Community Detection in Real World Complex Mobility Networks

Weicheng Hu, Stella Li, Shanelle Cao, Bohan Hou, TJ Bai, Anton Dahbura
Department of Computer Science
Johns Hopkins University

Abstract

In the past two years, we have observed the coronavirus (COVID-19) pandemic evolve into a public health crisis and challenge all over the world. To answer how the disease spreads, we need to first study how people interact. The demand for community detection in real-world mobility networks then thrived. Community detection in networks has always been an important research topic in computer science. However, in the context of mobility networks, the connections between people are unprecedentedly strong, making community detection in these networks particularly challenging. In this project, we tackle this problem by proposing innovative computer science methods and augmented classical algorithms with modern strategies to discover the community structures in large-scale mobility networks. Based on different evaluation metrics, the baseline algorithms with the best performance are Louvain and Spectral. Some of the visualizations of the partitioning of the Oklahoma are shown.

Introduction

Community detection, also called graph partitioning, is a family of algorithms designed to detect the underlying relationships between nodes in networks. A community is a cluster of nodes that are more closely related to each other than to the nodes outside the cluster. By partitioning the nodes in a graph into several groups, community detection algorithms have been applied to solve many real-world problems, from biology to sociology, depending on the network being investigated.

In our project, in order to investigate and simulate the unique pattern of disease spread in each group of people, we are interested in discovering the community structure behind the dividing the mobility network provided by SafeGraph data into convenience zones where the majority of the people move within the convenience zones, and few people move across zones. As people become increasingly interconnected, such a mobility network is unprecedentedly large in scale, density, and complexity. And because the mobility networks have weighted edges, many of the classical algorithms become either less effective or overwhelmingly computing intensive when dealing with real-world mobility networks.

In our project, various clustering algorithms are implemented and compared against a strong baseline. We proposed a convenience zone algorithm that effective divides real-life data and developed a novel evaluation metric that requires an equilibrium between inter-cluster movements and compactness within each cluster. Furthermore, we developed a more sophisticated clustering model that employs modern machine learning algorithms, which out-performs baseline standard clustering methods such as k-means.

Methods

Our proposed algorithm involves two main sections. We first cluster the datapoints and group them into convenience zones, then, we use the evaluation metric to quantitatively and qualitatively evaluate the clustering results.

I. Algorithms for Partitioning Optimization

Baseline Algorithms. Some of the classical community detection algorithms we have implemented and augmented are Louvain, Spectral, Walktrap, BRIM, and Leiden. Each of them are evaluated with the classical metrics such as modularity, and our proposed loss function.

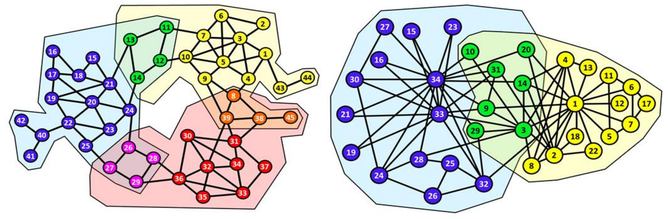

Figure 1. An illustration of community detection in networks[1]

Our Approach. By expressing our loss functions via matrix operation of the one-hot encoding of the partitioning, we are able to make the loss function differentiable, making it possible to optimize the partitioning by strategies such as stochastic gradient descent. We therefore designed a deep neural network (DNN) to optimize the partition starting from randomly initialized clusters. Our approach has the following key advantages:

- **Handling Large-Scale Dataset.** Deep learning methods are known for big data analysis. And our DNN works by manipulating the presence of each edge in the partitioning. Therefore, our approach can easily handle large-scale real-world mobility datasets.

- **Overlapping Clusters.** One unique advantage of using DNN is the overlapping of clusters, which is not allowed by most of the classical algorithms.

II. Evaluation Metric

When partitioning the datapoints into convenience zones, the cost of the partitions, which is the cross-cluster population movement, needs to be minimized, while the clusters need to be maintained at a reasonable size. Therefore, we propose a novel loss function that is sensitive to both requirements.

Novel Loss Function. The loss function consists of two parts: the sum of cross-cluster edges, and the compactness of each cluster. The former part makes sure our demand is fulfilled, and the latter part applies a constraint on the cluster size by avoiding putting loosely connected nodes in the same cluster to lower the inter-cluster movement. The equilibrium of both parts is the optimal partitioning we seek.

$$Loss = \sum_i \left(\sum_j interCluster(e_{ij}) \right) \times compactness(i)$$

where $interCluster(e_{ij})$ is the edges crossing clusters i and j, and $compactness(i)$ is the compactness of cluster i

Baseline Sample Results

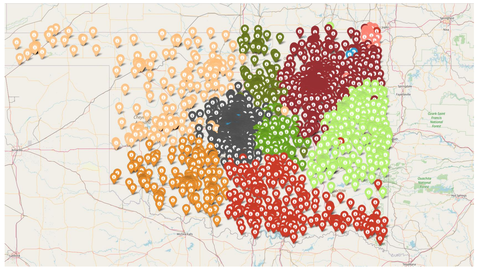

Figure 2. Clustering of CBGs in Oklahoma state using Louvain algorithm. This algorithm is strong in minimizing the inter-cluster movements.

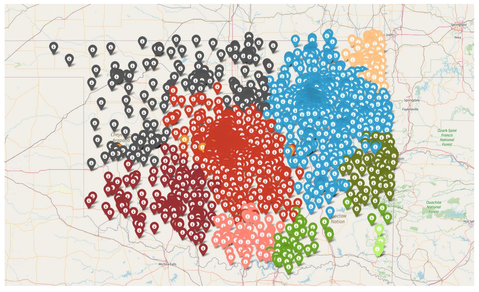

Figure 3. Clustering of CBGs in Oklahoma state using Walktrap algorithm. This algorithm is strong in maximizing the modularity.

Conclusions

It is possible to divide CBGs into clusters with our current objective. Meanwhile, we were able to implement multiple clustering algorithms to get to reasonable convenience zones. With the goal of minimizing inter-cluster population movement, we have developed a viable clustering method to effectively reduce the computation complexity for downstream tasks such as disease modeling.

Future Directions

After obtaining favorable results as summarized above, we aim to further extend out algorithm to better cluster the datapoints. Deepening the neural network could potentially lead to improved performance of the model. A more sophisticated loss function or a Pareto front optimization model are both fruitful areas of future research to better serve our clustering goals. Lastly, the possibility of overlapping convenience zones that better describes real-life population movement can allow for more flexible partitioning and further assist downstream applications.

Acknowledgement

This project and the research behind it would not have been possible without the exceptional support of our mentors: Dr. Kimia Ghobadi, Dr. Claus Aranha, Dr. Ali Madooei, and Dr. Eili Klein.

[1] Whang, J. J. (2015). Overlapping community detection in massive social networks (Doctoral dissertation).