formal analysis. **Siqin Wang:** Writing – original draft, Writing – review & editing. **Xiao Li:** Writing – original draft, Writing – review & editing, Investigation. **Di Yang:** Validation, Data curation, Supervision. **Yu Feng:** Methodology, Validation, Writing – review & editing. **Yang Xu:** Formal analysis, Conceptualization, Methodology. **Liao Zhu:** Conceptualization, Methodology, Visualization. **Biyu Chen:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

The Vice Chancellor for Research and Innovation Fund from the University of Arkansas.

(continued on next page)

Table A (continued)

Full names	Short names
New Jersey	NJ
New York	NY
Pennsylvania	PA
Texas	TX
Virginia	VA
Washington	WA
West Virginia	WV
Wisconsin	WI

References

- Alizadehtazi, B., Tangtrakul, K., Woerdeman, S., Gussenoven, A., Mostafavi, N., & Montalto, F. A. (2020). Urban Park usage during the COVID-19 pandemic. *Journal of Extreme Events*, 7(04), 2150008.
- An, R., & Xiang, X. (2015). Social vulnerability and leisure-time physical inactivity among US adults. *American Journal of Health Behavior*, 39(6), 751–760.
- Bae, S. H., Halperin, D., West, J., Rosvall, M., & Howe, B. (2013). Scalable flow-based community detection for large-scale network analysis. December. In *2013 IEEE 13th international conference on data mining workshops* (pp. 303–310). IEEE.
- Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939), 412–413.
- Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, 50(4), 495–510.
- Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19(6), 275.
- Blumenstock, J., & Fratamico, L. (2013). Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data. December. In *Proceedings of the 4th annual symposium on computing for development* (pp. 1–10).
- Bora, N., Chang, Y. H., & Maheswaran, R. (2014). Mobility patterns and user dynamics in racially segregated geographies of US cities. December. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 11–18). Cham: Springer.
- Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075), 462–465.
- Bullard, R. D., Johnson, G. S., & Torres, A. O. (1999). *Sprawl Atlanta: Social equity dimensions of uneven growth and development*. Atlanta, GA: Clark Atlanta University, The Environmental Justice Resource Center.
- Bureau, U. S. C. (2020). *Metropolitan and micropolitan statistical areas population totals and components of change: 2010–2019*. <https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-metro-and-micro-statistical-areas.html>. (Accessed 22 October 2020).
- Catlin-Groves, C. L. (2012). The citizen science landscape: From volunteers to citizen sensors and beyond. *International Journal of Zoology*, 2012.
- Centers for Disease Control and Prevention. (2021). *CDC SVI documentation 2018*. June 22. Centers for Disease Control and Prevention https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html.
- Cramer, A., Cucarere, J., Tran, M., Lu, A., & Reddy, A. (2009). Performance measurements on mass transit: Case study of New York City transit authority. *Transportation Research Record*, 2111(1), 125–138.
- Curranini, S., Jackson, M. O., & Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4), 1003–1045.
- De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011). Generalized louvain method for community detection in large networks. November. In *2011 11th international conference on intelligent systems design and applications* (pp. 88–93). IEEE.
- Deauna, J. D. L., Yatco, K. M. B., Villanoy, C. L., & Juinio-Meñez, M. A. (2021). Identification of priority sites to support management of commercially important sea cucumber species by applying infomap and habitat filters to larval dispersal data. In *Frontiers in Marine Science*.
- Farage, C., Edler, D., Eklöf, A., Rosvall, M., & Pilosof, S. (2021). Identifying flow modules in ecological networks using infomap. *Methods in Ecology and Evolution*, 12(5), 778–786.
- Ferguson, N. (2017). The false prophecy of hyperconnection: How to survive the networked age. *Foreign Affairs*, 96(5), 68–79.
- Galiana, L., Sakarovich, B., & Smoreda, Z. (2018). *Understanding socio-spatial segregation in French cities with mobile phone data*.
- Glaeser, E., & Vigdor, J. (2012). *The end of the segregated century*. Manhattan Institute for Policy Research. January, 23–26.
- Heine, C., Marquez, C., Santi, P., Sundberg, M., Nordfors, M., & Ratti, C. (2021). Analysis of mobility homophily in Stockholm based on social network data. *Plos one*, 16(3), Article e0247996.
- Hilbert, M. (2021). Information theory for human and social processes. *Entropy*, 23(1), 9. <https://doi.org/10.3390/e23010009>
- Hong, Y., & Yao, Y. (2019). Hierarchical community detection and functional area identification with OSM roads and complex graph theory. *International Journal of Geographical Information Science*, 33(8), 1569–1587.
- Horse, A. J. Y., Parkhurst, N. A. D., & Huyser, K. R. (2020). COVID-19 in New Mexico tribal lands: Understanding the role of social vulnerabilities and historical racisms. *FrontiersSociology*, 5.
- Huang, X., Li, Z., Lu, J., Wang, S., Wei, H., & Chen, B. (2020). Time-series clustering for home dwell time during COVID-19: What can we learn from it? *ISPRS International Journal of Geo-Information*, 9(11), 675.
- Huang, X., Lu, J., Gao, S., Wang, S., Liu, Z., & Wei, H. (2022). Staying at home is a privilege: Evidence from fine-grained Mobile phone location data in the United States during the COVID-19 pandemic. *Annals of the American Association of Geographers*. <https://doi.org/10.1080/24694452.2021.1904819>
- Jiang, B., Yin, J., & Zhao, S. (2009). Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2), Article 021136.
- Jiang, Y., Li, Z., & Cutter, S. L. (2021). Social distance integrated gravity model for evacuation destination choice. *International Journal of Digital Earth*, 1–15.
- Jones, K. K., Anderko, L., & Davies-Cole, J. (2020). Neighborhood environment and asthma exacerbation in WashingtonDC. *Annual review of nursing research*, 38(1), 53–72.
- Kang, Y., Gao, S., Liang, Y., Li, M., Rao, J., & Kruse, J. (2020). Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic. *Scientific Data*, 7(1), 1–13.
- Karatäş, A., & Şahin, S. (2018). Application areas of community detection: A review. In *2018 International congress on big data, deep learning and fighting cyber terrorism (IBIGELFT)* (pp. 65–70). IEEE.
- Khan, B. S., & Niazi, M. A. (2017). *Network community detection: A review and visual survey*. arXiv preprint. arXiv:1708.00977.
- Kramer, R., & Kramer, P. (2019). Diversifying but not integrating: Entropic measures of local segregation. *Tijdschrift voor Economische en Sociale Geografie*, 110(3), 251–270.
- Kwan, M. P. (1999). Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic Geography*, 75(4), 370–394.
- Lenormand, M., Samaniego, H., Chaves, J. C., da Fonseca Vieira, V., da Silva, M. A. H. B., & Eusékoff, A. G. (2020). Entropy as a measure of attractiveness and socioeconomic complexity in Rio de Janeiro metropolitan area. *Entropy*, 22(3), 368.
- Li, Z., Huang, X., Ye, X., Jiang, Y., Yago, M., Ning, H., Li, X., ... (2021). Measuring global multi-scale place connectivity using geotagged social media data. *Scientific Reports*, 11, 14694. <https://doi.org/10.1038/s41598-021-94300-7>
- Lieb, E. H., & Yngvason, J. (1998). A guide to entropy and the second law of thermodynamics. In *Statistical mechanics* (pp. 353–363). Berlin, Heidelberg: Springer.
- Limtanakool, N., Dijst, M., & Schwanen, T. (2006). The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium- and longer-distance trips. *Journal of Transport Geography*, 14(5), 327–341.
- Lu, F., Liu, K., Duan, Y., Cheng, S., & Du, F. (2018). Modeling the heterogeneous traffic correlations in urban road systems using traffic-enhanced community detection approach. *Physica A: Statistical Mechanics and Its Applications*, 501, 227–237.
- Massey, D. S., & Denton, N. A. (1987). Trends in the residential segregation of Blacks, Hispanics, and Asians: 1970–1980. *American Sociological Review*, 802–825.
- McCrea, R. (2009). Explaining sociospatial patterns in south East Queensland, Australia: Social homophily versus structural homophily. *Environment and Planning A*, 41(9), 2201–2214.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444.
- Mollaca, K. A., Gray, B., & Trevino, L. K. (2003). Racial homophily and its persistence in newcomers' social networks. *Organization Science*, 14(2), 123–136.
- Mora, R., & Ruiz-Castillo, J. (2011). Entropy-based segregation indices. *Sociological Methodology*, 41(1), 159–194.
- Morales, A. J., Dong, X., Bar-Yam, Y., & Pentland, A. S. (2019). Segregation and polarization in urban areas. *Royal Society Open Science*, 6(10), Article 190573.
- Mu, L., Wang, F., Chen, V. W., & Wu, X. C. (2015). A place-oriented, mixed-level regionalization method for constructing geographic areas in health data dissemination and analysis. *Annals of the Association of American Geographers*, 105(1), 48–66.
- Murugesan, S. (2007). Understanding web 2.0. *IT professional*, 9(4), 34–41.
- Pinto, T., Morais, H., & Corchado, J. M. (2019). Adaptive entropy-based learning with dynamic artificial neural network. *Neurocomputing*, 338, 432–440.
- Piorkowski, M. (2009). Sampling urban mobility through on-line repositories of GPS tracks. December. In *Proceedings of the 1st ACM international workshop on hot topics of planet-scale mobility measurements* (pp. 1–6).
- Poorthuis, A. (2018). How to draw a neighborhood? The potential of big data, regionalization, and community detection for understanding the heterogeneous nature of urban neighborhoods. *Geographical Analysis*, 50(2), 182–203.

- Proops, J. L. (1987). Entropy, information and confusion in the social sciences. *Journal of Interdisciplinary Economics*, 1(4), 225–242.
- Reardon, S. F., & Bischoff, K. (2011). Income inequality and income segregation. *American Journal of Sociology*, 116(4), 1092–1153.
- Reardon, S. F., & Firebaugh, G. (2002). Measures of multigroup segregation. *Sociological Methodology*, 32(1), 33–67.
- Reardon, S. F., & O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology*, 34(1), 121–162.
- Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, 178(1), 13–23.
- Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), 1118–1123.
- SafeGraph. (2019). *What about bias in the SafeGraph dataset?*. <https://www.safegraph.com/blog/what-about-bias-in-the-safegraph-dataset>.
- SafeGraph. (2020). *Social Distancing Metrics*. December <https://docs.safegraph.com/docs/socialdistancing-metrics>.
- Sethi, N., & Sharma, D. (2012). A new cryptology approach for image encryption. December. In *2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing* (pp. 905–908). IEEE.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, 142, 198–211.
- Silverman, R. M. (2005). Community socioeconomic status and disparities in mortgage lending: An analysis of metropolitan Detroit. *The Social Science Journal*, 42(3), 479–486.
- Theil, H., & Finizza, A. J. (1971). A note on the measurement of racial integration of schools by means of informational concepts. *The Journal of Mathematical Sociology*, 1 (2), 187–193.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1), 234–240.
- Tolnay, S. E. (2003). The african american “great migration” and beyond. *Annual Review of Sociology*, 29(1), 209–232.
- Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1), 1–12.
- Vanhoof, M., Schoors, W., Van Rompaey, A., Ploetz, T., & Smoreda, Z. (2018). Comparing regional patterns of individual movement using corrected mobility entropy. *Journal of Urban Technology*, 25(2), 27–61.
- Wang, C., Tang, W., Sun, B., Fang, J., & Wang, Y. (2015). Review on community detection algorithms in social networks. December. In *2015 IEEE international conference on progress in informatics and computing (PIC)* (pp. 551–555). IEEE.
- Wang, Q., Phillips, N. E., Small, M. L., & Sampson, R. J. (2018). Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences*, 115(30), 7735–7740.
- Wyczalkowski, C. K., Welch, T., & Pasha, O. (2020). Inequities of transit access: The case of Atlanta. *GA. JCULP*, 4, 654.
- Xu, Y., Belyi, A., Bojic, I., & Ratti, C. (2018). Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems*, 72, 51–67.
- Xu, Y., Santi, P., & Ratti, C. (2021). Beyond distance decay: Discover homophily in spatially embedded social networks. *Annals of the American Association of Geographers*. <https://doi.org/10.1080/24694452.2021.1935208>
- Yildirimoglu, M., & Kim, J. (2018). Identification of communities in urban mobility networks using multi-layer graphs of network traffic. *Transportation Research Part C: Emerging Technologies*, 89, 254–267.
- Zambon, I., Serra, P., Grigoriadis, E., Carlucci, M., & Salvati, L. (2017). Emerging urban centrality: An entropy-based indicator of polycentric development and economic growth. *Land Use Policy*, 68, 365–371.
- Zhao, K., Musolesi, M., Hui, P., Rao, W., & Tarkoma, S. (2015). Explaining the power-law distribution of human mobility through transportationmodality decomposition. *Scientific Reports*, 5(1), 1–7.
- Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28(11), 2178–2199.