Mobility Networks as a Predictor of Socioeconomic Status in Urban Systems

Devashish Khulbe[1(✉)], Alexander Belyi[1], Ondřej Mikeš[2], and Stanislav Sobolevsky[1,3,4]

[1] Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic
dk3596@nyu.edu

[2] RECETOX, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic

[3] Center for Urban Science and Progress, New York University, 370 Jay Street, Brooklyn, NY 11201, USA

[4] Institute of Law and Technology, Faculty of Law, Masaryk University, Veveri 70, 611 80 Brno, Czech Republic

Abstract. Modeling socioeconomic dynamics has always been an area of focus for urban scientists and policymakers, who aim to better understand and predict the well-being of local neighborhoods. Such models can inform decision-makers early on about expected neighborhood performance under normal conditions, as well as in response to considered interventions before official statistical data is collected. While features such as population and job density, employment characteristics, and other neighborhood variables have been studied and evaluated extensively, research on using the underlying networks of human interactions and urban structures is less common in modeling techniques. We propose using the structure of the local urban mobility network (weighted by commute flows among a city's geographical units) as a signature of the neighborhood and as a source of features to model its socioeconomic quantities. The network structure is quantified through node embedding generated using a graph neural network representation learning model. In the proof-of-concept task of modeling the location's median income and housing profile in two different cities, such network structure features provide a noticeable performance advantage compared to using only the other available social features. This work can thus inform researchers and stakeholders about the utility of mobility network structure in a complex urban system for modeling various quantities of interest.

Keywords: Mobility networks · socioeconomic modeling · Network representations · Graph Neural Networks

1 Introduction

Networks are an integral part of any urban system, with many intra-city dynamics that can be perceived as a structure of interconnected entities. Urban researchers

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023, LNCS 13957, pp. 453–461, 2023.
https://doi.org/10.1007/978-3-031-36808-0_32

have recognized the utility of network structures and topology to measure relationships and understand the organization of cities. Networks from the perspective of social, geographical, political, and cultural domains have been well established in the context of cities [1]. Their scales, structures, and development over time have informed urban studies in understanding complex city characteristics. Inter-city networks have also been an area of focus, with research informing broader policies like polycentric urban development [2] and long-term economic investments [2,3] in large urban regions. Local neighborhood characteristics like median income, employment rate, housing prices, etc., are some of the many variables that help urban scientists develop a "profile" of an area. These socioeconomic indicators help understand the local health of a neighborhood. Studying the evolution of economic health over time is particularly useful to policymakers with regard to decisions like investments, development potential, energy consumption, and migration trends within a large city [4–6]. Socioeconomic modeling has been studied extensively at both the individual and neighborhood levels. Contextual socioeconomic features such as population density, job density, ethnic profile, etc., are well established as indicators of a larger population standard of living [7]. In smaller urban areas where consistent official data could be hard to obtain, features like bank card transactions have been used to predict regional economic indices [8]. More recently, researchers employed aggregate mobility indicators for socioeconomic modelling [9]. Spatial variables like natural amenities, public parks, etc., have also been used in modeling [10] to understand the impact of physical geography. In the context of networks, these external variables have also been studied in conjunction with road network topology to model and understand traffic capacity in urban regions [11]. However, interactions among entities is not a common feature that is incorporated in modeling, even though relationships between a given neighborhood and other parts of a city could be an important aspect of measuring economic health. Traditional machine learning modeling approaches are limited by the inability to properly define a feature space for a region within a larger interconnected network. With the development and proven applicability of methods like Graph Neural Networks, there is a potential to leverage their power in urban contexts.

Large-scale mobility data has been used to study complex city dynamics while measuring the resilience of cities [12], although the network structure itself is not considered here to model socioeconomic performance indicators. Researchers have proposed using multitude of digital and physical variables like mobile app records, taxi and bike trips, and geo-tagged social media data as a measure of urban mobility [13]. Data-driven research has also emphasized modeling and studying dependencies among external features like 311 service requests data [14], showing the importance of local neighborhood context in modeling. Using social media networks, metrics like centrality and average clustering coefficient have also been used as features to model economic growth in neighborhoods as a result of government investments [15]. These methods rely on combining external node variables with pre-defined network metrics as input features rather than being able to learn the network representations of the whole urban area. Urban street networks using geo-located Twitter data have also been used to study

public spaces and evaluate their success [16], indicating that granular networks are useful in the analysis of sophisticated urban geographies. Networks based on twitter metadata have also proved to be useful in delineating urban neighborhoods [17]. However, social media-based networks are not the best proxy for urban interactions, as they may exclude large populations based on age and location. Dynamic mobility data may help decision-makers assess the distribution and/or dynamics of the socioeconomic performance across the city based on the available urban mobility data in nearly real-time, without having to wait for years until Census and/or other official statistics reflecting such performance gets collected or updated. Figure 1 illustrates the idea outline with New York City's mobility among zip-code neighborhoods.

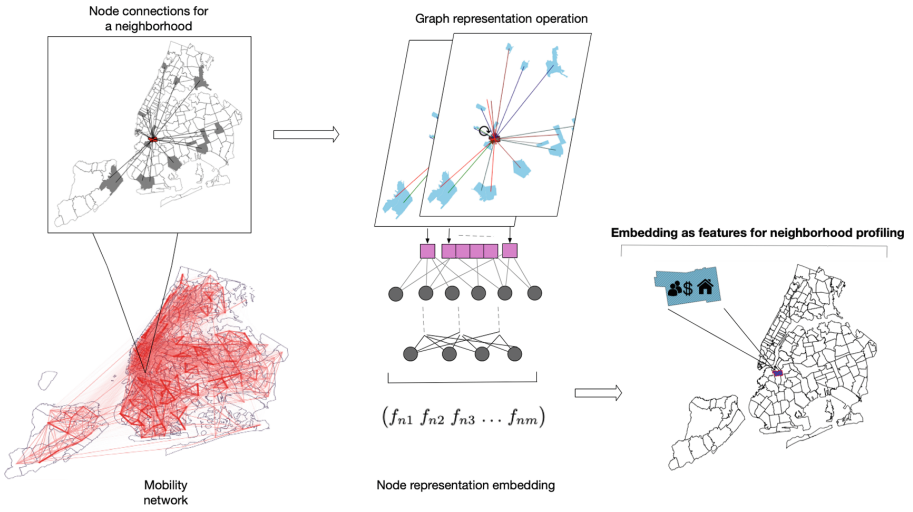

Fig. 1. Structure of the locations' mobility network as a characteristic of its socioeconomic performance - an illustration for NYC

2 Data and Methodology

2.1 Network Representation

The general idea behind constructing representations of a network is to capture the connectivity and topology through some transformations, which produces a representation feature set (embedding). With respect to a given node, a node embedding thus represents each node as a vector in a low dimensional space. In this regard, many techniques have been used in graphs to generate embedding, with graph-based applications such as node classification, node clustering, node retrieval/recommendation, link prediction, etc. [18]. However, in many of the

applications, the stability and robustness of the embedding become important. Techniques such as node2vec are found to be susceptible to sources of noise, stochastic effects, or specific parameter choices [19,20]. At the same time, GNN (and some variations) based embedding has shown promising results in generating stable and robust embedding [21,22]. We consider two techniques to produce node representations:

1. Personalized PageRank (PPR), which quantifies the significance of all nodes with respect to the given node. Formally, PPR embedding of a source node is the stationary probability distribution of a Markov chain that, with probability α, randomly transitions following the link structure of the network, and with probability $1 - \alpha$ teleports to a source node [23,24]. For a given node s in a weighted directed graph $G(V, E)$ with adjacency matrix A, we normalize neighboring edges by the total of edge weights among s and its connections to obtain a stochastic transition matrix P. The teleport probabilities are thus proportional to the relative "activity" between the origin and destination node.

2. Graph Neural Network (GNN) based embedding, which quantifies the node embedding as the series of network convolutions of certain initial embedding (e.g. one-hot encoding) of the interconnected nodes [25]. The goal is to learn a node function within the graph that preserves important structural properties of the network, such as local connectivity and symmetry. The formula for computing GNN embedding typically involves several iterations of message passing between nodes in the graph, which allows for the incorporation of local and global information. The message passing at the node level can be formulated as

$$h_i^k = \sigma(W_{self}^k h_i^{k-1} + W_{neigh}^k \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} h_j^{k-1} + b^k) \tag{1}$$

where h_i^k represents the embedding of node i at iteration k, $\mathcal{N}(i)$ represents the set of neighboring nodes of i, W^k and b^k are learnable weight and bias parameters at iteration k, and σ is a non-linear activation function such as ReLU or sigmoid. By repeating this message passing process, GNN embedding can capture increasingly complex relationships and structures in the graph. Many GNNs can also be succinctly defined using graph-level equations. In the case of a basic GNN, we can write the graph-level definition of the model as follows:

$$H^k = \sigma(H^{k-1} W_{self}^k + A H^{k-1} W_{neigh}^k + B^k) \tag{2}$$

where $H^k \in R^{V \times d}$ denotes the matrix of node representations at layer k in the GNN (with each node corresponding to a row in the matrix), B^k is the matrix of bias vectors b^k, and A is the normalized graph adjacency matrix. H^k is usually a lower dimensional matrix that encodes the network representations through the GNN framework. Embedding is thus learned as a part of the network representation task, where the goal is to reconstruct the original network edges.

2.2 Experiment Setup

We consider the mobility network as a directed graph $G(V, E)$, where the nodes V represent the geographical regions in a city and edges $E \subseteq V \times V$ are weighted by the mobility values between nodes. The graph adjacency matrix is therefore populated by the mobility values among the nodes, which then serve as the input to generate node representations. While PPR embedding is represented by a vector of probabilities for each node, GNN-based node representation is a d dimensional vector, where dimensionality is configured as part of representation learning of the mobility network. The output matrix (embedding) is then used as node features, with each vector corresponding to a node in the mobility network. Data from two cities are considered:

New York City, USA. To represent NYC mobility, we consider job flows obtained from The Longitudinal Employer-Household Dynamics (LEHD), a U.S. census bureau program [26]. U.S. Census also provides other neighborhood variables. The median income is considered as the target feature representing the socioeconomic profile of a node. We aggregate the mobility and median income on the zip-code level to get 173 nodes with significant mobility and socioeconomic information. For baseline models, we use population density and job density as contextual neighborhood features.

Brno, Czechia. Brno is the second largest city in Czechia, with a population of around 700,000 people in its metropolitan area. The city is an important cultural and administrative center and the largest city in the Moravia region of Czechia. Mobility information for Brno is retrieved from the commute information from T-mobile cell-phone-based location data. The data is provided as hourly aggregated values among 48 unique districts in the city. We further aggregate the commute flows over 14 available days to weight the network with districts as nodes. With economic characteristics like income not easily available for Brno, we consider housing profile as a measure of a district's economic status. The percentage of single-family houses in districts is taken as the target feature. Furthermore, the employment rate is used as input to train baseline models.

The embeddings and node features thus serve as inputs to the supervised learning models. Three approaches are considered – Regularized Linear model (Lasso), Random Forests, and a Vanilla Neural Network (VNN). These models are tuned to find the optimal hyperparameters for the modeling task.

3 Results

3.1 Network Embedding vs. Contextual Node Features

Table 1 presents out-of-sample results using various input features with the three models considered. The evaluation of model performance is done based on Root

Table 1. Out-of-sample RMSE values with supervised learning based on different model inputs

	NYC			Brno		
	Node features	PPR	GNN	Node features	PPR	GNN
Lasso	0.92	0.78	0.54	1.41	0.74	0.45
RF	0.63	0.59	0.42	1.06	0.68	0.42
VNN	0.63	0.47	0.40	1.01	0.60	0.42

Mean Squared Error (RMSE) values across a range of features, including contextual node features and network embeddings derived from PPR and GNN techniques. The input features are used individually to train the models, allowing for a comparison of the modeling capability of embeddings over neighborhood variables. The results show that both node embedding approaches, PPR and GNN, outperform external node features considered individually. In particular, the GNN embedding of mobility network nodes has the best results overall, with the highest out-of-sample performance in both cities. The use of contextual node features is also useful, although to a lesser extent than the network embedding. This suggests that mobility embedding alone is effective in capturing complex relationships between nodes and the larger interactions in a city. Additionally, we also notice that model fitting on Brno data is somewhat unstable with noticeable variance on the validation set RMSE values. Whereas the results for NYC are much more stable, highlighting the differences in data and network sizes between the cities. Nevertheless, the better performance with embeddings for both cities shows the importance of mobility in socioeconomic profiling.

Although the target features for NYC and Brno are different, they represent important aspects of an urban neighborhood. While median income is directly a measure of economic health, housing type could be a proxy for complex social dynamics, including poverty, transport accessibility, etc. Local neighborhood features could be an important signal for these characteristics. Thus, the superior performance of mobility embedding over node features is significant since node features do not capture the intricacies of the network structure and interactions between nodes.

3.2 Discussion

We investigate the utility of mobility networks over two very different cities. NYC and Brno differ vastly in scale, geography, economy, and demographics. Nevertheless, results show that mobility network embedding proves to have significant modeling capability for the socioeconomic profiling of neighborhoods in both cities. GNN-based embedding is particularly promising in this regard, with

the potential to improve the accuracy of various network modeling applications. The inherent difference in scale and lack of consistent socioeconomic features between the two cities could be a potential reason for somewhat unstable results for Brno. Brno's mobility network is much smaller, with data available among 48 districts. This potentially is an issue for supervised modeling, with not enough data samples to optimize model parameters. More granular data could help to stabilize the models. In general, data augmentation techniques could be explored for fixing data scarcity issues in smaller cities. There is a broader scope to evaluate the impact of other urban networks, such as social media, in urban profiling. While we have focused on median income and housing profile as the target features, there is room to investigate the methodology in other socioeconomic variables of interest, such as unemployment, poverty rate, education, etc.

4 Conclusion

Diverse networks, which can represent dynamic relationships among entities in urban systems, have proven to be quite useful for researchers. However, using their representations in modeling complex urban quantities is not commonplace. With interactions among individuals or urban entities becoming increasingly intense in the modern world, the importance of networks in urban modeling becomes more relevant. Applications such as link prediction, node classification, etc., can thus be extended in urban contexts, in which GNN-based methods have already proven to be useful. We demonstrate their ability to model the socioeconomic characteristics of urban geographies. While mobility is an essential aspect of any city's function, we show that its network structure representations can be useful estimators for urban modeling tasks, as demonstrated in the case of modeling the average income of urban locations. Additionally, there is potential for other urban networks, such as traffic networks, social media, etc., also to inform urban models.

Our broader work aims to find the most efficient way of incorporating the network structure with other available neighborhood local contextual features to determine the maximum modeling capability. We also aim to experiment more with supervised learning, and non-linear network representation approaches to fine-tune and improve the models. Overall, our work can inform scientists and stakeholders to study and model complex urban systems from the perspective of a city's real-time networks while also leveraging the utility of local geographical contexts.

Acknowledgements. 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). The work of Stanislav Sobolevsky was also partially supported by ERDF "CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence" (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

References

1. Pflieger, G., Rozenblat, C.: Introduction. Urban networks and network theory: the city as the connector of multiple networks. Urban Stud. **47**(13), 2723–2735 (2010)
2. Liu, X., Derudder, B., Wu, K.: Measuring polycentric urban development in China: an intercity transportation network perspective. Reg. Stud. **50**(8), 1302–1315 (2016)
3. Glaeser, E.L., Ponzetto, G.A., Zou, Y.: Urban networks: connecting markets, people, and ideas. Pap. Reg. Sci. **95**(1), 17–59 (2016)
4. Townsend, A.M.: Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia. W. W. Norton & Company, New York (2013)
5. Borjas, G.J.: Economic theory and international migration. Int. Migr. Rev. **23**(3), 457–485 (1989)
6. Swan, L.G., Ugursal, V.I.: Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew. Sustain. Energy Rev. **13**(8), 1819–1835 (2009)
7. Wen, M., Browning, C.R., Cagney, K.A.: Poverty, affluence, and income inequality: neighborhood economic structure and its implications for health. Soc. Sci. Med. **57**(5), 843–860 (2003)
8. Sobolevsky, S., Massaro, E., Bojic, I., Arias, J.M., Ratti, C.: Predicting regional economic indices using big data of individual bank card transactions. In: 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, pp. 1313–1318 (2017). https://doi.org/10.1109/BigData.2017.8258061
9. Xu, Y., Belyi, A., Bojic, I., Ratti, C.: Human mobility and socioeconomic status: analysis of Singapore and Boston. Comput. Environ. Urban Syst. **72**, 51–67 (2018)
10. Lee, S., Lin, J.: Natural amenities, neighbourhood dynamics, and persistence in the spatial distribution of income. Rev. Econ. Stud. **85**(1), 663–694 (2018)
11. Loder, A., Ambühl, L., Menendez, M., et al.: Understanding traffic capacity of urban networks. Sci. Rep. **9**, 16283 (2019). https://doi.org/10.1038/s41598-019-51539-5
12. Yabe, T., Rao, P.S.C., Ukkusuri, S.V.: Resilience of interdependent urban sociophysical systems using large-scale mobility data: modeling recovery dynamics. Sustain. Cities Soc. **75**, 103237 (2021). https://doi.org/10.1016/j.scs.2021.103237
13. Zhu, E., Khan, M., Kats, P., Bamne, S.S., Sobolevsky, S.: Digital urban sensing: a multi-layered approach. arXiv preprint arXiv:1809.01280 (2018)
14. Wang, L., Qian, C., Kats, P., Kontokosta, C., Sobolevsky, S.: Structure of 311 service requests as a signature of urban location. PLoS ONE **12**(10), e0186314 (2017). https://doi.org/10.1371/journal.pone.0186314
15. Zhou, X., Hristova, D., Noulas, A., Mascolo, C., Sklar, M.: Cultural investment and urban socioeconomic development: a geosocial network approach. R. Soc. Open Sci. **4**(9), 170413 (2017). https://doi.org/10.1098/rsos.170413
16. Agryzkov, T., Martí, P., Nolasco-Cirugeda, A., Serrano-Estrada, L., Tortosa, L., Vicent, J.F.: Analysing successful public spaces in an urban street network using data from the social networks Foursquare and Twitter. Appl. Netw. Sci. **1**(1), 1–15 (2016). https://doi.org/10.1007/s41109-016-0014-z
17. Sobolevsky, S., Kats, P., Malinchik, S., Hoffman, M., Kettler, B., Kontokosta, C.: Twitter Connections Shaping New York City (2018). https://doi.org/10.24251/HICSS.2018.127
18. Cai, H., Zheng, V.W., Chang, K.C.-C.: A comprehensive survey of graph embedding: problems, techniques, and applications. IEEE Trans. Knowl. Data Eng. **30**(9), 1616–1637 (2018). https://doi.org/10.1109/TKDE.2018.2807452

19. Hacker, C., Rieck, B.: On the Surprising Behaviour of node2vec. arXiv preprint arXiv:2206.08252 (2022)
20. Schumacher, T., Wolf, H., Ritzert, M., Lemmerich, F., Grohe, M., Strohmaier, M.: The effects of randomness on the stability of node embeddings. In: Kamp, M., et al. (eds.) ECML PKDD 2021. CCIS, vol. 1524, pp. 197–215. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-93736-2_16
21. Gallicchio, C., Micheli, A.: Fast and deep graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 3898–3905 (2020)
22. Agarwal, C., Lakkaraju, H., Zitnik, M.: Towards a unified framework for fair and stable graph representation learning. In: Uncertainty in Artificial Intelligence, pp. 2114–2124. PMLR (2021)
23. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford infolab (1999)
24. Gleich, D.F.: PageRank beyond the web. SIAM Rev. **57**(3), 321–363 (2015). https://doi.org/10.1137/140976649
25. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional network. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=SJU4ayYgl
26. United States Census Bureau: Longitudinal Employer-Household Dynamics (2021). https://lehd.ces.census.gov/. Accessed 16 Mar 2023