**Supplementary Information for Census Block Group Clustering**

**Methods**

**Data source**

Our community networks were developed using movement data from SafeGraph ([www.safegraph.com](http://www.safegraph.com)), a data company that aggregates anonymized location data from millions of mobile devices to provide insights about movement patterns. SafeGraph collects hourly data of visits to millions of points of interest (POIs) across the US, which include different types of businesses and establishments, such as retail stores, restaurants, schools, healthcare facilities, etc. The visit dataset includes details such as the number of visits aggregated by the home census block groups (CBGs) of the visitors. SafeGraph data have been used in different studies to investigate movement patterns,1,2 especially transmission of COVID-19 in communities.3,4

**Community networks**

Using SafeGraph movement data, a set of community networks were constructed to discover the structure and patterns of connectedness in the population.

*1. Data preparation*

As the spatial scope of this study was the Johns Hopkins Health System catchment area within the Maryland-DC region, the data was filtered for the POIs and CBGs within the study area. For privacy reasons, SafeGraph excludes CBGs for which fewer than five cellphone devices are registered. Furthermore, while SafeGraph categorizes POIs into more than 400 categories, to enable cross-category analyses, we aggregated sub-categories and certain top-categories of SafeGraph into the following nine categories: Retail, Parks, Religious, School, Health, Restaurants, Bars, Other, and Unknown. For this study, we excluded the Health POIs as we are interested in movements outside of the healthcare facilities. Overall, we analyzed visits to 112,207 POIs in 5852 CBGs within the Maryland-DC area from January 2019 to March 2021.

*2. Weekly flow matrices*

We used SafeGraph Weekly Places Patterns datasets to calculate the weekly flow between all pairs of CBGs in the study area. This dataset contains the weekly number of visits for each POI and the home CBGs of the visitors. Using the CBGs of POIs as destinations and the home CBGs of the visitors as origins, we aggregated the visit counts by CBG and calculated the weekly flow matrices for each of the 8 category of POIs (excluding Health):

|  |  |
| --- | --- |
| , | (1) |

where is the weekly total number of visitors from CBG that visited a POI of category in CBG during week . The average weekly flow matrix was then calculated as:

|  |  |
| --- | --- |
| , | (2) |

where is the number of weeks.

*3. Average interaction matrix*

Having calculated the average weekly flow matrix, we calculated the average normalized flow matrix, as follows:

|  |  |
| --- | --- |
| , | (3) |

where represents the average proportion of the visits that the population of CBG make to non-healthcare POIs in CBG in any week (, and is the number of registered cellphone devices (in SafeGraph system) in CBG , respectively.

Then, we calculated the interaction matrix that gives the average weekly probability of interaction between every pair of CBGs ():

|  |  |
| --- | --- |
| *.* | (4) |

where is the set of all CBGs.

Each component, , of the interaction matrix denotes the probability that any random individual from CBGs would contact a random individual from CBG (whether inside CBG , , or any other CBG). In Equation (4), the term denotes the joint probability that two random individuals from CBGs and would go to any CBG , and denotes the probability of interaction. The probability that an individual from CBG contacts an individual from CBG inside CBG depends on the active population mixing of CBG (here, active population refers to the population that actively visit POIs in CBG , and it does not refer to the population of CBG ). From all the population actively visiting POIs in CBG , , only are from CBG , hence the probability of a random interaction with a person from CBG equals .

It is noteworthy that the interaction matrix is not necessarily symmetric because the probability of interaction depends on the proportion of the visits that the populations of two CBGs make to all other CBGs, i.e.:

|  |  |
| --- | --- |
|  | (5) |

which implies for .

*4. Community detection*

For each CBG in the study area, the corresponding row of the interaction matrix indicates the level of connection with other CBGs in terms of the probability of interaction. We applied a set of community detection methods to discover clusters of CBGs in which residents are more likely to interact with each other rather than with people from other clusters.

First, we created the mobility network based on the average CBG interaction matrix, defined as a directed and edge-weighted graph , in which vertices () are the CBGs, edges () denote CBGs with any positive level of population interaction, and weights () are the probabilities of interaction between populations of pairs of CBGs. The result was a network with 2692 nodes and 164,556 weighted edges. For detecting community structures, we used the greedy Clauset-Newman-Moore5 and the Louvain algorithm,6 which are both based on modularity maximization and among the most commonly used community detection algorithms across different datasets. In case of isolated CBGs, i.e., communities consisting of only one CBG, we modified the resulted community structure by assigning these isolated CBGs to the closest community, identified based on the geodesic distances between the centroids of the CBGs.

We conducted sensitivity analyses on the resolution parameter, often called lambda (), of these algorithms. Gamma is a hyper parameter that sets an arbitrary tradeoff between intra-group and inter-group edges. Increasing the resolution parameter can result in smaller-sized clusters and more complex structures. We incrementally increased gamma from 1 to 2. In addition, as Louvain is a heuristic method, we repeated the analysis for each value of gamma and extracted a set of 100 possible community structures for the study area.

Given the probabilistic nature of community detection algorithms and their sensitivity to the value of the resolution parameter, and in the absence of ground truth, it is difficult to judge the reliability of the resulted community structures. To incorporate this uncertainty into our analysis, we applied the method of community cores7 to identify clusters of CBGs on which the results of the community detection analysis agree most of the time. The community cores methodology aims at finding the tendency (probability) of pairs of nodes to belong to the same cluster across multiple executions of non-deterministic community detection algorithms or either deterministic or probabilistic algorithms with varying hyper parameters.

To identify the community cores, we created an matrix, where is the number of nodes (CBGs), denoted by . Each component of this matrix, , is the fraction of executions in which CBGs and were classified in the same community (note that ). Based on matrix , we create a new undirected weighted network ), where the weight of the edge is . Finally, for any given threshold , the edges with are removed from to obtain the so called α-cores subgraph , which is non-overlapping clusters of CBGs that appear in at least in a fraction of all the resulted community structures.

**CBG Communities**

A sample of how varying resulted community structures using the greedy algorithm is shown in Figure S1. As the resolution parameter increased from 1 to 1.6, the number of detected communities increased from 8 to 13. For the community cores methods, as the threshold increased in value, the number of clusters increased such that for thresholds of 50%, 95%, and 99%, the number of clusters were 3, 74, and 179, respectively Figures S2 and S3..

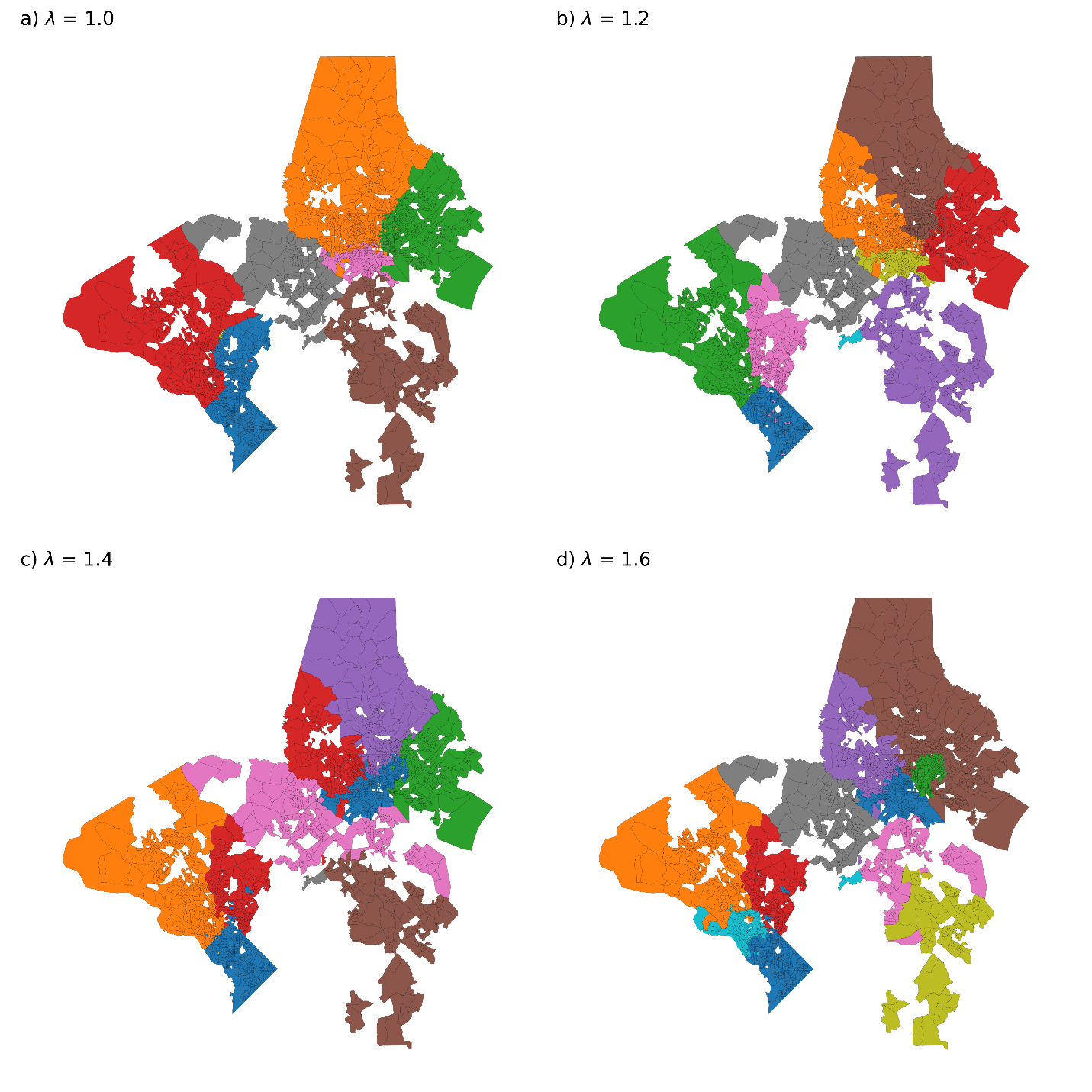


Figure S1. Examples of detected community structures with different resolution parameters.

Chart, line chart

Description automatically generated

Figure S2. Results of community cores methodology

Finally, there were a small number of CBGs that were not included in other identified community structures but rather as single CBG “communities”. This occurred either because (a) CBGs with small numbers of users were excluded from SafeGraph data, or (b) the information from SafeGraph did not highly associate them with other community structures using the methodology. For the purpose of aggregation to identify potential ESBL transmission clusters, these single community CBGs were clustered with the main identified community structures in the final analysis based on the geodesic distance to the nearest cluster. The final clusters used in the analysis are shown in Figure S3.

|  |  |
| --- | --- |
| A) |  |
| B) |  |
| C) |  |

Figure S3: Final community structures. The above community structures for (A) Greedy, (B) Louvain, and (C) community cores, were detected and used to cluster possible community transmission of ESBL organisms.

**Data availability**

Mobile phone mobility data from SafeGraph is freely available for research purposes through the SafeGraph COVID-19 Data Consortium (<https://www.safegraph.com/covid-19-data-consortium>).

**Acknowledgment**

The authors would like to thank SafeGraph for providing their data for COVID-19 related work.

**References**

1. Athey, S., Blei, D., Donnelly, R., Ruiz, F. & Schmidt, T. Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data. *AEA Pap. Proc.* **108**, 64–67 (2018).

2. Keith Chen, M. & Rohla, R. The effect of partisanship and political advertising on close family ties. *Science (80-. ).* **360**, 1020–1024 (2018).

3. Chang, S. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nat. 2020 5897840* **589**, 82–87 (2020).

4. Benzell, S. G., Collis, A. & Nicolaides, C. Rationing social contact during the COVID-19 pandemic: Transmission risk and social benefits of US locations. *Proc. Natl. Acad. Sci.* **117**, 14642–14644 (2020).

5. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* **70**, 6 (2004).

6. Nguyen, L. Van *et al.* Fast unfolding of communities in large networks You may also like The Baxter Q operator of critical dense polymers Alessandro Nigro-A Bayesian fusion model for space-time reconstruction of finely resolved velocities in turbulent flows from low resolution measurements Fast unfolding of communities in large networks. *J. Stat. Mech* 10008 (2008) doi:10.1088/1742-5468/2008/10/P10008.

7. Seifi, M., Junier, I., Rouquier, J. B., Iskrov, S. & Guillaume, J. L. Stable community cores in complex networks. *Stud. Comput. Intell.* **424**, 87–98 (2013).