Signed weighted graph community detection for spatial correlation in earthquake intensity measurements networks

Yilin Chen
yilinc2@stanford.edu

Yang Wang
leonwang@stanford.edu

Hongtao Sun
s3sunht@stanford.edu

Abstract

This project is to analyze different algorithms' performance on spatial community detection on signed weighted graph. The weight of link in the graph represents how much correlation coefficient between two nodes deviates from the expected correlation coefficient in the graph, where positive sign of link indicates that the pair of nodes has higher correlation and vice versa. We look to explore two community detection methods, namely, modified spectral clustering and modified Louvain, to identify areas and stations that have unusually high or low correlations. Adjustments are made on both algorithms to accommodate weighted signed graphs. We evaluate the performance of the algorithms by visualizing the spatial location of the detected communities and comparing them with geology map, because the graph is built of earthquake intensity data which have been well studied by seismologist and have been proved that it's highly dependent on geological condition. We also perform simulation based on detected communities using Stochastic Block Model (SBM) to further validate our results. Many potential applications can derives from this simulation.

1. Introduction

Spatial networks appear in many different fields, such as seismic networks, road networks, mobile networks and flight connections. In many applications, properties of nodes that are spatially closer have a greater probability of being correlated with nearby nodes. In the case of earthquake measurements networks, nodes represent different stations and edges represent positive and negative deviation of correlations between stations' earthquake intensity measurements from the expected correlations. Note that edges are weighted and signed to represent the strength of the correlation deviation.

The standing empirical model states that this correlation between stations is a function of distance only. However, reality is far more complicated than this. We look to utilize community detection methods to identify areas and stations that have unusually high or low correlations. Successfully detecting communities of earthquake stations allows as to uncover underlying reasons for measurements. Moreover, simulating earthquake data is of great practical use for both scientific research and civil applications.

This project aims to develop and evaluate two community detection methods that handle weighted and signed networks. We implemented two distinct algorithms to detect communities on signed weighted graphs based on spectral clustering and Louvain algorithm. Using this method, we are able to find the regional communities (i.e., regions that are abnormally higher/lower correlated compared to expected correlation) in earthquake measurements network. Our community detection results coupled with Stochastic Block Model (SBM) provides a new way to simulate spatially correlated earthquake data.

2. Related Work

2.1. Modularity

The common version of community detection tackles graphs that does not have weighted edges. One of the most used techniques in community detection algorithms is to use a quality function called modularity proposed by Newman and Girvan (2004).

The modularity is defined as

$$Q = \frac{1}{2w} \sum_{C \in P} \sum_{i,j, \in C} (A_{ij} - P_{ij}) \tag{1}$$

where $i, j \in C$ is a summation over pairs of nodes i and j belonging to the same community C of partition P, and A is the adjacency matrix and w is the total weight of the network. The most popular choice of P_{ij} proposed by Newman and Girvan (2004) is:

$$P_{ij} = w_i w_j / 2w \tag{2}$$

The weight sum w_i is defined as $w_i = \sum_k w_{ik}$, which is the sum of edge weights around node i. The total weights $w = \sum_k w_k = \sum_i \sum_j w_{ij}$. Larger modularity indicates better partitioning since it deviates more from the null case where the edges are generated randomly. However, maximizing modularity score is a NP-hard problem, and it is usually approximately solved by the Louvain algorithm (Blondel et al. (2008)).

The above notion generalizes naturally to positive edge weights. However, according to Gomez, Jensen, and Arenas (2008), naively plugging signed weights into the equations would result in mistakes. The authors thus generalized the modularity defined above and refined it into two parts. We will extend his method and use it in our proposed approach.

2.2. Spectral Clustering

Spectral clustering is a popular method for community detection tasks. Variations of spectral clustering usually solve a form of graph cut problem by exploiting the spectral properties of the adjacency matrix of the graph. However, the original versions of spectral clustering does not allow signed graphs. Kunegis et al. (2010) introduced a modified spectral clustering algorithm and provided some properties of the algorithm.

The paper shows that the **dominant** eigenvector of the *Signed Laplacian Matrix* $\bar{L}$ solves the signed ratio cut problem where (some further explanations are provided in section 4)

$$\bar{L} = \bar{D} - A \tag{3}$$

Here A is the signed adjacency matrix of the graph and $\bar{D}_{ii} = \sum_j |A_i j|$ is the modified degree matrix.

Similarly, the **dominant** eigenvector of matrix $D^{-1}A$ solves the signed normalized cut problem.

3. Data Processing

For every pair of stations (j, k), we select all earthquakes with suitable recordings at both stations, and use equation 4 to calculate the correlation coefficient in ground motion intensity measure $W_{i,k}$.

$$\hat{\rho}(j,k) = \frac{\sum_{i=1}^{n}(\delta W_{i,j} - \delta \bar{W}_{i,j})(\delta W_{i,k} - \delta \bar{W}_{i,k})}{\sqrt{\sum_{i=1}^{n}(\delta W_{i,j} - \delta \bar{W}_{i,j})^2}\sqrt{\sum_{i=1}^{n}(\delta W_{i,k} - \delta \bar{W}_{i,k})^2}} \tag{4}$$

where n is the number of earthquakes with pairs of recordings at the given stations. Figure 1 shows calculated correlation coefficients. An exponential function model is fitted to the averaged correlation coefficients to capture the relationship between the correlation coefficient of nodes and their distance in the graph. This model represents the expected correlation coefficient of a pair of nodes given their geographical distance in the graph. It can be seen that the expected correlation decreases with distance, as expected, although there is significant variation relative to the expected correlation coefficient at individual station pairs.

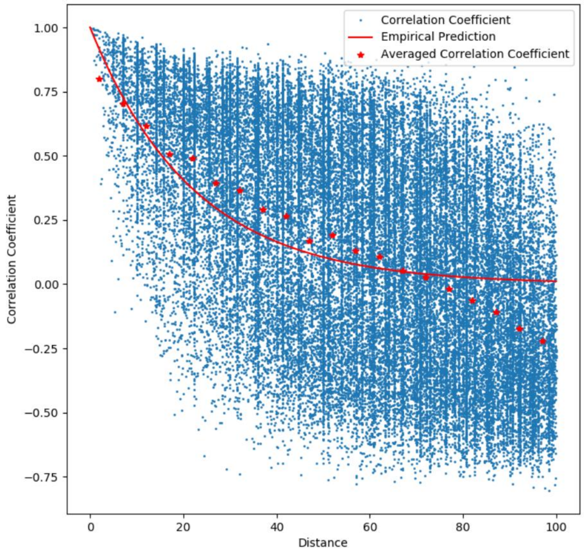

Figure 1: Correlation coefficients of all connected nodes as a function of nodes geographical distance.

We quantify these site-specific deviation of correlations relative to the expected correlation coefficient based on Fisher's z-transformation:

$$z_{\hat{\rho}} = \frac{1}{2}ln(\frac{1+\hat{\rho}}{1-\hat{\rho}}) \tag{5}$$

where $\hat{\rho}$ is the sample correlation coefficient between a pair of nodes. For a sample of observations, $z_{\hat{\rho}}$ is approximately normally distributed with mean $\frac{1}{2}ln(\frac{1+\rho}{1-\rho})$ and standard deviation $\frac{1}{\sqrt{n-3}}$, where ρ is the expected correlation coefficient and n is the number of paired observations.

Then we can define

$$e = (z_{\hat{\rho}} - z_{\rho}) \times \sqrt{n-3} \tag{6}$$

as the measure of correlation deviation. Under the above assumptions, e will follow the standard normal distribution. Therefore, e will be the weighted signed edge in our graph, which quantify the correlation deviation a pair of station relative to the expected correlation correlation in the graph.

Three earthquake datasets at Wellington, Los Angeles and Japan are used to construct the graphs. There are 18 nodes and 118 edges in the Wellington graph, 335 nodes and 42144 edges in the California graph and 382 nodes and 3373 edges in the Japan graph.

4. Technical Approach

4.1. Signed and weighted Spectral Clustering

We use a signed version of spectral clustering proposed by Kunegis et.al for the community detection task Kunegis et al. (2010). The signed weighted adjacency matrix A is defined as usual where A_{ij} is the edge weight between node i and j. The signed degree matrix is defined as:

$$\tilde{D}_{ii} = \sum_i |A_{ij}| \tag{7}$$

The signed Laplacian matrix is then defined as $\tilde{L} = \tilde{D} - A$, and the signed ratio cut between cluster X and Y is

$$\text{SignedRatioCut} = \text{scut}(X, Y)(\frac{1}{|X|} + \frac{1}{|Y|}) \tag{8}$$

where

$$\text{scut}(X, Y) = 2\text{cut}^+(X, Y) + \text{cut}^-(X, X) + \text{cut}^-(Y, X) \tag{9}$$

and

$$\text{cut}^+(X, Y) = \sum_{i \in X, j \in Y} A^+_{ij} \tag{10}$$

$$\text{cut}^-(X, Y) = \sum_{i \in X, j \in Y} A^-_{ij} \tag{11}$$

$$A^+_{i,j} = \max(0, A_{ij}), A^-_{i,j} = \max(0, -A_{ij}) \tag{12}$$

The signed cut $\text{scut}(X, Y)$ counts the number of positve edges that connect X, Y and number of negative edges that remain each of these groups.

It was shown by Kunegis et al. (2010) that the minimization problem for signed ratio cut is equivalent can be solved by finding the smallest eigenvectors of $\tilde{L}$.

A similar result shows that to minimize the signed normalized cut, we need to cluster based on the eigenvectors of $\tilde{D}^{-1}A$. In this project, we implement this algorithm with K-Means clustering on the eigenvectors.

4.2. Signed Louvain Algorithm

Gomez, Jensen, and Arenas (2008) defined the signed graph modularity as:

$$\begin{aligned}\tilde{Q} =& [\frac{1}{2w^+ + 2w^-} \sum_i \sum_j [w_{ij} - (\frac{w_i^+ w_j^+}{2w^+} - \frac{w_i^- w_j^-}{2w^-})] \\ & \times \delta(C_i, C_j)]\end{aligned} \tag{13}$$

where

$$w_{ij} = w^+_{ij} - w^-_{ij} \tag{14}$$

where $w^+_{ij} = \max\{0, w_{ij}\}$, $w^-_{ij} = \max\{0, -w_{ij}\}$, and

$$w_i^+ = \sum_j w^+_{ij}, w_i^- = \sum_j w^-_{ij}. \tag{15}$$

To optimize the modularity, the modularity gain can be calculated as:

$$\begin{aligned}\Delta\tilde{Q}(i \to C) =& \frac{2w^+}{2w^+ + 2w^-}\Delta\tilde{Q}^+ \\ & - \frac{2w^-}{2w^+ + 2w^-}\Delta\tilde{Q}^-\end{aligned} \tag{16}$$

$$\begin{aligned}\Delta\tilde{Q}^+ =& [\frac{\sum_{in^+} + k^+_{i,in}}{2w^+} - (\frac{\sum_{tot^+} + k^+_i}{2w^+})^2] - \\ & [\frac{\sum_{in^+}}{2w^+} - (\frac{\sum_{tot^+}}{2w^+})^2 - (\frac{k^+_i}{2w^+})^2]\end{aligned} \tag{17}$$

$$\begin{aligned}\Delta\tilde{Q}^- =& [\frac{\sum_{in^-} + k^-_{i,in}}{2w^-} - (\frac{\sum_{tot^-} + k^-_i}{2w^-})^2] - \\ & [\frac{\sum_{in^-}}{2w^-} - (\frac{\sum_{tot^-}}{2w^-})^2 - (\frac{k^-_i}{2w^-})^2]\end{aligned} \tag{18}$$

where w^+ and w^- is the sum of the positive/negative weight, $k^+_{i,in}$ and $k^-_{i,in}$ is the sum of positive/negative weights between i and C, k^+ and k^- is the sum of all positive/negative link weights of node k, $\sum_{in^+}$ and $\sum_{in^+}$ is the sum of positive/negative link weights between nodes in C, and $\sum_{tot^+}$ and $\sum_{tot^-}$ is the sum of all positive/negative link weights of nodes in C.

5. Results

We experimented on three datasets from three different places with different geological characteristics. Our signed Louvain algorithm performs better on the Japan dataset but on the other two datasets, spectral clustering obtained results that fits our prior knowledge better.

5.1. Wellington

The geology at south and north Wellington region are different. Intuitively, the community detection performed on this region should be consistent with this geology fact. From figure 3, the black community and white community almost recovered the two communities separated by the gulf.

As we can see from figure 4, Louvain performs relatively poorer than spectral clustering and we end up getting mixed groups that are not exactly mutually exclusive in geographic sense.

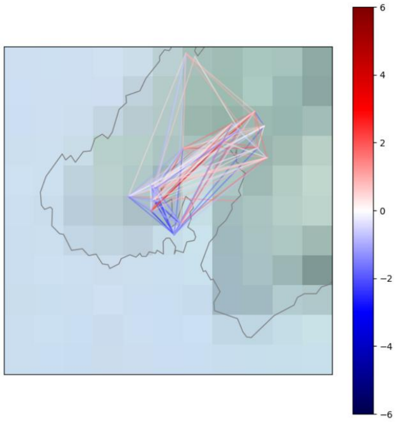

Figure 2: Edge weights in the Wellington graph. The weight of the edges are colored according to the value. Positive weights are displayed in red and negative weights in blue color

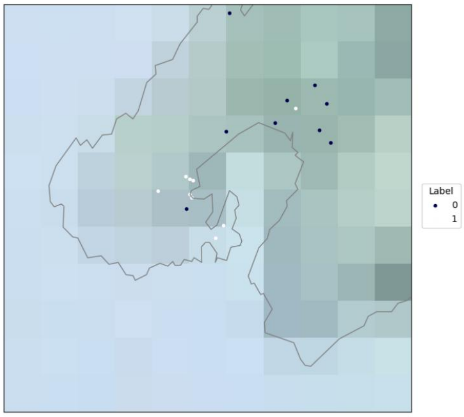

Figure 3: Nodes community assignment in the Wellington graph using spectral clustering

5.2. Los Angeles

For this dataset, we already know for a fact there is a basin at LA county, which can be seen in figure 5.

The original graph can be visualized by figure 6 where the blue edges are relatively low correlations and the red edges are relatively high correlations.

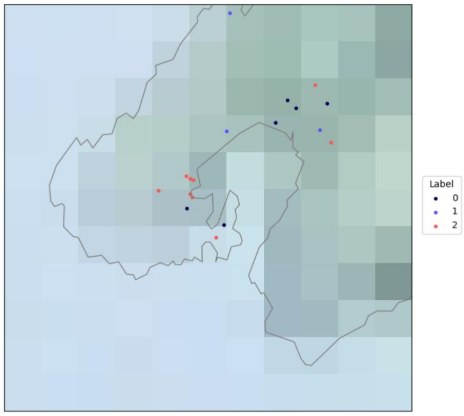

Figure 4: Nodes community assignment in the Wellington graph using Louvain

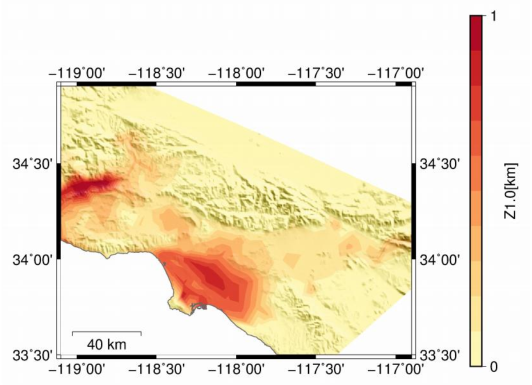

Figure 5: Map of basin depth value in south California region. Data from Small et al. (2017).

For this dataset, the communities detected by spectral clustering match accurately with the geographic geoplogy. As we can see from figure 7 when we set the number of communities to 5, the algorithm identifies three major communities, which correspond the basin and the mountainous region outside. When we increase the number of communities, we can see from figure 8, the algorithm is also able to identify more precise community, and it still make geological sense.

Signed Louvain algorithm is able to detect two communities that are separated from the middle. However, signed Louvain stops before it further identifies any other geological structures such as the basin. Therefore, in this case, the signed Louvain is less flexible and provides less insight into the data.

5.3. Japan

The third dataset we have is the earthquake intensity measurements in Japan (figure 10). This graph is much

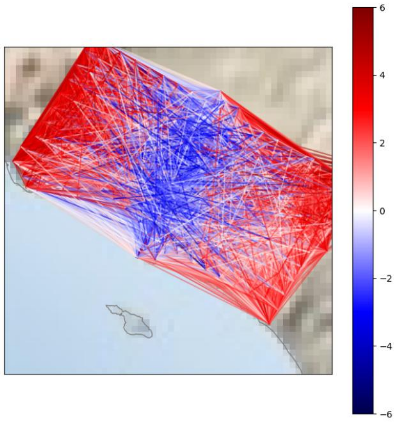

Figure 6: Edge weights in the LA graph

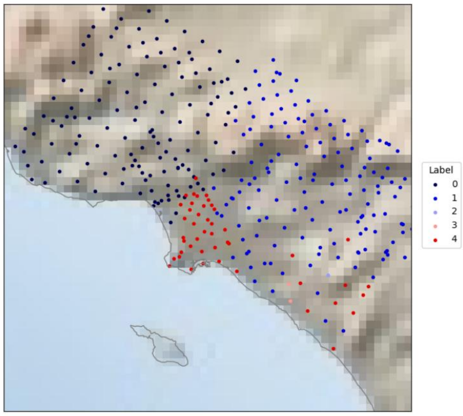

Figure 7: Nodes community assignment in LA graph using spectral clustering, 5 clusters

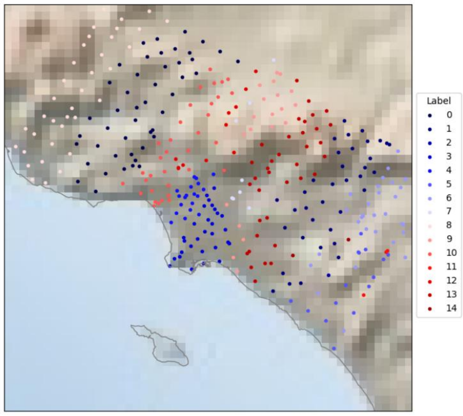

Figure 8: Nodes community assignment in the LA graph using spectral clustering, 15 clusters

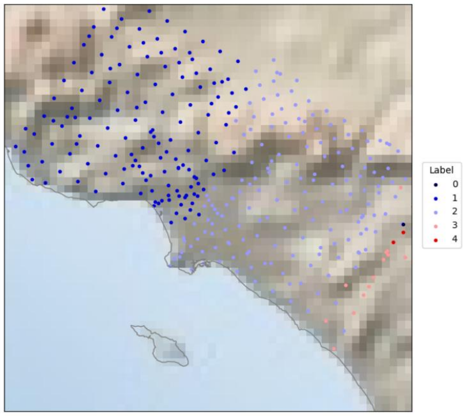

Figure 9: Nodes community assignment in the LA graph using Louvain

more complex. Since the data covers a large spatial area, it potentially contains many communities where correlations are unusually high or low. We applied the signed spectral clustering model to the Japan earthquake measurement correlation graph. The result is visualized on figure 11. We experimented on using both the signed ratio cut and signed normalized cut as our objective function. It is worth noting that for different cluster number k, the spectral clustering algorithm always produces a large community and the algorithm fails to further divide the community.

Figure 12 shows the detected communities using adjusted Louvain algorithm. Compared with spectral clustering results, the number of communities generated by Louvain is larger. It detected 43 communities and it is noticeable that most of these communities have similar size and small extent, which makes more geological sense.

On this complex dataset with many communities, Louvain is able to cluster nodes that are close geographically without using any distance attributes. Spectral clustering in this case, however, will group lots of nodes together, giving less insights.

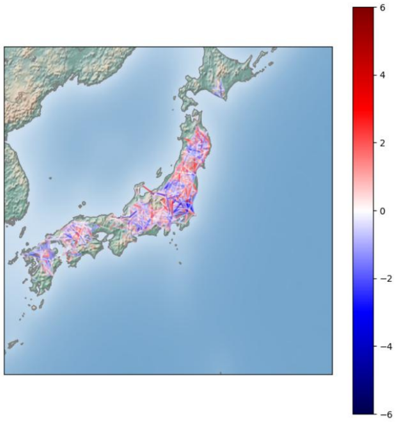

Figure 10: Edge weights in the Japan graph

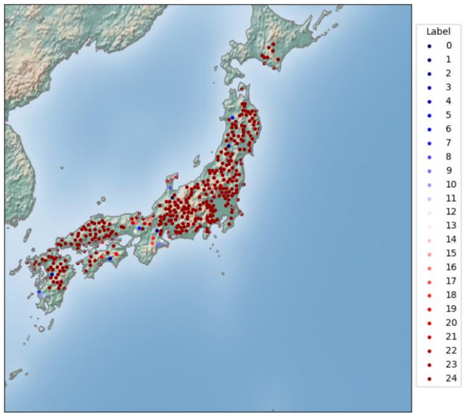

Figure 11: Nodes community assignment in Japan graph using spectral clustering

6. Evaluation and SSBM

We extended the notion of SBM in Holland, Laskey, and Leinhardt (1983), instead of computing connection probability within and between groups, we computed the mean strength and variances of each blocks, which is similar to Aicher, Jacobs, and Clauset (2014), and we assumed normal distribution of edge weigths within and between groups.

6.1. Visualization

We would also like to visually validate the clustering results based on blocks of the adjacency matrix. We expect nodes within the same community have higher than normal correlations. This information can be visualized by plotting the rearranged adjacency matrix.

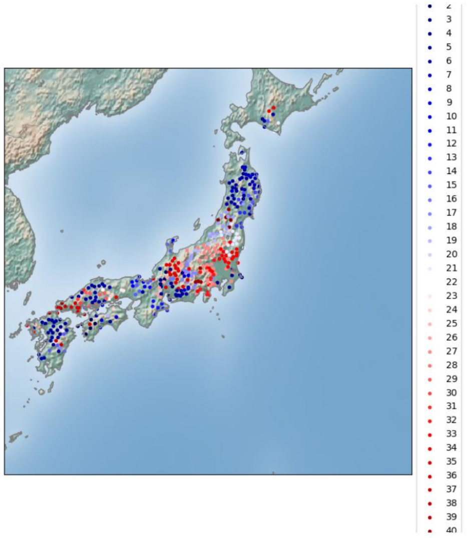

Figure 12: Nodes community assignment in Japan graph using Louvain

6.1.1 Los Angeles

Rearrangging the adjacency matrix based on clustering results. We have figure 13 and 14.

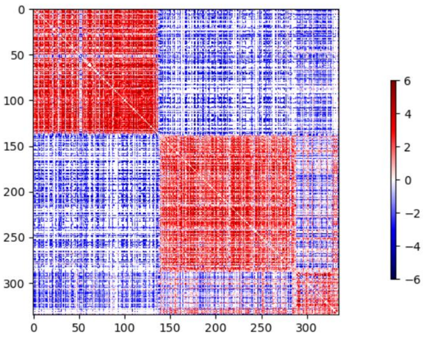

Figure 13: Block adjacency matrix of LA graph, 5 clusters

From the rearranged block matrix, we observe that within groups and between groups, there are clearly block patterns, which can be used to simulate graph using SSBM model.

6.1.2 Japan

Similarly, we have rearranged adjacency matrix from Japan data and obtained 15. Comparing with the adjacency ma-

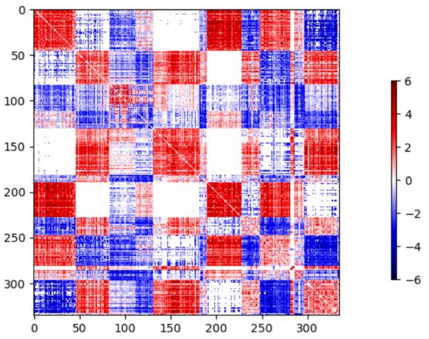

Figure 14: Block adjacency matrix of LA graph, 15 clusters

trix of Los Angeles, we observed weaker within groups and between groups edge strength.

As seen from 16, the spectral clustering method only picks up two clusters with no edges or edges has near zero correlations. However, within groups, the edge values are randomly distributed.

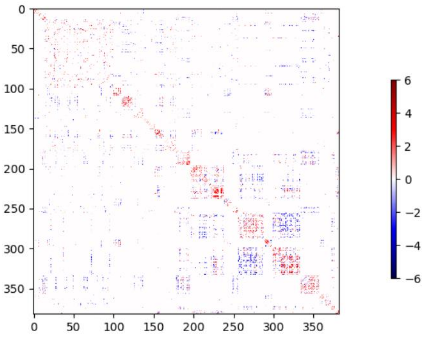

Figure 15: Block adjacency matrix of Japan graph, Louvain

6.2. Simulation and Link Prediction

There has been research done on link prediction on weighted signed networks Kumar et al. (2016). Here we conduct link prediction and network simulation based on SSBM models. We extract the parameter estimations of SSBM model by computing edge means and variances within groups and between groups based on our clustering results, and random variable is simulated by the normal distribution with the extracted mean and variance.

6.2.1 Los Angeles

The simulated SSBM in figure 17 and 18 are similar to their original counterparts.

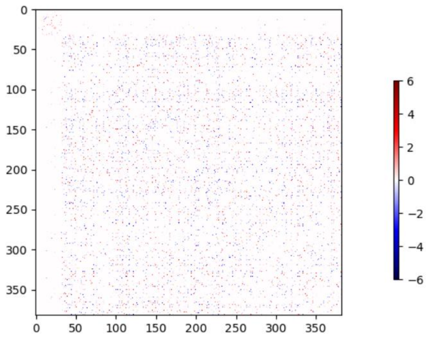

Figure 16: Block adjacency matrix of Japan graph. Spectral Clustering

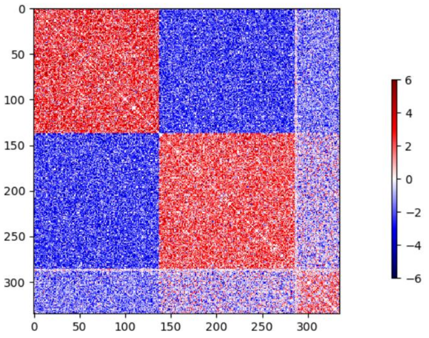

Figure 17: Simulated SBM, 5 clusters

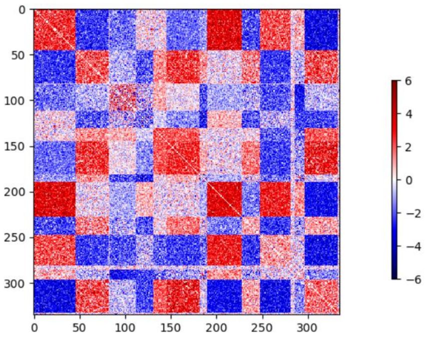

Figure 18: Simulated SBM, 15 clusters

6.2.2 Japan

The simulated SSBM for Japan data does not resemble the original adjacency matrix. An obvious reason is that the original graph have relatively sparse connections between nodes, however, when we simulate adjacency matrix from

SSBM, we will generate all edges from each nodes to every other node.

Comparing the simulated SSBM adjacency matrix from Louvain and Spectral, we can also observe that spectral clustering gives a near noise adjacency matrix whereas Louvain is able to find more reasonable groups.

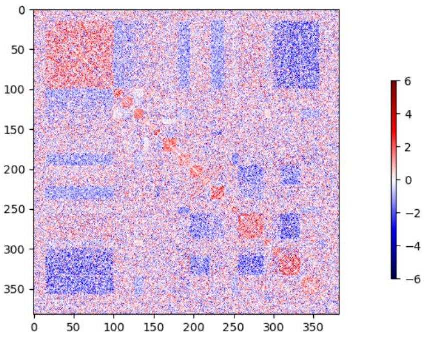

Figure 19: Simulated SBM, Japan, Louvain, 15 Clusters

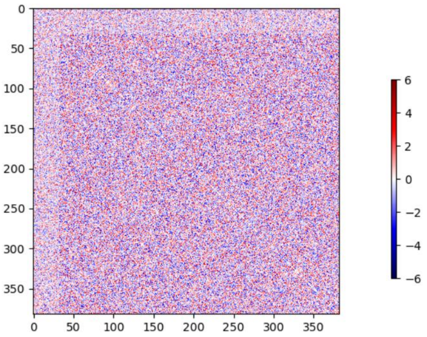

Figure 20: Simulated SBM, Japan, Spectral Clustering

7. Discussion

7.1. Complexity of the underlying spatial network

Both Wellington and Los Angeles datasets have relatively simple ground truth and fewer number of communities. In such cases, spectral clustering is able to produce very good result and recovers communities. However, for a graph like Japan, the underlying community assignment is much more complex and obscure. It also has many more communities in the dataset. In this case, spectral clustering does not produce reasonable results while signed Louvain is able to detect reasonable communities as illustrated in the previous section.

It has been shown by Nadler and Galun (2007) that the first few eigenvectors of adjacency matrices cannot successfully cluster datasets that contain structures at different scales of size and density. For the Japanese earthquake dataset, the network have different densities across different regions. Therefore, spectral clustering is unlikely to produce optimal result. Since spectral clustering algorithm is designed to solve a graph cut problem by splitting the graph into two clusters. When we want to produce more than two clusters, we use a K-means clustering algorithm with appropriate eigenvectors as features. In this case, it is intuitive to assume that when we have relatively few clusters, spectral clustering will be a good approximation. However, when number of clusters grows, the information provided by the eigenvectors is less likely to accurately separate clusters when fed into K-means. However, the Louvain algorithm overcomes this problem as it iteratively maximizes modularity until a local maximum is found.

7.2. Flexibility of Louvain

One of the downside of Louvain is also revealed from our experiment. In the Los Angeles dataset, signed Louvain stops after two clusters are identified. However, more insightful results can be found when we assign more communities. Although we can change the stop condition of the Louvain Algorithm to adjust the number of clusters, it still does not give us as much freedom as spectral clustering, for which we can choose number of clusters manually.

8. Conclusion

In our project, we explored modified spectral clustering and signed Louvain algorithm's performance on signed spatial networks. We identified that for more local and graphs with fewer communities spectral clustering gives very good community recovery results. For more global and graphs with many communities, Louvain outperforms spectral clustering.

We also provided a way to simulate earthquakes using community detection results and symmetric stochastic block model(SSBM). This method both validates our community detection result and has other application in the study of simulating spatially correlated earthquake data.

The code of this project can be found at https://github.com/yilinchen0911/cs224wProjectPublic.git

References

Aicher, Christopher, Abigail Z. Jacobs, and Aaron Clauset (2014). "Learning Latent Block Structure in Weighted Networks". In: *arXiv e-prints*, arXiv:1404.0431, arXiv:1404.0431. arXiv: `1404.0431 [stat.ML]`.

Blondel, Vincent D et al. (2008). “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10, P10008.

Gomez, S., P. Jensen, and A. Arenas (2008). “Analysis of community structure in networks of correlated data”. In: *ArXiv e-prints*. arXiv: 0812.3030 [physics.soc-ph].

Holland, Paul W., Kathryn Blackmond Laskey, and Samuel Leinhardt (1983). “Stochastic blockmodels: First steps”. In: *Social Networks* 5.2, pp. 109–137. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(83)90021-7. URL: http://www.sciencedirect.com/science/article/pii/0378873383900217.

Kumar, Srijan et al. (2016). “Edge Weight Prediction in Weighted Signed Networks”. In: *Data Mining (ICDM), 2016 IEEE International Conference on*.

Kunegis, Jérôme et al. (2010). “Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization”. In: *SDM*.

Nadler, Boaz and Meirav Galun (2007). “Fundamental Limitations of Spectral Clustering”. In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, pp. 1017–1024. URL: http://papers.nips.cc/paper/3069-fundamental-limitations-of-spectral-clustering.pdf.

Newman, M. E. and M. Girvan (2004). “Finding and evaluating community structure in networks”. In: 69.2, 026113, p. 026113. DOI: 10.1103/PhysRevE.69.026113. eprint: cond-mat/0308217.

Small, Patrick et al. (2017). “The SCEC unified community velocity model software framework”. In: *Seismological Research Letters* 88.6, pp. 1539–1552.