RESEARCH

Leveraging Network Representation Learning and Community Detection for Analyzing the Activity Profiles of Adolescents

Saket Gurukar[1*], Bethany Boettner[2], Christopher Browning[3], Catherine Calder[4] and Srinivasan Parthasarathy[1]

[*]Correspondence: gurukar.1@osu.edu
[1]Computer Science and Engineering, The Ohio State University, Columbus, USA
Full list of author information is available at the end of the article

Abstract

Human mobility analysis plays a crucial role in urban analysis, city planning, epidemic modeling, and even understanding neighborhood effects on individuals' health. Often, these studies model human mobility in the form of co-location networks. We have recently seen the tremendous success of network representation learning models on several machine learning tasks on graphs. To the best of our knowledge, limited attention has been paid to identifying communities using network representation learning methods specifically for co-location networks. We attempt to address this problem and study user mobility behavior through the communities identified with latent node representations. Specifically, we select several diverse network representation learning models to identify communities from a real-world co-location network. We include both general-purpose representation models that make no assumptions on network modality as well as approaches designed specifically for human mobility analysis. We evaluate these different methods on data collected in the Adolescent Health and Development in Context (AHDC) study. Our experimental analysis reveals that a recently proposed method (LocationTrails) offers a competitive advantage over other methods with respect to its ability to represent and reflect community assignment that is consistent with extant findings regarding neighborhood racial and socio-economic differences in mobility patterns. We also compare the learned activity profiles of individuals by factoring in their residential neighborhoods. Our analysis reveals a significant contrast in the activity profiles of individuals residing in white-dominated vs. black-dominated neighborhoods and advantaged vs. disadvantaged neighborhoods in a major metropolitan city of United States. We provide a clear rationale for this contrastive pattern through insights from the sociological literature.

Keywords: Mobility Analysis; Activity Profiles; Co-location networks; GPS

Introduction

The ability to capture the location of individuals using GPS-enabled devices has allowed researchers to analyze human mobility with unprecedented precision. Beyond individual mobility trajectories, data on spatially delimited groups of individuals

has provided the opportunity to estimate bipartite, *co-location* networks where users and locations are treated as nodes, and location visits are treated as edges. These co-location networks, however, do not necessarily indicate direct contact between individuals at specific geographic locations; instead, they capture the potential for shared experiences and exposures. Co-location networks uncover the structure of shared exposure in a collective sense, illuminating the potential for contagion (social or viral), cohesion, and related outcomes such as health and crime [1, 2].

Recent and emerging research suggests that the extraction of communities (consisting of individuals) from such co-location networks that model human activity spaces can provide important information about the functioning of a city and its neighborhoods [3, 4, 5]. For instance, understanding the community structure of co-location networks can shed light on systematic patterns of urban racial and socioeconomic segregation in everyday routines beyond those identified by an exclusive focus on residential sorting [1]. Estimating communities based on shared routines also helps identify indirect or higher-order location exposures that may be relevant for contagion processes (but not necessarily rooted in spatial proximity).

The numerous applications of co-location networks warrant careful consideration of appropriate methods for their construction. One could adopt a *structured data* collection approach, followed by the Los Angeles Family and Neighborhood Study [6], in which one first samples individuals/households from a region/city and then prompt subjects for the location of typical routine activity destinations such as workplaces, schools, or grocery stores; the co-location network is then constructed based on the locations provided from survey-style instrumentation. An alternative method is to adopt an *unstructured approach* in which one could provide GPS-enabled devices to the sampled individuals/households from a region/city, record the spatial location of the individual at a short interval, find the stationary locations where the individuals spend a significant time and then construct a co-location network between individuals and stationary locations.

The Adolescent Health and Development in Context (AHDC) study [7] follows both structured and unstructured data collection approaches to capture individuals' mobility in Franklin County, Ohio. To collect structured data on mobility, the AHDC study surveys caregivers of adolescents about their location visits and then forms a co-location network (one can denote this network as a *coarser-grained co-location network*). The unstructured approach is based on the spatial coordinates of adolescents over regular intervals and then forms a co-location network (one can denote this network as a *finer-grained co-location network*).

In this study, we focus on extracting community structure from the fine-grained co-location network. Since there is no ground truth available, we assess the extent to which alternative approaches to community detection align with previous findings on neighborhood racial and socioeconomic differences in mobility patterns [1, 8]. Our approach for extracting communities relies on computing a meaningful vector representation of each node in the co-location network (for all location and user nodes). These vectors can then be utilized by any off-the-shelf clustering algorithms (such as K-means [9] or Gaussian Mixture Models [10]) to identify meaningful communities of users and their shared exposure locations. We evaluate the use of several state-of-the-art approaches for computing the representation of each node within the fine-grained AHDC co-location network. These include:

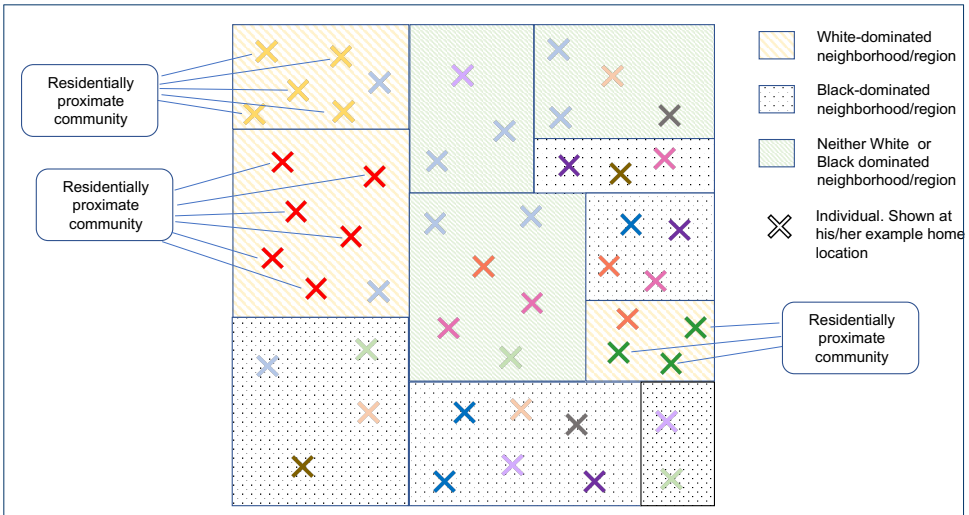


Figure 1: Toy example showing an example city (outlined by a box) and its neighborhoods (outlined by sub-boxes). Individuals are shown by cross marks. Individuals are placed at their illustrative home locations in an neighborhood. Individuals shown in same color belong to same community. Residentially proximate community refers to group of individuals who reside in the same neighborhood and share same community.

- A previous effort by, Xi et al. [1], that focused on identifying communities from the *coarser-grained* AHDC co-location network.
- Several neural network based models that have recently shown to be highly effective for the learning of node representations from such network data. These include efforts such as DeepWalk [11], and LINE [12].
- A recently proposed low-resource (efficient) neural approach called LocationTrails [13]. Unlike other neural methods, LocationTrails explicitly leverages the sequential ordering of a user's visits to specific locations that is available in such fine-grained co-location networks.

We present a toy example in Figure 1 that defines the terminology we use to explain our findings. A neighborhood is dominated by a given race if its percent population is higher than 70% [14, 15]. In Figure 1, note the presence of residentially proximate communities in the white-dominated neighborhoods and the lack of residentially proximate communities in the black-dominated neighborhoods.

Our key findings can be summarized as follows. First, a qualitative examination of the communities extracted by different methods suggests that the community structures extracted by LocationTrails identify patterns that are consistent with our understanding of urban racial and socioeconomic segregation in everyday routines. Second, among the other neural approaches (DeepWalk [11] and LINE [12]) appear to offer the strongest performance, although these methods do appear to be biased towards residentially proximate community structures, potentially mischaracterizing the routine activity patterns of more segregated and socioeconomically disadvantaged neighborhoods [8]. Third, several important patterns identified by LocationTrails and the other neural models largely agree with the results Xi et

al. observed from the coarse-grained AHDC co-location network analysis study [1]. However, our qualitative analysis suggests that [1] was less effective on the fine-grained co-location network data, when compared to LocationTrails. Fourth, a quantitative examination of the activity profiles of the individuals residing in neighborhoods with different characteristics (white-dominated vs. black-dominated, advantaged vs. disadvantaged) reveals that individuals who reside in white-dominated neighborhoods are more likely to share the same cluster than their black counterparts. While individuals who reside in black-dominated neighborhoods often do not share the same cluster as they seem to have dissimilar activity profiles.

The rest of the paper is organized as follows. The next section describes the data collection, data cleaning, and formation of the fine-grained co-location network from the AHDC study - an important contribution of this study. The Methods section overviews related work and summarizes the selected methods utilized for our evaluation. The Results section presents the analysis of the selected methods on the AHDC fine-grained co-location network dataset. We present the conclusions and contributions of our work in the Conclusions section.

AHDC Activity Pattern Data

Overview

The Adolescent Health and Development in Context (AHDC) study [7] is an ongoing longitudinal data collection study. The goal of the AHDC study is to explore the relationship between aspects of the social and spatial contexts of everyday routines and the health and wellbeing of urban youth. To that end, the AHDC study collects data on multiple contexts of youth development from a representative sample of 1,347 adolescents (age 11-17 years old) residing within Franklin County (contains the city of Columbus – Ohio's largest city) using a prospective cohort design. Franklin County is racially and ethnically diverse – White (Non-Hispanic) (62%), Black or African American (Non-Hispanic) (22.9%), Asian (Non-Hispanic) (5.38%), and White (Hispanic) (3.25%) [16]. In terms of social and economic characteristics, the Columbus metropolitan area is representative of the average US metropolitan area [17]. The data collection from youth and their caregivers occurs in two waves (Wave 1 and 2) separated over one year period. In this work, we focus on data collected in Wave 1. Wave 1 data collection took place between April 2014 and July 2016. The data collection design is as follows. The AHDC study first performs an Entrance Survey – the *structured data* collection approach – with the adolescents and their caregivers. The survey covers a broad range of topics related to demographic and socioeconomic background, household composition, family structure and marital status, employment and income, health, social support, and alcohol/substance use. The entrance survey included a "location generator" [18] in which caregivers and adolescents provided information about the locations of the youth's everyday routine activities (e.g., school, work, grocery shopping, etc.).

Xi et al. [1] construct a co-location network from the above mentioned Entrance survey where the reported locations are aggregated to the census block group. The authors [1] perform data cleaning based on the missing data and the density of caregivers in a block group. The resultant *coarser-grained co-location network* consists of 1307 caregivers (out of 1405 caregivers) and has 883 block groups. Census block

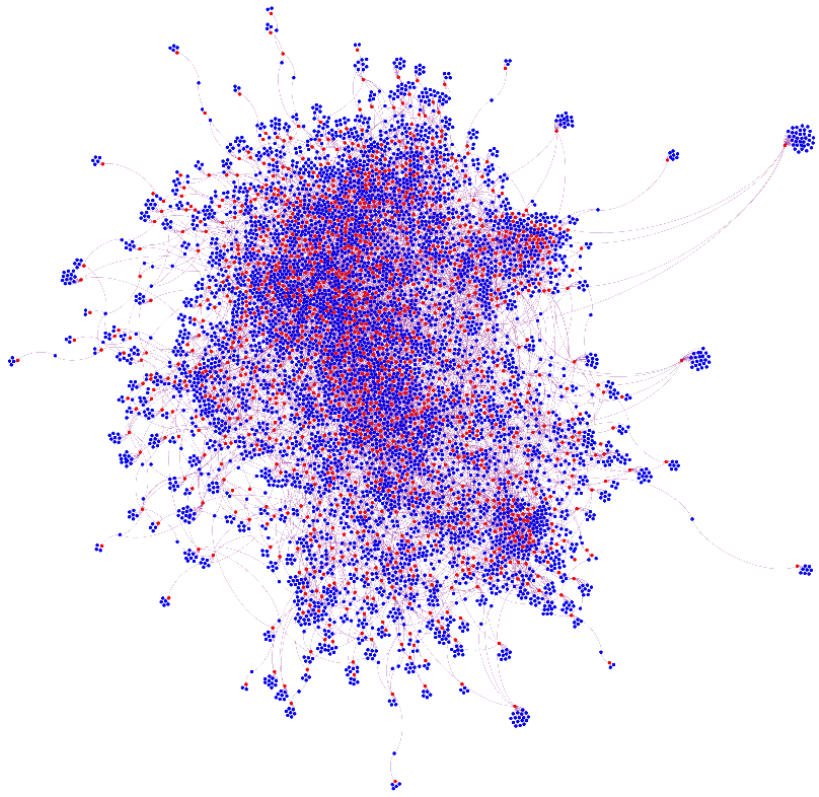


Figure 2: Visualization of the AHDC co-location network. Adolescents are shown in red color nodes while the locations are shown in blue color nodes. An edge represents a location visit by an adolescent. Here, edge is considered to be unweighted. We use ForceAtlas2 algorithm [21] to visualize the co-location network.

groups are statistical divisions of census tracts and are generally defined to contain between 600 and 3,000 people [19].

The Entrance Survey of the AHDC study was followed by geographically explicit ecological momentary assessment (GEMA) [20] for a period of seven days – the *unstructured data* collection approach. During this period, adolescents carried a study provided GPS-enabled smartphone that collected real-time assessments of locations, activities, and experiences as well as near-continuous Global Positioning System (GPS) coordinate data. The spatial coordinate data were collected through GPS satellites every 30 seconds. However, if no GPS satellite coordinates were collected for a period of 10 minutes or more, location coordinates were recorded from the cell network tower connection every minute to obtain an approximate location.

Next, we describe the data cleaning and construction of the *finer-grained co-location network* from the unstructured data collection approach.

Deriving Finer-Grained Co-Location Network from the Unstructured Activity Data

The collected GPS data are subject to error and contain noise [22]. We process the collected GPS using the convex hull-based binning algorithm [23] to capture an

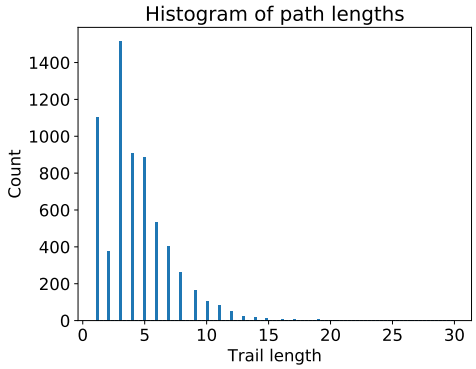


Figure 3: Histogram of trail lengths. A trail corresponds to the sequence of locations visited by an individual in a day.

# Adolescents	# Locations	# Edges	Mode length	Mean length	# of Trails
1,347	1,347 (home) + 4,225 (activity)	10,057	4	4.33	6,483

Table 1: The statistics of the trails on the AHDC dataset. Mode and Mean are computed on the distribution of length of all the trails.

accurate estimate of the location. The algorithm gives us the stationary and travel periods of the adolescents [24] and the convex hull centroid over the stationary periods GPS coordinates is estimated as the visited locations. The visited locations are then presented to the adolescents on a map using a recall-aided interactive space-time budget application [24]. The application has a graphical user interface (GUI) showing Google Maps and has several other data collection functionalities. Using the application, the adolescents in the AHDC study corroborate the estimated visited location and also provide the labels of the location. The collected latitude and longitude values of stationary locations need to be converted to a location id so that we could form a co-location graph between user-ids and location-ids. This conversion process is known as reverse geocoding, and we utilize the OpenStreetMap API [1] for this purpose.

The visualization of the constructed co-location network is shown in Figure 2. In Figure 2, we observe that there exist several locations (such as schools) commonly visited by most adolescents. We also observe that at the periphery there are few locations (such as a relative's house) that are visited by a small number of adolescents. The statistics of the constructed *fine-grained co-location network* are shared in Table 1. We also share the location visits statistics in the table. A trail represents the number of locations visited by adolescents in a day. The mean and mode of the trails are 4.33 and 4, while the histogram of trail lengths is shown in Figure 3. From the visualization, one can observe that there are certain locations (shown in blue) that were only visited by few adolescents. These locations could be the home of the adolescents, their relative's house, or local stores that were not visited by other adolescents in the study. We also observe a significant number of locations (such as schools, shopping malls) that were visited by multiple adolescents. The anonymized

[1]https://photon.komoot.io/

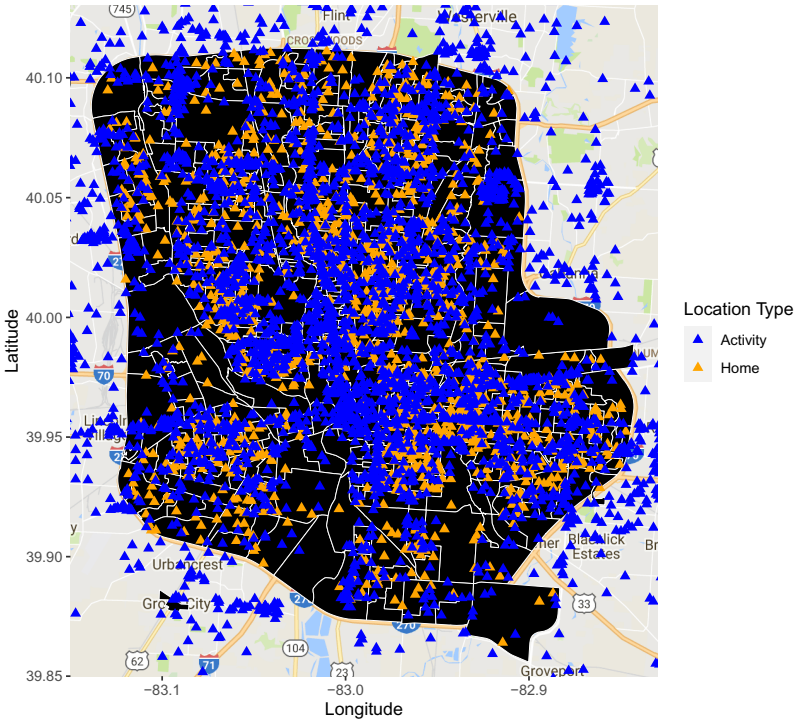


Figure 4: The activity and home locations of the adolescents. The locations are anonymized (anonymization process is explained below (see Figure reference in main text)).

home locations and activity locations of the adolescents are shown in Figure 4. Here, the locations are anonymized as follows: given the latitude and longitude of the location, we first identify the block group of the location and then set the home location of the adolescent to be a random point in the block group.

Methodology

The extent to which activity spaces—the collection of an individual's routine activity locations—overlap with those of their neighbors or those with similar backgrounds provides important information about the functioning of a city and its neighborhoods. The identification of communities from the co-location network can provide additional insight into the structure of shared urban routines. In this work, we evaluate both network representation learning (NRL) methods [11, 12, 13, 1, 25] and standard network science methods [26, 27, 28, 29] to identify such communities. In the case of NRL methods, the first step is to identify a meaningful representation of individuals (adolescents in our co-location networks) as well as that of the routine areas they visit (locations in our co-location network). To compute the representation of nodes within a two-mode co-location network, we draw on exemplars from general-purpose network representation learning and human-mobility network representation learning. In the case of standard network science methods, we select

two popular methods that rely on pre-defined metrics to identify communities. We discuss both NRL and network science methods in the next sections.

Network Representation Learning

The network representation learning (NRL) models aim to learn a representation of nodes such that the similarity of nodes in graph space is approximated by the closeness of nodes in the representation space. One of the initial network representation learning models is Laplacian Eigenmaps [30] which learns node representations by preserving the first-order proximity between the nodes – connected nodes should have node representations with low L2 distance. Inspired by the effectiveness of neural networks, Perozzi et al. [11] proposed Deepwalk that performs truncated random-walk on the graphs and then applies skip-gram [31] objective function on the random-walks to learn the node representations. Node2vec [32] proposed an approach to bias the random-walks and then adopt the Deepwalk strategy to learn the node representations. LINE [12] proposes two objective functions that preserve both first-order and second-order proximity – nodes with similar neighbors should have node representations with low L2 distance – for learning the node embeddings. NetMF [33] argues that the skip-gram based models with negative sampling optimization such as Deepwalk [11], Node2vec [32], LINE [12] and PTE [34] are implicitly factorizing matrices formed with graph laplacians. Recently, Huang et al. [35] provided an analytical framework for random-walk based graph embedding methods and categorizes several existing random-walk based methods.

Given the plethora of network representation learning methods, Gurukar et al. [36] performed an experimental analysis of the popular network representation learning methods to understand the scientific progress in this field. They found that if one tunes the parameter of the Deepwalk method [11] it performs in a competitive manner on both node classification and link prediction tasks. Given the competitive nature of Deepwalk, we select it as one of the approaches to learn meaningful representation of individuals and locations. We also select LINE [12] as one of the approaches for representation learning, as it was found to be both efficient (in terms of running time) and effective (in terms of predictive tasks) [36]. We also performed experiments with BiNE method [25], a network representation learning method designed for bipartite networks. These results are presented in the supplementary section (see section "Cluster Analysis: BiNE") along with a rationale for its relatively poor performance. The summaries of the selected methods are also presented in the supplementary section (see section "Methods summary").

Human Mobility Network Representation learning

The human mobility network representation learning model focuses on a form of co-location network constructed from the human mobility dataset. These models learn representations such that one can efficiently perform human mobility-related downstream tasks such as location prediction [37], location recommendation [38], and travel time estimation [39]. LBSN2vec [37] focuses on Location-Based Social Networks to study user mobility and their social relationships using a hypergraph-based random walk approach to learn user and location embeddings. However, such an approach requires the social network of users, which is not always available.

Location2vec [40] collects the Geo-tagged tweets to learn the location representation and employ skip-gram model [41] on the collected corpus. The representations of Point of Interest (POI) are learned by Yan et al. [38] by proposing a novel method of training corpus generation based on augmented spatial contexts for word2vec model [41]. Note that both Location2vec [40] and the approach by Yan et al. [38] focus on only learning representations of locations and not individuals. Hence, we focus on the following two approaches – one based on Latent Dirichlet Allocation [42] and another based on the sequence of location visits (LocationTrails) – to learn representations of both individuals and locations. The summaries of these selected approaches are present in the supplementary section (see section "Methods summary").

Clustering representations for community assignment

The learned representations of adolescents can be clustered with any off-the-shelf clustering algorithm. The adolescents belonging to the same cluster are then assigned to the same community. In this work, we present the results with Gaussian Mixture Models (GMMs) [10] clustering method. However, we have also experimented with other clustering methods such as K-means [9], and Bisecting K-means [43] and found the results obtained to be consistent with GMMs. GMMs are probabilistic models that assume the data is generated from a mixture of Gaussians with unknown parameters where the parameters are identified with the Expectation-Maximization (EM) algorithm. The output of GMMs is the community-membership probability matrix [2] that contains the probability of an adolescent i belonging to a cluster (community) k. The adolescent is assigned the community that has the highest probability in the community-membership matrix. We utilize GMMs on the representations learned by Deepwalk, LINE, and LocationTrails. Xi et al. [1], on the other hand, directly learn the community-membership affiliation probabilities via the Latent Dirichlet process.

Network Science Methods for Community Identification

The network science methods for identifying communities in both homogeneous and bipartite networks rely on pre-defined community metrics such as normalized cuts [44, 45] or ratio cuts [46, 47]. We consider two popular community identification methods: Metis [26] and Graclus [27]. These methods are multi-level algorithms and consist of three phases: i) coarsening phase in which graph is repeatedly transformed into smaller graphs by combining set of nodes and their corresponding edges, ii) base-clustering phase in which clustering is performed on the coarsest graph. Here, clustering is efficient due to the small size of the coarsest graph and the ability of the coarsest graph to capture the global structure of the graph [48], and iii) refinement phase in which identified clusters are propagated to the larger graphs till the clusters are identified for the input graph. We also performed experiments with a network science method BRIM [29] that is designed for bipartite networks. However, we found that BRIM performs poorly (like BINE) on our dataset. Hence, we do not include BRIM in our analysis. The readers are encouraged to refer to

[2] The clustering output from GMMs is a probability vector - similar to the approach utilized by Xi et al.[1] – another reason for using GMMs to cluster users in our study.

the papers for the detailed algorithm. We apply these methods to our undirected co-location network and analyze the identified adolescents clusters.

Method's Hyperparameters

For all the experiments, we tune the parameters of the methods Deepwalk (walk length = [10, 20, 40], number of walks = [40, 80], context window = [3, 5, 10]), LINE (negative samples = [3, 10], number of samples = [5 billion, 10 billion]), LDA (Gibbs: number of iterations = [10,000, 100,000]), Metis (cut objectives=['normalized cut', 'volume']), and Graclus (cut objectives=['normalized cut', 'ratio association']), and report the best observed results. Note that the mobility pattern related inferences drawn for the methods are consistent across hyper-parameters (more details in the supplementary section "Cluster Analysis: Hyper-parameter results"). We have also included a map of Columbus, Ohio and map of frequently mentioned regions in the supplementary (see Figure 1 and Figure 2) to help the reader locate the neighborhoods referenced in the analysis.

Ground Truth

Precise ground truth for our study is not available. We note that the lack of ground truth is a common problem in community discovery literature (see Hennig [49] for a detailed discussion). Often, the ground truth is ill-defined. Hennig echos this point as "In most cluster analysis literature, however, explanations of what 'true' or 'real' clusters are, are rather hand-waving". The deficiencies in the current clustering evaluation are also pointed out by Von Luxburg et al. [50]. They point out that "whether a clustering of a particular data set is good or bad cannot be evaluated without taking into account what we want to do with the clustering once we have it.". In this work, we want to study human mobility with the help of clustering, hence we rely on existing studies on human mobility (Xi et al. [1], Browning et al. [8]) as well as the sociological studies to assess clustering quality [51, 52, 53]. We describe the sociological studies in the next section.

Sociological studies on the activity profiles

To access the quality of clustering, we would like to bring forth two sets of sociological findings. The first set of findings is from the "activity space" literature in which individuals' activity locations (within and beyond the neighborhood) are the focus of measurement. Studies in this literature have found that many activity locations lie outside of the individual's residential neighborhood unit. For instance, Basta et al. [51] found that the adolescents spent 70% of the non-home time outside their residential neighborhood. Sastry et al. [52] found that only 16% of individuals' routine grocery stores and only 12% of individuals' places of worship lie in their residential neighborhood. Our own findings from the AHDC study suggest that youth spend about 6% of their waking-time in their neighborhood but not at home, 60% at home, and 34% outside their home neighborhood [8]. These studies offer evidence that the clusters of adolescents identified based on their activity locations should not always be residentially proximate – it is not necessary that adolescents who reside in the same neighborhood will share the same cluster, provided they are clustered based on their activity locations.

The second set of findings is drawn from research examining mobility for the purpose of accessing organizational resources [3]. Small and McDermott [53] found that as the proportion of blacks in the neighborhood increases, the number of establishments decreases. In analyses of the AHDC data, Browning et al (2021) find that segregated, higher poverty neighborhoods had fewer schools present within the neighborhood, indicating that youth from these neighborhoods are more likely to be regularly traveling outside the neighborhood to reach school locations. AHDC data indicate that black youth residing in high proportion black neighborhoods encountered more heterogeneous exposures to neighborhood racial composition than other youth and spend a nontrivial proportion of their time in *low* proportion black neighborhoods, largely in the context of organizational resource seeking [8]. Therefore, we expect that for adolescents residing in black-dominated neighborhoods, the probability of falling in residentially proximate clusters will be lower. Moreover, adolescents who reside in the same black-dominated neighborhood will have a higher probability of not sharing the same cluster, provided they are clustered based on their activity profiles.

Neighborhood nomenclature: We collect demographic information on neighborhoods from 2009-2013 American Community Survey data. A neighborhood is considered to be dominated by ethnicity if its percent population is higher than 70%. A neighborhood is considered advantaged if the poverty index is lower than 20% and is considered disadvantaged if the poverty index is greater than 40% [54].

Results

In this section, we evaluate the efficacy of the methods to identify communities [4] on the *finer-grained co-location network*. Next, we perform experiments to study if the identified communities can help in understanding the neighborhood's functioning.

Community Analysis

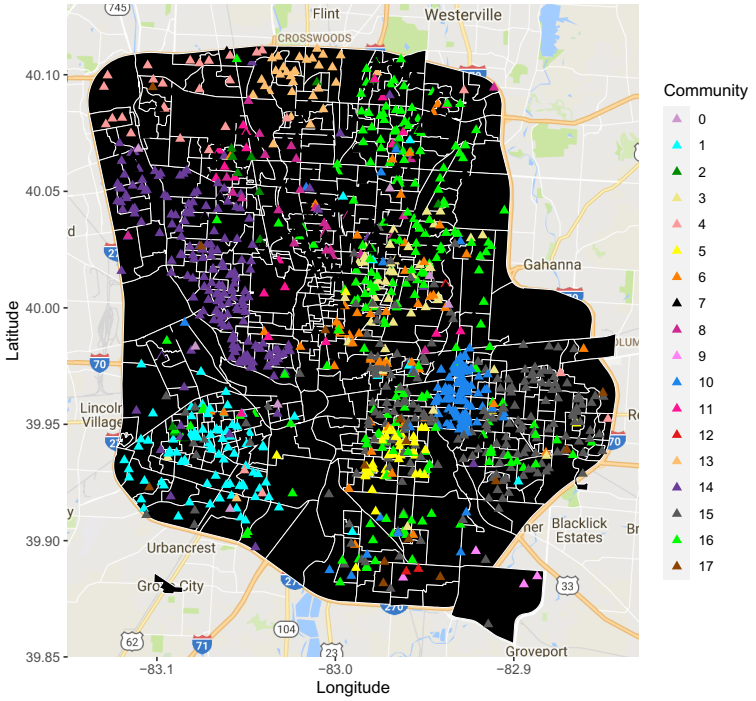
In this section, we perform the community analysis of the adolescent representations learned by all the selected representation learning methods. We render the identified adolescent communities on the Columbus map, where each adolescent is represented through their approximate home location. We select the number of communities to be 18 – similar to the one reported in Xi et al. [1] – and also observe the perplexity metric [42] value with 18 number of communities to be one of the lowest. The identified communities for Deepwalk, LINE, LocationTrails, LDA (Xi et al. [1]), Metis and Graclus are shown in Figure 5a, Figure 5b, Figure 6b, Figure 6a, Figure 7a and Figure 7b respectively. Next, we analyze the identified communities from a sociological lens.

Qualitative Holistic Analysis of Results

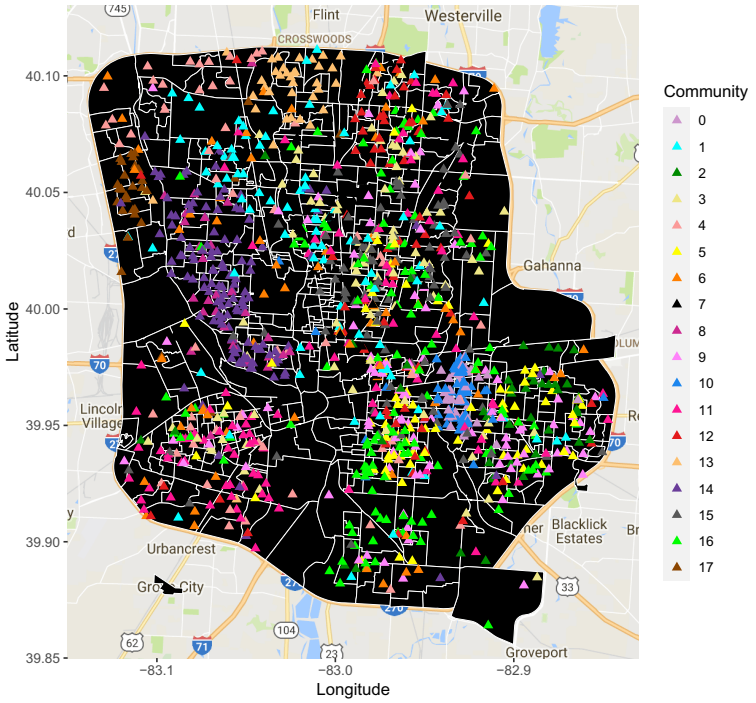
We observe that in white-dominated neighborhoods the evaluated methods often identify residentially proximate communities (refer Figure 1). For instance, we observe that all methods identify a community present at Bexley, Ohio (community

[3] Organizational resources refers to the establishments which have a physical location and offer services or sells goods essential to day-to-day living

[4] We use the term community and cluster interchangeably.

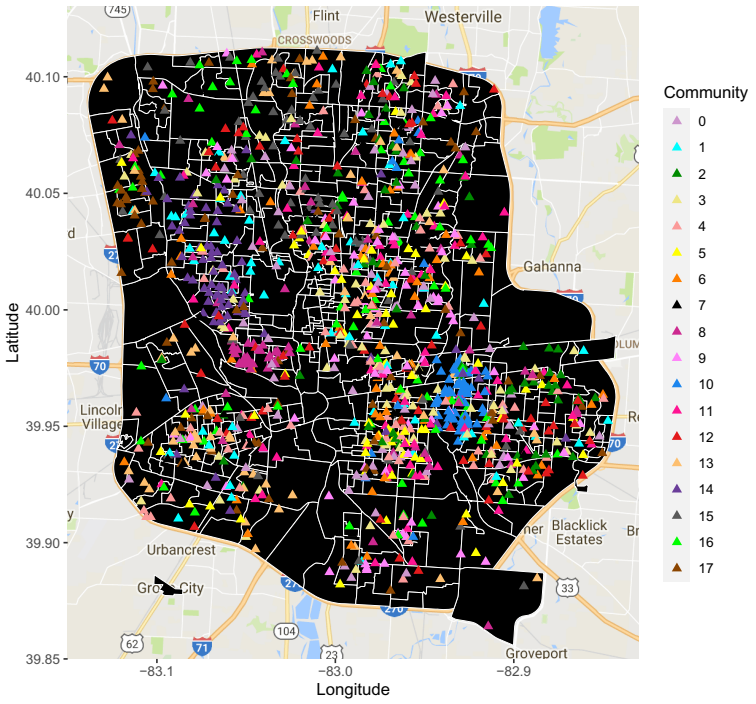


(a) Deepwalk

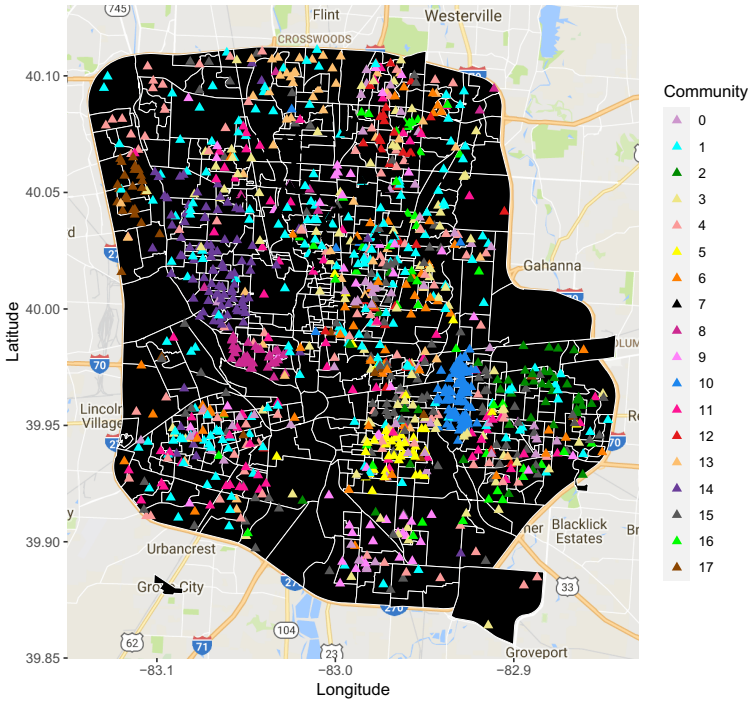


(b) LINE

Figure 5: Number of clusters = 18. (Home location anonymized)

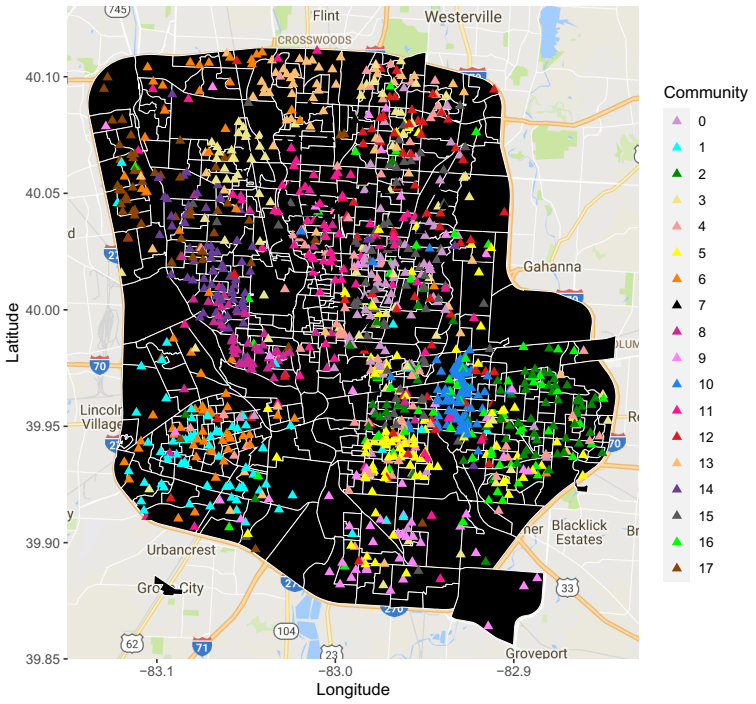


(a) LDA

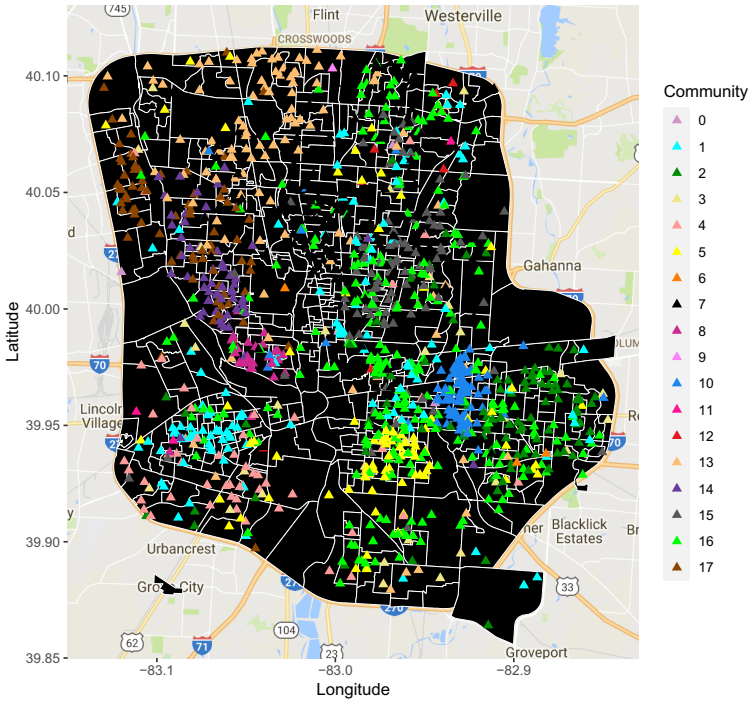


(b) LocationTrails

Figure 6: Number of clusters = 18. (Home location anonymized)



(a) Metis



(b) Graclus

Figure 7: Number of clusters = 18. (Home location anonymized)

number: 10, color: blue). Bexley is a white-dominated area (86.5% of its population is white). The median household income of its residents is double than that of residents living in Columbus city. Bexley is also rich in organizational resources and was historically considered a relatively insular community given its spatial embeddedness in a largely lower-income context. The emergence of the Bexley community shows that many of its residents share the same activity profiles, and this might be due to the abundance of organizational resources (an advantaged neighborhood). Moreover, a few white-dominated neighborhoods such as Upper Arlington, Grandview Heights, and Worthington are commonly identified by Deepwalk, LINE, LocationTrails, Metis, and Graclus.

A few of the methods (Deepwalk, LINE, Metis, and Graclus) that rely solely on the graph structure place adolescents in the same community if they reside in the same black-dominated neighborhoods (such as Near East Side (Census Tract 29 and 36, Franklin, OH) and Milo Grogan (Census Tract 15 and 23, Franklin, OH)). This result does not align well with existing sociological studies [51, 52, 8, 53]. These studies mention that the lack of organizational resources (grocery stores, schools) in black-dominated neighborhoods result in adolescents spending a nontrivial proportion of their time outside of their residential neighborhoods and they encounter more heterogeneous exposure to neighborhood racial composition than other adolescent [8]. This often results in dissimilar activity profiles among adolescents residing in these disadvantaged neighborhoods. Hence, it is surprising that few methods (Deepwalk, LINE, Metis, and Graclus) identify residentially proximate communities in black-dominated neighborhoods. LocationTrails, which relies on the sequence of locations visited by the adolescents, does not identify residentially proximate communities in black-dominated neighborhoods. We present a detailed community analysis of each method in the next few sections.

Community Analysis: LocationTrails

The communities identified by LocationTrails on the *finer-grained co-location network* are consistent with the ones identified by the peer reviewed study done by Xi et al. [1] on the AHDC *coarser-grained co-location network* constructed using a structured data collection approach. Specifically, we observe that LocationTrails places adolescents in the same clusters who reside in Grandview Heights (cluster number: 8, color: light green), Upper Arlington (cluster number: 2, color: black), and Worthington (cluster number: 7, color: green). All these regions have more than 90% white residents, and the median household income of the residents in these regions is double that of residents living in Columbus. These communities share similar characteristics as that of Bexley, however, Deepwalk, LINE, and LDA methods are unable to find these communities. For the adolescents living in the black-dominated neighborhoods, LocationTrails place them in different communities. Specifically, the adolescents who reside in Near East Side (Census Tract 29 and 36, Franklin, OH), Milo Grogan (Census Tract 15 and 23, Franklin, OH) are placed in different communities. The median household income of residents in these regions is less than that of residents living in Columbus. The adolescents in these disadvantaged neighborhoods need to travel further, on average, to access organizational resources and have few common activity profiles. Therefore, LocationTrails assigned them to different communities.

Community Analysis: Deepwalk and LINE

From Figure 5a, we observe that Deepwalk and LINE identify communities that are often residentially proximate – adolescents who reside in the same neighborhood often share the same communities. The identified residentially proximate communities are present for most of the neighborhoods (both white-dominated and black-dominated). This result runs counter to expectations in that residentially proximate communities are less likely to occur in high poverty neighborhoods. As mentioned previously, youth from high poverty neighborhoods often spend a nontrivial proportion of their time outside of their residential neighborhoods and encounter more heterogeneous exposure to neighborhood racial composition than other youth (in order to seek organizationally-based resources) [8]. This often results in dissimilar activity profiles among youth residing in these disadvantaged neighborhoods. Drilling down on the raw activity profiles of individuals in this community, we find that they do indeed have activity profiles that differ and are quite heterogeneous. The results observed here suggest that LINE and Deepwalk are pre-disposed (biased) to identifying residentially proximate neighborhoods.

The reason both Deepwalk and LINE identify residentially proximate communities even for the segregated high poverty neighborhoods can be explained as follows. Both these methods rely solely on the structure of the graph to learn the node representations. Deepwalk relies on the random walks on the co-location network, while LINE relies on both explicit (first-order proximity) and implicit (second-order proximity) connectivity between nodes to learn the node representations. Hence, if two adolescents residing in the same neighborhood visit few common locations (e.g. local stores) present in that neighborhood, these methods would put a high constraint on learning similar representations of those adolescents, as there exists an implicit link between those adolescents. The clustering method would then assign these two adolescents in the same cluster as they would have similar representations.

Community Analysis: LDA

From Figure 6a, we observe that LDA identifies clusters at Bexley (cluster number: 10, color: blue) and Upper Arlington (cluster number: 2, color: black). However, it failed to identify clusters in white-dominated, advantaged neighborhoods that were identified by LocationTrails.

Community Analysis: Metis and Graclus

The communities identified by standard network science algorithms Metis [26] and Graclus [27] are shown in Figure 7a and Figure 7b, respectively. We observe that Metis and Graclus identifies clusters that are residentially proximate for both white-dominated and black-dominated neighborhoods. Metis and Graclus clustered adolescents residing in black-dominated neighborhoods such as South Columbus, south of Grandview Heights in the same communities. As mentioned earlier, these clusters are not aligned with the sociological findings mentioned in the section "Sociological studies on the activity profiles".

To summarize, the above analysis of the identified communities suggests that a method that is cognizant to the sequence of locations visited by the adolescents while learning node representations (LocationTrails [13]) is effective in identifying higher-quality communities from the co-location networks.

	Deepwalk	LINE	LocationTrails	LDA	Metis	Graclus
Deepwalk	1.00	0.53	0.47	0.27	0.49	0.51
LINE	0.53	1.00	0.48	0.25	0.44	0.44
LocationTrails	0.47	0.48	1.00	0.25	0.38	0.40
LDA	0.27	0.25	0.25	1.00	0.24	0.21
Metis	0.49	0.44	0.38	0.24	1.00	0.46
Graclus	0.51	0.44	0.40	0.21	0.46	1.00

Table 2: Normalized mutual information between communities identified by methods on white-dominated neighborhoods.

	Deepwalk	LINE	LocationTrails	LDA	Metis	Graclus
Deepwalk	1.00	0.37	0.24	0.18	0.42	0.35
LINE	0.37	1.00	0.33	0.20	0.36	0.27
LocationTrails	0.24	0.33	1.00	0.18	0.26	0.23
LDA	0.18	0.20	0.18	1.00	0.19	0.15
Metis	0.42	0.36	0.26	0.19	1.00	0.36
Graclus	0.35	0.27	0.23	0.15	0.36	1.00

Table 3: Normalized mutual information between communities identified by methods on black-dominated neighborhoods.

Quantitative analysis of the communities

We measure the overlap between the identified communities by the methods using Normalized Mutual Information (NMI) [55]. From the qualitative analysis, we observe that adolescents who reside in white-dominated neighborhoods, often share the same cluster. This clustering pattern is observed across different methods. In our quantitative analysis, we focus on the adolescents who reside in white-dominated neighborhoods. We then identify their clusters with different methods and present the NMI between the identified clusters in Table 2. A similar analysis for adolescents residing in black-dominated neighborhoods are shown in Table 3. We observe the NMI between clusters identified Deepwalk, LINE, LocationTrails, Metis, and Graclus in the white-dominated neighborhood is relatively high. The relatively high NMI coupled with visual analysis of identified clusters suggest that adolescents who reside in white-dominated neighborhoods often share the same cluster. In black-dominated neighborhoods, the NMI value between clusters identified by Deepwalk, LINE, Metis, and Graclus is relatively higher than NMI between these methods and LocationTrails. The relatively high NMI of Deepwalk, LINE, Metis, and Graclus in black-dominated neighborhoods coupled with visual analysis of identified clusters suggest that these methods are identifying clusters even in black-dominated neighborhoods. As mentioned earlier, this suggestion does not align well with existing sociological studies. We will shortly discuss in the context of neighborhood affinity that further amplifies this point. Note that NMI of LDA is relatively lower in both Table 2 and Table 3. The NMI between identified clusters of adolescents residing in all the neighborhoods is shared in the supplementary (see section "Quantitative analysis").

Quantitative Analysis: Neighborhood Affinity

In this section, we quantitatively analyze the communities present in the neighborhoods. Following the literature [1], we consider the census tract as a proxy for neighborhood and compute the percentage of adolescents who reside in a census

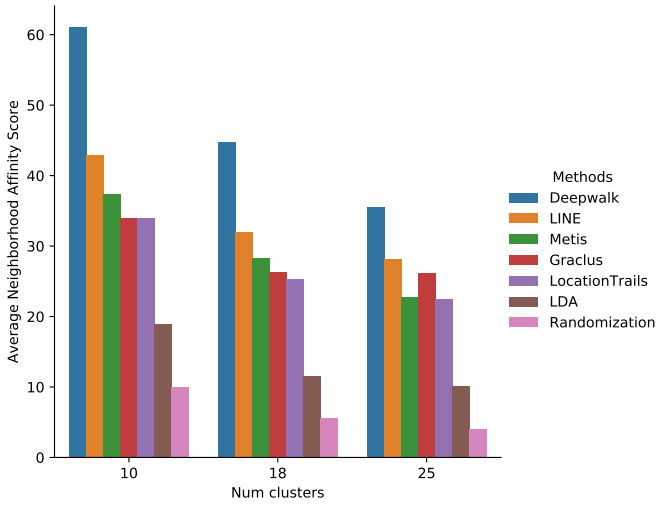


Figure 8: Average Neighborhood Affinity scores.

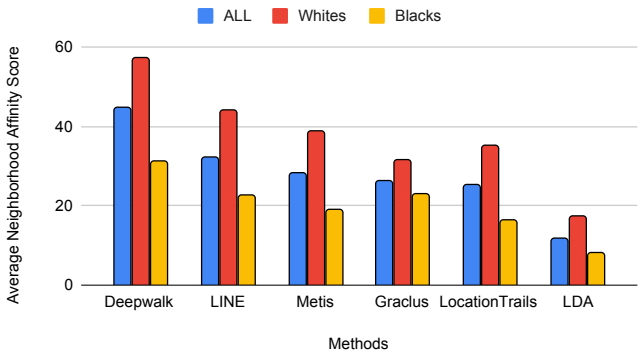


Figure 9: Average neighborhood affinity score in white and black dominated neighborhoods. Number of communities=18.

tract and share the same cluster. The neighborhood affinity of a neighborhood is the probability that two randomly selected adolescents who reside in the same census tract also share the same cluster. Since there are multiple neighborhoods, we report the average neighborhood affinity over all the neighborhoods. While computing the average neighborhood affinity, we filter out the neighborhoods that have fewer than five residents. The average neighborhood affinity scores of different methods are shown in Figure 8. We also report the average neighborhood affinity scores of the Randomization method to know the expected average neighborhood affinity score under uniform community assignment. In Randomization method, we assign adolescents to communities at random in a uniform manner over 1000 times and then compute the average of average neighborhood affinity score.

From Figure 8, we observe that the average neighborhood affinity score of the Deepwalk method is the highest, irrespective of the number of communities. LINE also identifies residentially proximate clusters and has the second highest average

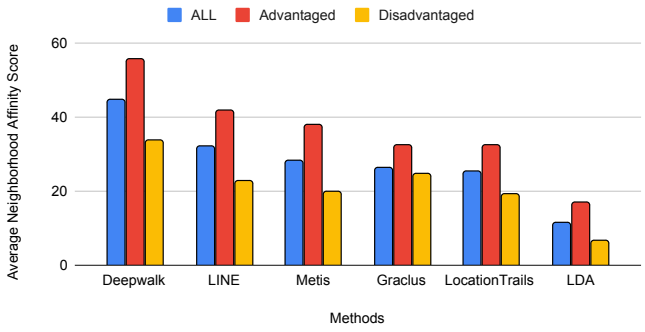


Figure 10: Average neighborhood affinity score in the advantaged (Poverty $\leq$ 20%) and disadvantaged neighborhoods (Poverty $\geq$ 40%) [54]. Number of communities=18.

neighborhood affinity score, irrespective of the number of communities. The high-affinity score of Deepwalk and LINE quantitatively show that they find residentially proximate clusters. LocationTrails affinity score is lower than Deepwalk as LocationTrails places adolescents who reside in black-dominated disadvantaged neighborhoods in different communities. On the other hand, LocationTrails affinity score is higher than LDA, as LocationTrails identifies more clusters with similar characteristics (white-dominated, advantaged neighborhoods). The difference between the average neighborhood affinity score of LDA and Randomization is statistically significant at significance level 0.01 (Z-score $\geq$ 26.0 for all clusters).

Next, we compare the average neighborhood affinity score across white vs. black dominated neighborhoods and advantage vs. disadvantaged neighborhoods. The results are shown in Figure 9 and Figure 10. The average neighborhood affinity score is multiplied by 100. We observe that the average neighborhood affinity score of the adolescents living in the white-dominated neighborhood is higher than that of i) black-dominated neighborhoods and ii) all the neighborhoods, for the four representation learning methods (Deepwalk, LINE, LocationTrails, and LDA). We also observe that the average neighborhood affinity score of the adolescents living in the advantaged neighborhood is higher than that of i) disadvantaged neighborhoods and ii) all the neighborhoods, for the same four representation learning methods. This analysis suggests that white adolescents or adolescents residing in advantaged neighborhoods tend to share more similar activity profiles than their black or disadvantaged neighborhood counterparts. The average neighborhood affinity score of black-dominated/disadvantaged neighborhoods is lower than that of all the neighborhoods. This is because adolescents who reside in these neighborhoods are less likely to have common activity patterns, and this non-commonality in activity patterns might be due to a lack of organizational resources in the black-dominated/disadvantaged neighborhoods.

Drilldown analysis of communities: LocationTrails

In this section, we present a drilldown analysis of communities identified by LocationTrails and provide commentary on the activity profiles of adolescents placed in

a community. We do not disclose the name of the locations that adolescents visit to preserve their privacy. The information about the types of public and private schools in the United States are provided in these articles [56, 57]. The population statistics, economic and political information of Franklin county and the below-mentioned neighborhoods can be found on several web portals [58, 59].

We observe that several communities identified by LocationTrails are residentially proximate. Specifically, Communities 0 and 3 (Upper Arlington), 2 and 17 (Clintonville), 6 (Hillard), 7 (Whitehall), 10 and 15 (Bexley), 13 (East of German village), 14 (Worthington), and 16 (Grandview Heights). Communities 0, 3, 6, 10, 14, 15, and 16 are present in white-dominated neighborhoods with rich organizational resources. Whitehall has a more diverse racial composition (43% white and 39% black residents) and is moderately affluent. Adolescents in residentially proximate Community 13 commonly visit one public magnet high school in East of German village and two public parks within 6 miles from East of German village.

We see that Community 0 and 3 both fall in Upper Arlington, but the adolescents in Community 0 are middle school students and commonly visit two middle schools in Upper Arlington while the adolescents in Community 3 are high school students and commonly visit one high school in Upper Arlington. Essentially, LocationTrails is able to distinguish the middle vs. high school adolescents based on their activity profiles even though their home locations lie in the same neighborhood. We also note that community 10 is extremely cohesive and centered in Bexley (students attending the local high school) whereas community 15 is also largely centered in the Bexley area, but it does have a spread of adolescents with neighborhood homes from largely advantaged neighborhoods in the rest of Franklin county. Drilling down, we observe that the rationale for this is largely driven by the fact that many of the students with shared activity profiles in this cluster attend one of several expensive private schools situated in Bexley. *We point both of these out (two distinct clusters in Upper Arlington and Bexley) as this type of fine-grained analysis is not immediately visible when examining communities identified by the other methods in our study.* Next, we observe that there are a few communities such as Community 5, 8, 11, and 12 in which the home locations of adolescents are spread out over Columbus city. We observe that in these communities, the adolescents often visit schools that have an open enrollment policy and often serve as magnet schools (for STEM, STEAM, and the Arts) or alternative high schools – the policy allows adolescents residing in one school district area to attend schools in another district area. Specifically,

- Adolescents in Community 5 commonly visit one arts middle school near Downtown and a public magnet school near Downtown.
- Adolescents in Community 8 commonly visit three public magnet high schools (one near Clintonville, one north of North Linden and one in Marion-Franklin).
- Adolescents in Community 11 commonly visit one stem school in South Linden and a public-magnet alternative high school in North Linden.
- Adolescents in Community 12 commonly visit two public magnet high schools (one between Worthington and Easton and another near downtown) and one public-magnet alternative high school (with intensive arts curriculum).

Finally, we note that community 4 is spread out over Columbus city as the adolescents in those communities share non-school activities such as a popular swimming club, visiting community centers, malls and church.

Conclusion

We focus on the problem of identifying communities in the co-location networks by using latent representation learning models and community detection methods. Our analysis revealed that the network representation learning model, LocationTrails [13], which relies on the sequence of location visits of adolescents, can identify high-quality communities that are consistent with extant knowledge regarding urban racial and socio-economic differences in neighborhood functioning and activity spaces. We observe that other neural approaches such as Deepwalk [11] and LINE [12] identify residentially proximate clusters – if the adolescents reside in the same or nearby neighborhoods, these methods would often assign them to the same community.

To study the neighborhood functioning of the city, we compare the activity profiles of individuals through an average neighborhood affinity score – the probability of two adolescents sharing the same cluster given that they reside in the same neighborhood. We then compare the average neighborhood affinity score across neighborhoods with different characteristics. Our analysis reveals that the individuals residing in the white-dominated and advantaged neighborhoods have similar activity profiles. Hence, they are assigned to the same clusters by most of the models. In contrast, individuals residing in black-dominated and disadvantaged neighborhoods are often assigned to different clusters. This is because individuals residing in black-dominated/disadvantaged neighborhoods encounter more heterogeneous exposures to neighborhood racial composition than other individuals and spend a nontrivial proportion of their time in *low* proportion black/disadvantaged neighborhoods [8], largely in the context of organizational resource seeking, thereby resulting in dissimilar activity profiles.

Abbreviations

NRL- Network Representation Learning, AHDC - Adolescent Health and Development in Context.

Declarations

Availability of data and materials

AHDC data will be deposited to Inter-university Consortium for Political and Social Research (ICPSR) [60] in publicly available and restricted access forms, beginning in Summer 2022. The location data needed to construct the co-location networks presented in the submission will only be available in the restricted access version of the data. The exact geographic coordinates cannot be released publicly due to concerns of participant privacy and maintenance of data confidentiality. Qualified researchers will be able to submit an application to ICPSR for access.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the National Institute on Drug Abuse (R01DA032371), Eunice Kennedy Shriver National Institute of Child Health and Development (R01HD088545; OSU P2CHD058484; UT Austin P2CHD042849), NSF CNS 2112471, and the William T. Grant Foundation.

Authors' contributions

CB, CC, and SP conceived and supervised the study. SG constructed the co-location network from cleaned data supervised by the AHDC data collection team, including BB, CB, and CC. SG, SP, and CC setup the experimental design for the methodology. SG performed the experiments and an initial analysis of the results. CC, SP and CB provided substantive interpretations of the experimental findings that were reviewed by all authors. SG drafted the initial manuscript. All authors edited and refined the manuscript and approved the final version.

Acknowledgements

We would like to acknowledge Dr. Wenna Xi for sharing the source code Xi et al. [1]. We would like to thank Goonmeet Bajaj for providing valuable feedback on the initial draft.

Ethics approval and consent to participate

The study design and procedures were approved by the institutional review board at the authors' university before fieldwork began. Written parental permission and youth assent to participate in the study was obtained by interviewers prior to the beginning of the initial in-home interview.

Author details

[1]Computer Science and Engineering, The Ohio State University, Columbus, USA. [2]Institute for Population Research, The Ohio State University, Columbus, USA. [3]Sociology, The Ohio State University, Columbus, USA. [4]Statistics and Data Sciences, The University of Texas at Austin, Austin, USA.

References

1. Xi, W., Calder, C.A., Browning, C.R.: Beyond activity space: Detecting communities in ecological networks. Annals of the American Association of Geographers **110**(6), 1787–1806 (2020)
2. Sampson, R.J., Raudenbush, S.W., Earls, F.: Neighborhoods and violent crime: A multilevel study of collective efficacy. science **277**(5328), 918–924 (1997)
3. Zhong, C., Arisona, S.M., Huang, X., Batty, M., Schmitt, G.: Detecting the dynamics of urban structure through spatial network analysis. International Journal of Geographical Information Science **28**(11), 2178–2199 (2014)
4. He, M., Glasser, J., Pritchard, N., Bhamidi, S., Kaza, N.: Demarcating geographic regions using community detection in commuting networks with significant self-loops. PloS one **15**(4), 0230941 (2020)
5. Fujishima, S., Fujiwara, N., Akiyama, Y., Shibasaki, R., Sakuramachi, R.: The size distribution of 'cities' delineated with a network theory-based method and mobile phone gps data. International Journal of Economic Theory **16**(1), 38–50 (2020)
6. Sastry, N., Ghosh-Dastidar, B., Adams, J., Pebley, A.R.: The design of a multilevel survey of children, families, and communities: The los angeles family and neighborhood survey. Social Science Research **35**(4), 1000–1024 (2006)
7. Christopher Browning: Adolescent Health and Development in Context. https://sociology.osu.edu/browning-adolescent-health-and-development-context
8. Browning, C.R., Calder, C.A., Boettner, B., Tarrence, J., Khan, K., Soller, B., Ford, J.L.: Neighborhoods, activity spaces, and the span of adolescent exposures. American Sociological Review **86**(2), 201–233 (2021)
9. Bishop, C.M.: Pattern recognition. Machine learning **128**(9) (2006)
10. Reynolds, D.A.: Gaussian mixture models. Encyclopedia of biometrics **741**, 659–663 (2009)
11. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710 (2014)
12. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1067–1077 (2015)
13. Gurukar, S., Parthasarathy, S., Ramnath, R., Calder, C., Moosavi, S.: Locationtrails: a federated approach to learning location embeddings. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 377–384 (2021)
14. Quillian, L.: Why Is Black–White Residential Segregation So Persistent?: Evidence on Three Theories from Migration Data. Social Science Research **31**(2), 197–229 (2002). doi:10.1006/ssre.2001.0726. Accessed 2022-03-29
15. DeLuca, S., Rosenbaum, J.E.: If low-income blacks are given a chance to live in white neighborhoods, will they stay? Examining mobility patterns in a quasi-experimental program with administrative data. Housing Policy Debate **14**(3), 305–345 (2003). doi:10.1080/10511482.2003.9521479. Publisher: Routledge _eprint: https://doi.org/10.1080/10511482.2003.9521479. Accessed 2021-03-18
16. Data USA, Franklin County, OH. https://datausa.io/profile/geo/franklin-county-oh. Accessed: 2021-10-18
17. U.S. Census Bureau, Franklin County, OH, QuickFacts. https://www.census.gov/quickfacts/fact/table/franklincityohio,US/PST045219. Accessed: 2021-10-18 (2021)
18. Browning, C.R., Pinchak, N.P., Calder, C.A.: Human mobility and crime: Theoretical approaches and novel data collection strategies. Annual Review of Criminology **4**, 99–123 (2021)
19. Brown, R.H., Barram, D.J.: Geographic areas reference manual (1994)
20. Kirchner, T.R., Shiffman, S.: Spatio-temporal determinants of mental health and well-being: advances in geographically-explicit ecological momentary assessment (gema). Social psychiatry and psychiatric epidemiology **51**(9), 1211–1223 (2016)
21. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one **9**(6), 98679 (2014)
22. Modsching, M., Kramer, R., ten Hagen, K.: Field trial on gps accuracy in a medium size city: The influence of built-up. In: 3rd Workshop on Positioning, Navigation and Communication, vol. 2006, pp. 209–218 (2006)
23. Shareck, M., Kestens, Y., Gauvin, L.: Examining the spatial congruence between data obtained with a novel activity location questionnaire, continuous gps tracking, and prompted recall surveys. International journal of health geographics **12**(1), 1–10 (2013)
24. Boettner, B., Browning, C.R., Calder, C.A.: Feasibility and validity of geographically explicit ecological momentary assessment with recall-aided space-time budgets. Journal of Research on Adolescence **29**(3), 627–645 (2019)
25. Gao, M., Chen, L., He, X., Zhou, A.: Bine: Bipartite network embedding. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 715–724 (2018)
26. Karypis, G., Kumar, V.: Metis: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices (1997)
27. Dhillon, I.S., Guan, Y., Kulis, B.: Weighted graph cuts without eigenvectors a multilevel approach. IEEE transactions on pattern analysis and machine intelligence **29**(11), 1944–1957 (2007)

28. Satuluri, V., Parthasarathy, S.: Scalable graph clustering using stochastic flows: applications to community discovery. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 737–746 (2009)
29. Barber, M.J.: Modularity and community detection in bipartite networks. Physical Review E **76**(6), 066102 (2007)
30. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation **15**(6), 1373–1396 (2003)
31. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. NeurIPS'13 (2013)
32. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)
33. Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., Tang, J.: Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In: WSDM (2018)
34. Tang, J., Qu, M., Mei, Q.: Pte: Predictive text embedding through large-scale heterogeneous text networks. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1165–1174 (2015)
35. Huang, Z., Silva, A., Singh, A.: A broader picture of random-walk based graph embedding. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 685–695 (2021)
36. Gurukar, S., Vijayan, P., Srinivasan, A., Bajaj, G., Cai, C., Keymanesh, M., Kumar, S., Maneriker, P., Mitra, A., Patel, V., et al.: Network representation learning: Consolidation and renewed bearing. arXiv preprint arXiv:1905.00987 (May, 2019)
37. Yang, D., Qu, B., Yang, J., Cudre-Mauroux, P.: Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach. In: TheWeb (2019)
38. Yan, B., *et al.*: From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In: SIGSPATIAL (2017)
39. Derrow-Pinion, A., She, J., Wong, D., Lange, O., Hester, T., Perez, L., Nunkesser, M., Lee, S., Guo, X., Wiltshire, B., et al.: Eta prediction with graph neural networks in google maps. arXiv preprint arXiv:2108.11482 (2021)
40. Shoji, Y., Takahashi, K., Dürst, M.J., Yamamoto, Y., Ohshima, H.: Location2vec: Generating distributed representation of location by using geo-tagged microblog posts. In: International Conference on Social Informatics, pp. 261–270 (2018). Springer
41. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NeurIPS (2013)
42. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
43. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques (2000)
44. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22**(8), 888–905 (2000)
45. Zha, H., He, X., Ding, C., Simon, H., Gu, M.: Bipartite graph partitioning and data clustering. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 25–32 (2001)
46. Chan, P.K., Schlag, M.D., Zien, J.Y.: Spectral k-way ratio-cut partitioning and clustering. IEEE Transactions on computer-aided design of integrated circuits and systems **13**(9), 1088–1096 (1994)
47. Billionnet, A.: Solving a cut problem in bipartite graphs by linear programming: Application to a forest management problem. Applied mathematical modelling **34**(4), 1042–1050 (2010)
48. Liang, J., Gurukar, S., Parthasarathy, S.: Mile: A multi-level framework for scalable graph embedding. arXiv preprint arXiv:1802.09612 (2018)
49. Hennig, C.: What are the true clusters? Pattern Recognition Letters **64**, 53–62 (2015)
50. Von Luxburg, U., Williamson, R.C., Guyon, I.: Clustering: Science or art? In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 65–79 (2012). JMLR Workshop and Conference Proceedings
51. Basta, L.A., Richmond, T.S., Wiebe, D.J.: Neighborhoods, daily activities, and measuring health risks experienced in urban environments. Social science & medicine **71**(11), 1943–1950 (2010)
52. Sastry, N., Pebley, A., Zonta, M.: Neighborhood definitions and the spatial dimension of daily life in los angeles. UCLA CCPR Population Working Papers (2004)
53. Small, M.L., McDermott, M.: The presence of organizational resources in poor urban neighborhoods: An analysis of average and contextual effects. Social Forces **84**(3), 1697–1724 (2006)
54. Jargowsky, P.A.: Concentration of Poverty in the New Millennium: Changes in Prevalence, Composition, and Location of High-Poverty Neighborhoods. Technical report, Century Foundation and the Center for Urban Research and Education, New York and Camden, NJ (2013). http://tcf.org/bookstore/detail/concentration-of-poverty-in-the-new-millennium
55. Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. IEEE Transactions on neural networks **20**(2), 189–201 (2009)
56. Grand Canyon University: Types of Public and Private Schools to Consider in the U.S. https://www.gcu.edu/blog/teaching-school-administration/public-and-private-schools
57. Public School Review: A Quick Guide to U.S. Public and Private School Options. https://www.publicschoolreview.com/blog/a-quick-guide-to-us-public-and-private-school-options
58. Data USA: Data USA. https://datausa.io/
59. US Census: Quick Facts. https://www.census.gov/quickfacts
60. ICPSR: Inter-university Consortium for Political and Social Research. https://www.icpsr.umich.edu/web/pages/