PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

Urban delineation through the lens of commute networks: Leveraging graph embeddings to distinguish socioeconomic groups in cities

Devashish Khulbe[1*], **Stanislav Sobolevsky**[2,3,4]

1 Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno, Czech Republic, **2** Center for Urban Science+Progress, New York University, Brooklyn, New York, United States of America, **3** Center for Interacting Urban Networks, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates, **4** Institute of Computer Science, Masaryk University, Brno, Czech Republic

* dk3596@nyu.edu

OPEN ACCESS

Citation: Khulbe D, Sobolevsky S (2025) Urban delineation through the lens of commute networks: Leveraging graph embeddings to distinguish socioeconomic groups in cities. PLOS Complex Syst 2(8): e0000061. https://doi.org/10.1371/journal.pcsy.0000061

Editor: Haroldo V. Ribeiro, Universidade Estadual de Maringa, BRAZIL

Received: March 31, 2025

Accepted: July 15, 2025

Published: August 11, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000061

Data availability statement: The data used in this study can be found in a Zenodo repository:

Abstract

Delineating areas within metropolitan regions stands as an important focus among urban researchers, shedding light on the urban perimeters shaped by evolving population dynamics. Applications to urban science are numerous, from facilitating comparisons between delineated districts and administrative divisions to informing policymakers of the shifting economic and labor landscapes. In this study, we propose using commute networks sourced from the census for the purpose of urban delineation, by modeling them with a Graph Neural Network (GNN) architecture. We derive low-dimensional representations of granular urban areas (nodes) using GNNs. Subsequently, nodes' embeddings are clustered to identify spatially cohesive communities in urban areas. Our experiments across the U.S. demonstrate the effectiveness of network embeddings in capturing significant socioeconomic disparities between communities in various cities, particularly in factors such as median household income. The role of census mobility data in regional delineation is also noted, and we establish the utility of GNNs in urban community detection, as a powerful alternative to existing methods in this domain. The results offer insights into the wider effects of commute networks and their use in building meaningful representations of urban regions.

Author summary

Understanding how cities are structured is important for both policymakers and researchers, as it guides decisions about public services, transportation, and economic development. Traditional ways of dividing a city into multiple regions often rely on official boundaries, which typically remain static over time. However, population dynamics and demands in urban areas constantly change, highlighting the need for a more flexible

https://doi.org/10.5281/zenodo.11494208. Specifically, the data contains origin-destination daily commute flow information among the census tracts in all 12 U.S. cities considered in this work.

Funding: This research was supported by the MUNI Award in Science and Humanities (MASH Belarus) of the Grant Agency of Masaryk University under the Digital City project (MUNI/J/0008/2021 to SS). This work was also partially supported by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award (CG001 to SS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and data-driven approach to defining urban boundaries. In this work, we introduce a new method that analyzes commuting patterns within cities using data from the U.S. Census, which records where people live and where they work. We represent this information as a network, where each area is connected based on the flow of commuters, and then apply deep learning techniques to model these networks. By clustering the learned representations of urban areas, we can identify groups of neighborhoods that share common commuting behaviors. When we compare these groups with information on household income, we observe meaningful differences, such as the separation of wealthier areas from lower-income ones. Our method provides a new way to delineate regions within cities using widely available data and may help reveal how commuting patterns reflect broader social and economic divides.

Introduction

Urban delineation plays a crucial role in addressing a wide range of questions in urban science. City administrators rely on well-defined boundaries to allocate resources effectively and plan essential services, while urban scientists and researchers analyze evolving urban boundaries to gain insights into shifting demographic patterns, spatial dynamics, and the changing nature of urban environments. Recent research has concentrated on discerning business hubs, regional administrative precincts, and even delineations of entire metropolitan areas.

Dynamic network data has long been proposed by researchers for this purpose. Cell-phone records have proved quite effective in regional delineation, modeling on GPS trajectories and origin-destination movement networks [1,2], while others have also considered urban movement derived from social media interactions [3], where larger regions have also been delineated on country level [4]. However, a data source that is widely available in cities is not yet widely known. We propose using mobility networks from Origin-Destination commute data available from the census. Such a dataset offers comprehensive coverage, facilitating studies throughout the country, and eliminating the reliance on alternative, potentially biased mobility data providers that may exclude significant portions of the population. Census data boasts extensive spatial reach, enables sophisticated network analysis, and constitutes a vast repository covering the majority of the country's inhabitants.

Network partitioning has been tackled with methods such as community detection [5], kernel density estimation [6], and partitioning based on spatial constraints [3], which are generally based on optimizing a network metric such as modularity. While self-supervised deep learning based network embeddings have not been evaluated for detecting communities, they have been widely successful in many downstream tasks with urban networks. With mobility networks, GNN based modeling have shown to be successful in socioeconomic modeling [7], while various GNN model architectures have been employed for numerous tasks like edge prediction and community detection in urban graphs [8,9]. They have also been useful in heterogeneous urban networks like population-facilities interaction graphs [10]. Using census data-derived commute networks as input, we train a GNN model to obtain low-dimensional embedding vectors of census tracts (nodes) in urban networks. The model is learned in a self-supervised fashion with the objective to reconstruct the original commute flows (edges) in the network. To obtain communities, we further cluster the embedding vectors and notice spatially homogeneous communities in all metro areas. Next, we evaluate the socioeconomic profiles of the clusters, and observe varying differences in income status among communities within cities. The income status is directly related to commute ability and flexibility, which is revealed by the changing community structure.

While GNN-based embeddings tend to yield more homogeneous and coherent community structures, graph clustering and community detection have long been active areas of research. Notably, graph embeddings have also been proposed as a powerful approach for community detection [11], often demonstrating performance comparable to traditional methods. However, such techniques have rarely been evaluated in the context of urban mobility networks, where the spatial and socioeconomic dimensions introduce additional complexity. Furthermore, it remains important to assess these methods across real-world networks of varying scales to understand their generalizability and robustness. Also, it is crucial to have some sort of comparative analysis to existing methods, perhaps with a network-based metric commonly used for community detection evaluation purposes. Therefore, we also compare our results with two widely used community detection approaches–1. A network modularity optimization-based method, and 2. A probabilistic generative Stochastic Block Model.

Our main findings can be summarized in two key points:

1. We show the effectiveness of general-purpose GNN-based network embeddings in enabling efficient and socioeconomically meaningful urban delineation, and find them to be at least as good, and in some cities, to be better than traditional community detection methods.
2. We demonstrate the practical value of census-derived mobility data for urban community delineation, highlighting its accessibility and relevance in diverse geographic contexts.

We present our results for the 12 largest U.S. metropolitan areas, with varying numbers of delineations for each area. We compare the results with community detection methods, and also present a qualitative evaluation of differences in communities from a socioeconomic perspective. Notably, we observe delineation of high-income density neighborhoods from low-income ones. The distinction is interesting, as it reveals commute patterns and gaps, which may be exacerbated by commuters' income and other socioeconomic factors.

Data overview

Diverse data have been proposed for the purpose of urban and regional delineation. Mobile phone data has been used in many recent studies to forge mobility networks for various downstream tasks [12]. Mobile phone data is good for highly granular analysis, but not consistently available everywhere. Other studies have also frequently used the social media footprint of people across urban areas and POIs [3]. In general, mobility datasets have been widely used for modeling and as a key feature for urban modeling, such as socioeconomic analysis and regional delineation[4,13]. Phone-based mobility and social media data may provide more dynamic records, but comes with the condition of not being comprehensive across locations and populations. In this study, we thus propose census-derived mobility data for networks in cities. Census data has long been a valuable resource for researchers, widely utilized in studies ranging from population dynamics to urban sprawl, land use, and development [14,15]. Although census-derived mobility data remains relatively underexplored, prior research has demonstrated the effectiveness of census-based variables—such as income and unemployment rates—as indicators of socioeconomic status [16,17].

For the U.S. cities considered in our analysis, we retrieve the Origin-Destination (O-D) commute data from the Longitudinal Employer-Household Dynamics (LEHD) [18]. Leveraging administrative records from the U.S. Census Bureau, LEHD provides a comprehensive view of worker flows in all cities across the U.S., and thus offers rich insights into the

dynamics of labor markets and commuting patterns. Populating the network with LEHD commute flows hence provides a comprehensive picture of mobility across cities. We also use U.S. census data to retrieve income data for all cities considered. Median income is considered a proxy to evaluate the socioeconomic profile of areas in cities.

Table 1 shows the population and commute network statistics for 12 U.S. metro regions considered in the experiments. The biggest metro areas in the country spans across every geography and covers a total of approximately 30 million people. Using census data here is key, as it is consistently available across the nation spanning across every spatial granularity. This makes our experiments possible across every major urban area in the country. Thus, the accessibility and comprehensiveness of the mobility information from the census is hard to match by other sources.

Methods

The underlying idea behind delineation in urban networks is of finding similar nodes in graphs based on network properties. There have been many methods proposed in the literature for this purpose. For comparison with our methods, we will primarily look into the results from community detection methods, which have been widely successful in urban networks in recent years [19,20].

Our approach mainly work by deriving the embedding of mobility networks. Mobility is a crucial urban metric and is the core ability under which many lifestyle choices depend of people. Some recent studies have also found that mobility is vastly impacted by commuters' income status [21,22]. Thus, embedding as a low-dimensional representation of the larger mobility network can be particularly useful for modeling socioeconomic indicators of regions, including income.

Our methods derive the idea from using positional encoding in models, with recent models with graph modeling with transformer-based models [23]. We propose a self-supervised learning framework, with GNN being the model architecture for network information propagation. We also present a MLP-based model, as a deep-learning baseline model to compare how GNN-based embedding compares to a standard MLP-based representation learning methodology. Thus, we first derive low-dimensional representation vectors (embedding) for the commute network i.e. $N \times d$ matrix for N spatial units (nodes), d being the embedding dimensionality. Then we cluster the embedding vectors using the K-means method to get communities in the urban area.

Table 1. Commute network statistics for 12 U.S. cities.

City	Nodes	Non-zero Edges	Avg. edge weight
New York	2157	976832	0.69
Chicago	1318	439553	1.06
Boston	520	127357	3.5
Austin	218	34777	8.63
Dallas	529	129352	2.83
Los Angeles	2341	1171362	0.65
San Antonio	366	83192	4.77
San Diego	627	180781	2.97
San Jose	372	81938	4.73
Philadelphia	384	68119	2.57
Phoenix	916	349894	2.10
Houston	786	290496	2.50

https://doi.org/10.1371/journal.pcsy.0000061.t001

Positional and Structural Encodings (PSE)

We derive the utility of encoding methods to capture positional and structural information in networks. Laplacian Eigenvectors (LE) has been used as positional encodings of nodes in recent models on Graph Transformers (GT) [24]. Other methods include shortest-path distances [25], and SVD [26], which has also been used in Graph Representation learning tasks [27]. SVD graph embedding has also proven to be useful in community detection, with promising results for unipartite and bipartite graphs [28]. Structural encodings have been developed to capture rich local and global connectivity patterns in networks. Random Walk encoding [29], has proven to be quite useful in capturing structural information in graphs [30]. They also have been successfully used as structural encoding in GNN based models [31,32]. Recent works on graph-based models (GNNs and Graph Transformers) use many of these existing embedding methods as positional inputs to the models. In urban mobility networks, network encodings have been proven to be useful for downstream tasks using heterogeneous networks [33,34].

Hence, the underlying idea behind experimenting with PSE is the evaluation of delineation capabilities of the encodings, and whether the resulting urban communities are spatially homogeneous. The presumption with using PSE is also that any modeling involves significantly larger parameters and training, whereas PSE are relatively easier and trivial to obtain. Additionally, we also evaluate the capabilities of PSE in discerning neighborhoods based on their socioeconomic profiles.

Representation learning

Graph representation learning aims to generate low-dimensional embeddings for each node (or sub-region in our context), from which the original graph structure can be reconstructed or downstream tasks (e.g., community detection, node classification) can be performed. Deep neural networks, particularly GNNs, have been highly successful in capturing meaningful graph representations by leveraging graph connectivity and neighborhood structure. Below, we outline our approaches for node embedding—focusing on our GNN-based method—along with a brief overview of a baseline MLP-based approach.

GNN-based embedding. Graph Neural Networks (GNNs) extend conventional neural networks to graph-structured data, enabling node (or sub-region) embeddings to be learned by aggregating and transforming information from local neighborhoods. In our setup, we utilize a two-layer Graph Neural Network (GNN) architecture, which has been shown to be effective for representation learning across a wide range of graph-related tasks [35,36]. Two-layer GNNs are particularly well-suited for capturing local structural information while maintaining computational efficiency. Moreover, increasing the number of layers in GNNs often leads to the oversmoothing phenomenon, where node representations become indistinguishable, ultimately degrading model performance [37,38].

Architecture. A two-layer GNN can be viewed as:

$$H^l = \sigma\left(W^l \hat{A}\, H^{l-1} + B^l\right), \tag{1}$$

where:

- H^{l-1} is the node embedding matrix at layer $(l–1)$,
- $\hat{A}$ is the adjacency matrix (normalized and containing self-loops),
- W^l and B^l are learnable weight and bias terms, and
- $\sigma(\cdot)$ is a non-linear activation (ReLU).

Each node's representation is updated by aggregating information from its neighbors. A more explicit illustration of the update rule at the node level can be written as:

$$h_i^l = \sigma\left(\sum_{j \in N(i)} \frac{1}{|N(i)|} W^l h_j^{l-1}\right), \tag{2}$$

where $N(i)$ is the set of neighbors of node i. This formulation underscores that each node embedding h_i^l is derived from the aggregation of its neighbors' embeddings at the previous layer.

Normalization. In practice, we use degree-based normalization to stabilize training and to ensure consistent scaling across different parts of the graph. This is particularly beneficial when sub-regions (i.e., nodes) vary significantly in their degrees.

Training objective. We train the GNN in a self-supervised manner to reconstruct the adjacency matrix A of the mobility network. After the second GNN layer, we obtain the final node embedding matrix H^2. We then apply a suitable decoder as a reconstruction function (additional feedforward layer) to predict $\hat{A}$. The reconstruction loss, Mean Squared Error (MSE), is:

$$\mathcal{L} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(A_{ij} - \hat{A}_{ij}\right)^2. \tag{3}$$

By minimizing this loss, the GNN learns node embeddings that capture both global and local connectivity patterns in the graph.

MLP-derived embedding (baseline). For comparison, we also train a simpler Multi-Layer Perceptron (MLP)—referred to here as a Vanilla Neural Network (VNN)—to learn node embeddings that can reconstruct the adjacency matrix. This baseline method initializes each node with a d-dimensional trainable embedding and transforms pairwise concatenations of these embeddings through a feedforward network to predict edge values. Formally, the process involves:

1. **Initialization:** Each node i has an embedding $e_i \in \mathbb{R}^d$.
2. **Pairwise Interactions:** For all node pairs (i,j), we concatenate learnable embedding vectors e_i and e_j to form an input feature vector.
3. **MLP:** A 3-layer feedforward network processes each pairwise feature vector to yield $\hat{A}_{ij}$.
4. **Training:** MSE loss is minimized on the adjacency reconstruction, $\sum_{i,j}(A_{ij} - \hat{A}_{ij})^2$.

While this MLP-based baseline can learn embeddings by directly modeling pairwise relationships, it does not leverage graph convolution or neighbor aggregation, potentially limiting its ability to capture higher-order structural properties in the mobility network. Nonetheless, it serves as a computationally straightforward point of comparison and can sometimes provide surprising effectiveness, especially when the graph is not too large or has strong pairwise signals.

In the following sections, we present experimental results comparing these approaches in terms of community detection and further socioeconomic analysis of the communities. Our primary focus lies in understanding how well graph-based embeddings capture the network communities relative to community detection methods, and also how these communities differ from each other. We evaluate the resulting communities with PSE, Graph embeddings, and community detection methods in 12 cities under consideration. The PSE for all mobility networks were retrieved choosing appropriate embedding dimensionality for different metro

regions. Graph representation learning models (VNN, GNN) were trained for 500 epochs with logMSE objective in reconstructing their respective adjacency matrices. The community detection method was run multiple times to ensure stability.

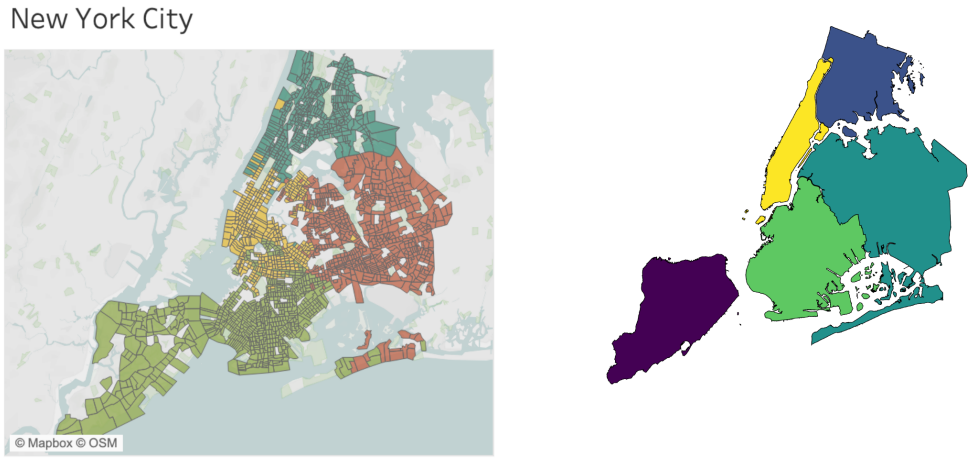
Results and discussion

Community structures

Borough/county comparison with clusters. One challenge in evaluating the detected communities is the absence of a universally recognized "ground truth" for urban delineation. Nevertheless, in some major cities such as New York–comparisons can be drawn against well-established administrative units like boroughs or counties within the city. Since our analysis is performed at the census tract level, we can assess how closely the discovered communities align with these larger administrative boundaries, providing a meaningful benchmark for validating the coherence of our results.

An inspection of the GNN-embedding derived communities in Fig 1 reveals that community clusters do not strictly follow the official borough boundaries. For instance, certain areas of southwestern Brooklyn are grouped together with Staten Island, forming a single community despite belonging to different boroughs. Likewise, Manhattan appears partitioned into multiple clusters, indicating diverse mobility and interaction patterns within what is officially one borough. Queens also shows distinct subdivisions rather than forming a single unified cluster. These overlaps and divergences underscore that the detected communities are driven more by patterns in the underlying data (e.g., mobility interaction strengths) than by administrative lines, highlighting both the utility and the potential complexity of such data-driven delineations.

In the subsequent section, we therefore investigate how these communities diverge across key indicators, focusing particularly on socioeconomic metrics, to illuminate the underlying factors that shape these data-driven partitions.



(a) GNN-based embedding communities (b) NYC's official borough boundaries

Fig 1. Comparison of GNN communities and borough boundaries in NYC (*Basemap in (a) based on OpenStreetMap data, ©OpenStreetMap contributors (https://www.openstreetmap.org), licensed under the Open Data Commons Open Database License (ODbL) 1.0 (https://opendatacommons.org/licenses/odbl/1.0/).*)

https://doi.org/10.1371/journal.pcsy.0000061.g001

Comparisons with community detection methods. Community detection methods have been commonly used to delineate districts and find urban communities in a variety of networks. Since there is no "ground-truth" as such in terms of urban areas, we can still compare the communities derived from network embeddings to commonly used community detection methods from the literature.

Traditionally, community detection methods have focused on optimizing modularity, a heuristic measure that quantifies the density of links inside communities compared to links between them [39]. More recent methods have addressed community detection from a principled statistical perspective using Stochastic Block Models (SBMs). SBMs are generative models that assume the observed network structure arises from an underlying block structure, where nodes are assigned to latent groups (or blocks), and edges are generated according to probabilities that depend solely on the group memberships of the nodes [40]. In practice, inference is performed by minimizing the description length of the model using Bayesian model selection techniques, such as the Minimum Description Length (MDL) principle.

Fig 2 also shows a comparison with two distinct and well-established approaches for community detection in networks. We evaluate and compare the communities obtained from 1. COMBO method [41], a well-established benchmark for community detection in networks based on modularity optimization, and 2. An SBM model based on [42] with Monte Carlo optimization. In all cities in our analysis, we notice GNN-embedding based communities closely align with those we get from COMBO method. This is also quantitatively shown by evaluating the modularity for communities from both methods. The community delineations produced by the SBM approach differ notably from those of other methods, often resulting in smaller and spatially incohesive communities in most cities. In New York City, for example, while SBM successfully identified neighborhoods in Manhattan and the Bronx (similar to GNN and COMBO), it also isolated a single census tract as its own community, highlighting its tendency toward fragmented partitions. In Chicago, SBM's results diverged more significantly, yielding scattered and irregularly shaped communities that contrast sharply with the more coherent patterns detected by GNN and COMBO. A similar trend was observed in Boston and Chicago, where the communities derived from SBM were consistently smaller and structurally different from those produced by the other two methods. Interestingly, for Austin and San Antonio, SBM assigns all nodes of the network to a single community. This contrasts GNN and COMBO methods, which have 3 cohesive communities for both cities.

Table 2 presents the modularity scores for the communities identified in each city. Across the 12 cities analyzed, we find that communities derived from GNN-based embeddings achieve modularity scores comparable to those produced by COMBO. In contrast, the spatial fragmentation of SBM-derived communities results in substantially lower modularity scores. Similarly, communities obtained from MLP and PSE-based embeddings also lack spatial coherence and correspondingly exhibit significantly reduced modularity.

Socioeconomic evaluation

To further investigate inter-cluster differences, we analyze the distribution of a key socioeconomic indicator–median neighborhood income–across the identified communities. Median income is a widely recognized proxy for residents' socioeconomic status and provides a meaningful basis for comparison. Our analysis shows that the communities delineated using network embeddings—particularly those generated by GNN-based methods—exhibit distinct income profiles. As shown in Fig 3, we analyze the income distributions of census tracts and examine the resulting disparities across all communities. While not all communities are

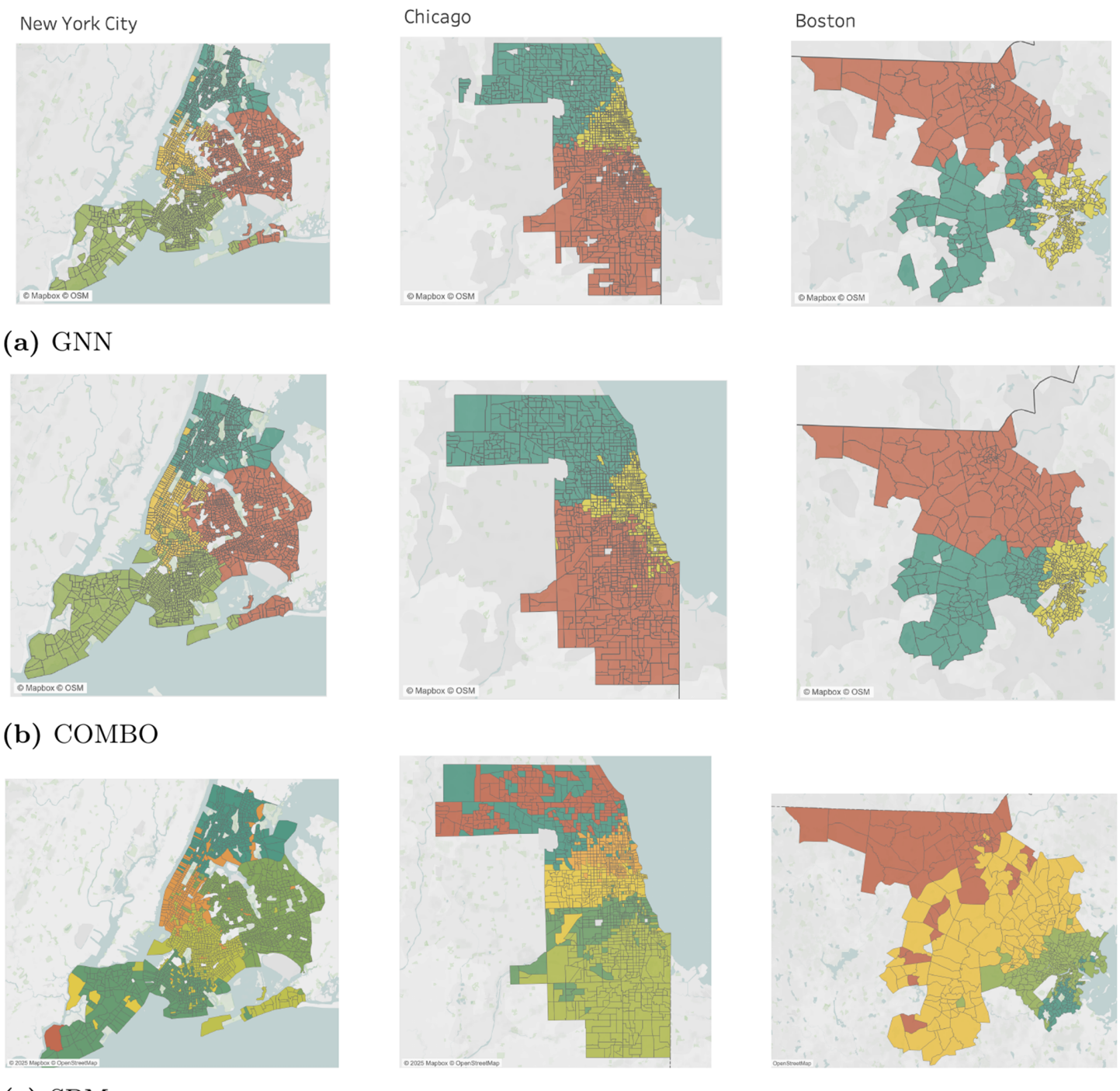


Fig 2. Community structures resulting from a) GNN-embedding based communities, b) COMBO method [41], and c) Stochastic Block Model (SBM). The GNN-derived community boundaries are similarly aligned to modularity optimization based COMBO. Notably, SBM's results have more communities in all cities. Moreover, SBM-derived community structures are observed to be smaller and not spatially cohesive. (*Basemaps based on OpenStreetMap data, ©OpenStreetMap contributors (https://www.openstreetmap.org), licensed under the Open Data Commons Open Database License (ODbL) 1.0 (https://opendatacommons.org/licenses/odbl/1.0/).*)

https://doi.org/10.1371/journal.pcsy.0000061.g002

Table 2. Modularity scores for communities resulting from different embeddings compared with COMBO-based network partitions.

City	Number of communities	COMBO	SBM	GNN	VNN
New York	4	0.27	0.16	0.259	0.12
Chicago	3	0.275	0.13	0.264	0.10
Boston	3	0.272	0.14	0.259	0.16
Austin	3	0.162	0.0	0.156	0.07
Dallas	3	0.187	0.08	0.187	0.05
Los Angeles	4	0.354	0.19	0.351	0.18
San Antonio	3	0.169	0.0	0.168	0.07
San Diego	4	0.30	0.18	0.29	0.17
San Jose	4	0.196	0.10	0.188	0.11
Philadelphia	4	0.185	0.12	0.179	0.09
Phoenix	3	0.24	0.14	0.24	0.13
Houston	3	0.25	0.18	0.25	0.07

https://doi.org/10.1371/journal.pcsy.0000061.t002

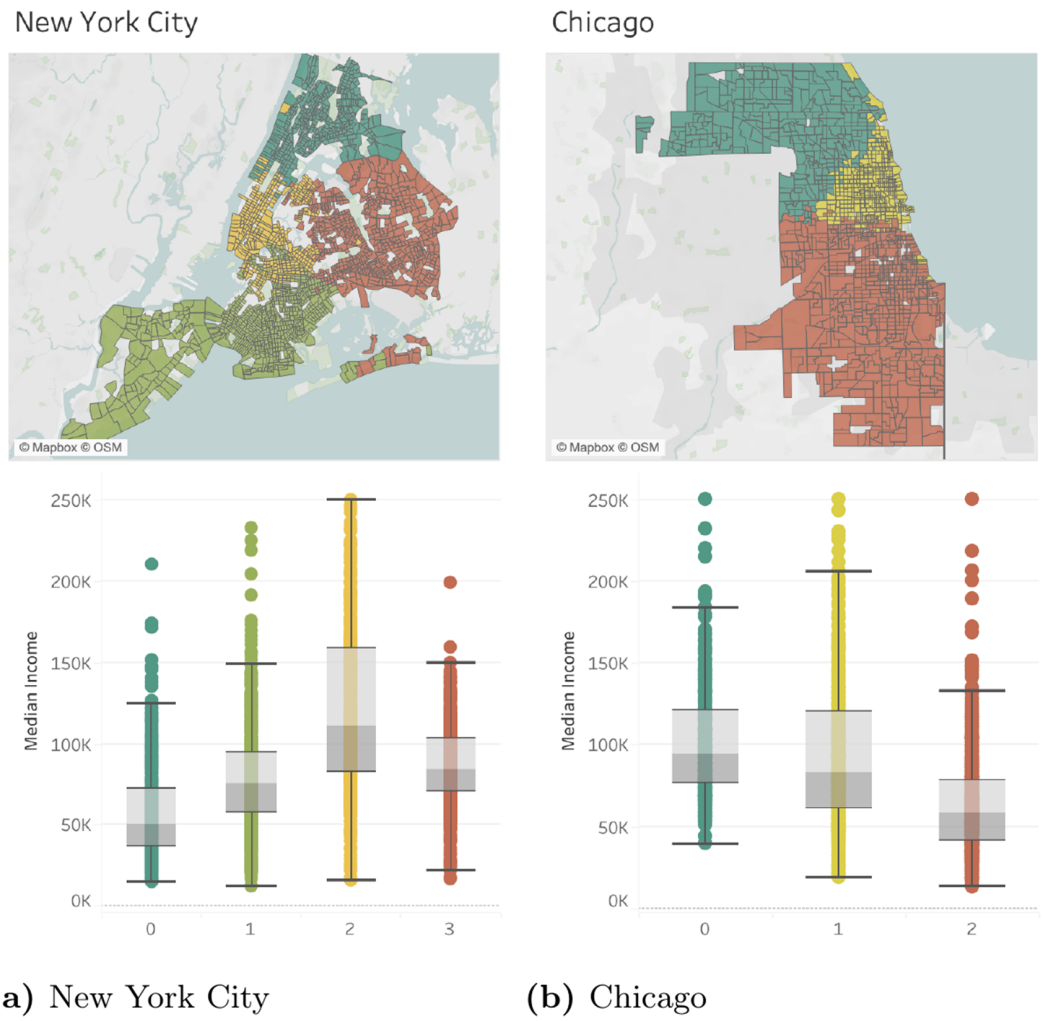


Fig 3. GNN-embedding based communities—shown along with income distributions within each community. (*Basemaps based on OpenStreetMap data, ©OpenStreetMap contributors (https://www.openstreetmap.org), licensed under the Open Data Commons Open Database License (ODbL) 1.0 (https://opendatacommons.org/licenses/odbl/1.0/).*)

https://doi.org/10.1371/journal.pcsy.0000061.g003

sharply differentiated, there is a clear separation between high- and low-income areas, with many of them forming distinct and internally consistent clusters.

Across several focal cities, community-detection results align closely with the spatial distribution of income density. As shown in Fig 4 in New York, for instance, the core of Manhattan emerges as one or two distinct high-income communities that sharply contrast with outlying, lower-income clusters. In Chicago, higher-income density is concentrated primarily along the northern corridor, with that region split between two overlapping communities–both capturing slightly different slices of affluence. Some cities reveals a larger, multi-community overlap surrounding high-income enclaves. Boston's high-income center is more tightly contained; however it is spanned by two communities enclosing the downtown area. By contrast, in Houston and Phoenix, wealthier zones sprawl outward from the central core. Dallas also exhibits a substantial high-income cluster largely confined to one community, but with a few additional pockets identified by neighboring communities. These individual cases demonstrate how community detection can trace the contours of both concentrated and dispersed affluence across different urban environments. Since our analysis is based on commute mobility data, the spatial clustering of high-income neighborhoods within a single community indicates that individuals from higher income groups tend to reside and travel within the same localized areas. Notably, a comparison of community sizes reveals that high-income communities are generally smaller across all cities, suggesting that affluent populations exhibit more spatially concentrated and possibly denser mobility patterns. In contrast, low-income communities tend to span larger areas, implying that less affluent individuals are more likely to commute over longer distances. This observation aligns with previous studies that have examined the relationship between income levels and commuting behavior [22,43]. The income differences among communities are quantified by Jensen–Shannon (J–S) divergence score, a symmetric measure of similarity between two distributions. This metric, bounded between 0 and 1, effectively captures the degree of disparity in income distributions across communities. Divergence based metrics, including J-S divergence, have been used in the literature to assess differences in income and other economic factors among the population.[44–46]. It thus provides us with a metric to compare the income distribution among the delineated districts in our analysis. Table 3 presents a comparison of high- versus low-income community distributions—as measured by the J–S divergence. It reveals considerable variation in how sharply these two groups diverge across different cities. In places such as Philadelphia and Austin, the high-income and low-income communities exhibit noticeably distinct distributional profiles, suggesting stark socioeconomic contrasts. By contrast, Chicago and Los Angeles show more moderate divergence values, implying relatively more overlap—or less pronounced differences—between the characteristics of their highest- and lowest-income communities. Cities like New York and Boston fall somewhere in between: although high- and low-income distributions differ substantially, those differences are not as extreme as in Philadelphia, nor as subtle as in LA. In general, these divergences highlight how, in some urban areas, economic disparities manifest as more distinct communities than in others.

Interestingly, when comparing the income profiles of resulting communities from COMBO, we notice that our method achieve at least the same, and in some cities, higher J-S divergence scores. This suggest that communities identified through our embedding-driven framework are more sharply demarcated in terms of socioeconomic attributes, likely due to the method's ability to preserve fine-grained node attribute correlations during representation learning. This is particularly seen for major metro cities like NYC, Chicago, and Los Angeles, which are the three biggest in the country. In case of SBM, we see some higher J-S divergence scores for some cities like New York, Chicago, and Los Angeles. This is because the method

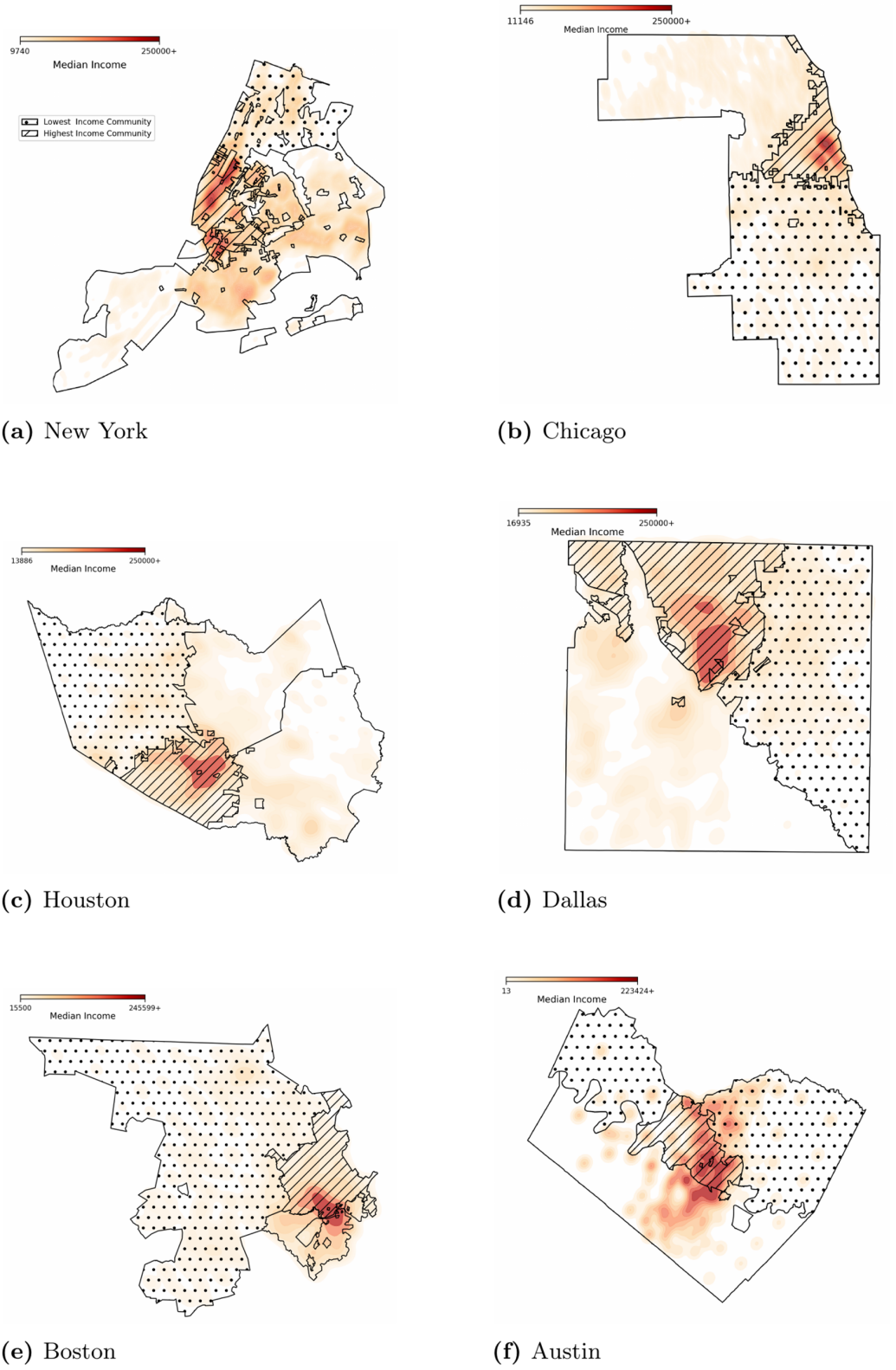


Fig 4. Communities are distinguished by their socioeconomic status—highest income density areas are captured by a community in most cities. Some cities—like Boston, Austin, and Dallas—have high-income areas lying in multiple different communities. Generally, the delineated communities capturing wealthier income groups tend to be smaller compared to other communities in all cities.

https://doi.org/10.1371/journal.pcsy.0000061.g004

Table 3. J-S divergence scores and median income differences between highest and lowest income communities.

City	J-S divergence			Median income delta (USD)
	COMBO	SBM	GNN embeddings	
New York	0.61	0.79	0.62	58,656
Chicago	0.49	0.79	0.54	27,793
Boston	0.65	0.64	0.65	42,839
Austin	0.79	0.0	0.79	67,462
Dallas	0.59	0.54	0.59	28,058
Los Angeles	0.34	0.72	0.38	14,400
San Antonio	0.67	0.0	0.67	30,001
San Diego	0.66	0.65	0.66	36,630
San Jose	0.73	0.71	0.74	46,498
Philadelphia	0.76	0.75	0.76	40,440
Phoenix	0.48	0.43	0.48	20,834
Houston	0.47	0.47	0.47	7,657

https://doi.org/10.1371/journal.pcsy.0000061.t003

delineates some very small, highly affluent neighborhoods as a single community. Thus, this increased socioeconomic differentiation from SBM comes at the expense of generating overly fragmented and spatially scattered communities.

Overall, GNN-derived communities across all cities successfully capture clear distinctions between high- and low-income areas. Interestingly, some cities exhibit a single, contiguous high-income community that encloses their principal area of elevated median income, whereas others reveal two or more communities collectively spanning these high-income clusters. This pattern suggests that while a single "hub" of wealthy neighborhoods dominate in certain urban areas, others have multiple pockets of high-income concentration (for instance in Austin). Importantly, the ability of our approach to isolate these communities–even when they are spatially dispersed–emphasizes its utility in analyzing urban income patterns.

Computational advantage

The adoption of general-purpose graph embeddings for community detection offers substantial computational advantages over traditional methods, particularly when scalability is critical. Conventional approaches, such as the COMBO algorithm, often suffer from high time complexity due to their reliance on iterative pairwise node comparisons or modularity maximization. For instance, executing COMBO on the New York City (NYC) network—a moderately sized graph with around two thousand nodes—requires approximately 4.8 seconds per run on a standard CPU. In contrast, embedding-based methods decouple the computationally intensive graph representation phase from the clustering step. Downstream clustering algorithms (e.g., k-means or spectral clustering) can operate on the embeddings with remarkable efficiency. Clustering the NYC network in the embedded space reduces runtime to just 32 milliseconds, achieving a 150-fold speedup over COMBO. While training a GNN entails an initial computational cost, these embeddings can be reused for multiple analyses without retraining or re-running community detection from scratch. Although raw speedup alone is unlikely to shift practical priorities in urban delineation–where real-time computation is rarely a strict requirement, this framework supports extensibility and efficiency for applied urban analytics, particularly when integrated with other tasks such as prediction, classification, or temporal analysis.

Conclusion

This study demonstrates the effectiveness of mobility network embeddings in uncovering urban community structures and revealing their underlying socioeconomic dynamics. While previously established community detection approaches often have a network metric to optimize, our methods show that a general purpose embedding resulting from a self-supervised trained model can provide meaningful urban communities. The analysis across multiple cities shows that the communities detected via GNN-based embeddings capture meaningful patterns that often can diverge from traditional administrative boundaries. For example, while official delineations in cities like New York suggest clear borough divisions, our results indicate that neighborhoods with similar mobility and interaction patterns may span multiple boroughs, suggesting a more nuanced urban fabric.

When benchmarked against established methods, the network embedding approach achieved comparable modularity scores, underscoring its robustness in delineating spatially cohesive clusters. Furthermore, socioeconomic evaluation reveals that data-driven communities align closely with key indicators such as income distribution. The differences in communities among cities vary, with some cities showing sharp delineation in high-income and low-income density areas, whereas in others the differences are more moderate. The variation in income densities among clusters underscores how economic disparities manifest differently depending on local mobility dynamics. This provides valuable insights into the spatial distribution of socioeconomic status based on where people commute within the city.

Another critical insight from our work is the role of commute networks in shaping these urban communities. The flow of commuters and the structure of transit routes play a significant role in connecting disparate neighborhoods, influencing both the formation of community clusters and the observed socioeconomic patterns. By incorporating commute networks into our analysis, we capture a more comprehensive picture of urban mobility, underscoring their importance in the delineation and evolution of community structures. While prior studies have predominantly relied on dynamic mobility data derived from cell phones and social media, our analysis is based on census-based mobility datasets. These datasets offer the distinct advantage of being universally accessible across spatial geographies and demographic groups. Consequently, our findings underscore the value of census-derived mobility data as a critical resource for urban analysis, particularly in contexts where obtaining large-scale mobility data from alternative sources is challenging or cost-prohibitive.

Author contributions

Conceptualization: Stanislav Sobolevsky.

Data curation: Devashish Khulbe.

Formal analysis: Devashish Khulbe.

Funding acquisition: Stanislav Sobolevsky.

Investigation: Devashish Khulbe, Stanislav Sobolevsky.

Methodology: Devashish Khulbe.

Supervision: Stanislav Sobolevsky.

Visualization: Devashish Khulbe.

Writing – original draft: Devashish Khulbe.

Writing – review & editing: Stanislav Sobolevsky.

References

1. Dong L, Duarte F, Duranton G, Santi P, Barthelemy M, Batty M, et al. Defining a city — delineating urban areas using cell-phone data. Nat Cities. 2024;1(2):117–25. https://doi.org/10.1038/s44284-023-00019-z
2. Li K, Niu X. Delineation of the shanghai megacity region of china from a commuting perspective: study based on cell phone network data in the yangtze river delta. J Urban Plann Dev. 2021;147(3). https://doi.org/10.1061/(asce)up.1943-5444.0000702
3. Jia T, Yu X, Shi W, Liu X, Li X, Xu Y. Detecting the regional delineation from a network of social media user interactions with spatial constraint: a case study of Shenzhen, China. Phys A: Statist Mech Appl. 2019;531:121719. https://doi.org/10.1016/j.physa.2019.121719
4. Ratti C, Sobolevsky S, Calabrese F, Andris C, Reades J, Martino M, et al. Redrawing the map of Great Britain from a network of human interactions. PLoS One. 2010;5(12):e14248. https://doi.org/10.1371/journal.pone.0014248 PMID: 21170390
5. Fortunato S. Community detection in graphs. Physics Reports. 2010;486(3–5):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
6. Yu W, Ai T, Shao S. The analysis and delimitation of Central Business District using network kernel density estimation. Journal of Transport Geography. 2015;45:32–47. https://doi.org/10.1016/j.jtrangeo.2015.04.008
7. Khulbe D, Belyi A, Mikeš O, Sobolevsky S. Mobility networks as a predictor of socioeconomic status in urban systems. Computational Science and Its Applications. 2023. p. 453–61.
8. Liu C, Han Y, Xu H, Yang S, Wang K, Su Y. A Community Detection and Graph-Neural-Network-Based Link Prediction Approach for Scientific Literature. Mathematics. 2024;12(3):369. https://doi.org/10.3390/math12030369
9. Sobolevsky S, Belyi A. Graph neural network inspired algorithm for unsupervised network community detection. Appl Netw Sci. 2022;7(1). https://doi.org/10.1007/s41109-022-00500-z
10. Mishina M, et al. Prediction of urban population-facilities interactions with graph neural network. In: Computational Science and Its Applications – ICCSA 2023. 2023.
11. Tandon A, Albeshri A, Thayananthan V, Alhalabi W, Radicchi F, Fortunato S. Community detection in networks using graph embeddings. Phys Rev E. 2021;103(2–1):022316. https://doi.org/10.1103/PhysRevE.103.022316 PMID: 33736102
12. Bogomolov Y, Belyi A, Sobolevsky S. Urban delineation through a prism of intraday commute patterns. Front Big Data. 2024;7:1356116. https://doi.org/10.3389/fdata.2024.1356116 PMID: 38504749
13. Xu Y, Belyi A, Bojic I, Ratti C. Human mobility and socioeconomic status: analysis of Singapore and Boston. Computers, Environment and Urban Systems. 2018;72:51–67. https://doi.org/10.1016/j.compenvurbsys.2018.04.001
14. Logan JR. Relying on the census in urban social science. City & Community. 2018;17(3):540–9. https://doi.org/10.1111/cico.12331
15. Martinuzzi S, Gould WA, Ramos González OM. Land development, land use, and urban sprawl in Puerto Rico integrating remote sensing and population census data. Landscape and Urban Planning. 2007;79(3–4):288–97. https://doi.org/10.1016/j.landurbplan.2006.02.014
16. Cohen SS, Mumma MT, Ellis ED, Boice JD Jr. Validating the use of census data on education as a measure of socioeconomic status in an occupational cohort. Int J Radiat Biol. 2022;98(4):587–92. https://doi.org/10.1080/09553002.2018.1549758 PMID: 30451561
17. Geronimus AT, Bound J. Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. Am J Epidemiol. 1998;148(5):475–86. https://doi.org/10.1093/oxfordjournals.aje.a009673 PMID: 9737560
18. U.S. Census Bureau. Longitudinal Employer-Household Dynamics (LEHD) Data, Snapshot Release S2023. Washington, DC: U.S. Census Bureau, Center for Economic Studies; 2025. https://lehd.ces.census.gov/
19. Sobolevsky S, Szell M, Campari R, Couronné T, Smoreda Z, Ratti C. Delineating geographical regions with networks of human interactions in an extensive set of countries. PLoS One. 2013;8(12):e81707. https://doi.org/10.1371/journal.pone.0081707 PMID: 24367490
20. Expert P, Evans TS, Blondel VD, Lambiotte R. Uncovering space-independent communities in spatial networks. Proc Natl Acad Sci U S A. 2011;108(19):7663–8. https://doi.org/10.1073/pnas.1018962108 PMID: 21518910
21. He M, Bogomolov Y, Khulbe D, Sobolevsky S. Distance deterrence comparison in urban commute among different socioeconomic groups: A normalized linear piece-wise gravity model. Journal of Transport Geography. 2023;113:103732. https://doi.org/10.1016/j.jtrangeo.2023.103732

22. Bogomolov Y, He M, Khulbe D, Sobolevsky S. Impact of income on urban commute across major cities in US. Procedia Computer Science. 2021;193:325–32. https://doi.org/10.1016/j.procs.2021.10.033
23. Rampášek L, Galkin M, Dwivedi VP, Luu AT, Wolf G, Beaini D. Recipe for a general, powerful, scalable graph transformer. NeurIPS. 2022;35:14501–15.
24. Dwivedi VP, Bresson X. A generalization of transformer networks to graphs. arXiv preprint 2020.https://arxiv.org/abs/2012.09699
25. Ying C, Cai T, Luo S, Zheng S, Ke G, He D, et al. Do transformers really perform badly for graph representation? In NeurIPS. 2021.
26. Klema V, Laub A. The singular value decomposition: its computation and some applications. IEEE Trans Automat Contr. 1980;25(2):164–76. https://doi.org/10.1109/tac.1980.1102314
27. Abu-El-Haija S, Mostafa H, Nassar M, Crespi V, Ver Steeg G, Galstyan A. Implicit SVD for graph representation learning. Advances in Neural Information Processing Systems. 2021;34: 8419–31.
28. Sarkar S, Dong A. Community detection in graphs using singular value decomposition. Phys Rev E Stat Nonlin Soft Matter Phys. 2011;83(4 Pt 2):046114. https://doi.org/10.1103/PhysRevE.83.046114 PMID: 21599247
29. Adams K, Pattanaik L, C W C. Learning 3D representations of molecular chirality with invariance to bond rotations. arXiv preprint 2021. https://arxiv.org/abs/2110.04383
30. Brüel-Gabrielsson R, Yurochkin M, Solomon J. Rewiring with positional encodings for graph neural networks. arXiv preprint 2022. https://arxiv.org/abs/2201.12674
31. Li P, Wang Y, Wang H, Leskovec J. Distance encoding: design provably more powerful neural networks for graph representation learning. In: NeurIPS. 2020. p. 4465–78.
32. Dwivedi VP, Luu AT, Laurent T, Bengio Y, Bresson X. Graph neural networks with learnable structural and positional representations. In: International Conference on Learning Representations. 2022.
33. Xu F, Lin Z, Xia T, Guo D, Li Y. SUME. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2020;4(3):1–25. https://doi.org/10.1145/3411807
34. Chandra DK, Leopold J, Fu Y. NodeSense2Vec: Spatiotemporal Context-Aware Network Embedding for Heterogeneous Urban Mobility Data. In: 2021 IEEE International Conference on Big Data (Big Data). 2021. p. 2884–93. https://doi.org/10.1109/bigdata52589.2021.9672072
35. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint 2016. https://arxiv.org/abs/1609.02907
36. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks?. arXiv preprint 2018. https://arxiv.org/abs/1810.00826
37. Li Q, Han Z, Wu X. Deeper insights into graph convolutional networks for semi-supervised learning. AAAI. 2018;32(1). https://doi.org/10.1609/aaai.v32i1.11604
38. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. In: NeurIPS. 2017.
39. Chen M, Kuzmin K, Szymanski BK. Community detection via maximization of modularity and its variants. IEEE Trans Comput Soc Syst. 2014;1(1):46–65. https://doi.org/10.1109/tcss.2014.2307458
40. Lee C, Wilkinson DJ. A review of stochastic block models and extensions for graph clustering. Appl Netw Sci. 2019;4(1):122. https://doi.org/10.1007/s41109-019-0232-2
41. Sobolevsky S, Campari R, Belyi A, Ratti C. General optimization technique for high-quality community detection in complex networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2014;90(1):012811. https://doi.org/10.1103/PhysRevE.90.012811 PMID: 25122346
42. Peixoto TP. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. Phys Rev E Stat Nonlin Soft Matter Phys. 2014;89(1):012804. https://doi.org/10.1103/PhysRevE.89.012804 PMID: 24580278
43. Fournier N, Christofa E. On the impact of income, age, and travel distance on the value of time. Transportation Research Record: Journal of the Transportation Research Board. 2020;2675(3):122–35. https://doi.org/10.1177/0361198120966603
44. Kirkley A. Information theoretic network approach to socioeconomic correlations. Phys Rev Research. 2020;2(4). https://doi.org/10.1103/physrevresearch.2.043212
45. Oczki J, Wędrowska E. The use of Csiszár's divergence to assess dissimilarities of income distributions of EU countries. Quantitative Methods in Economics. 2014;15(2):167–76.
46. Magdalou B, Nock R. Income distributions and decomposable divergence measures. Journal of Economic Theory. 2011;146(6):2440–54. https://doi.org/10.1016/j.jet.2011.06.017