Social Network Clustering

Kyle Luh, Peter Elliott, and Raymond Ahn

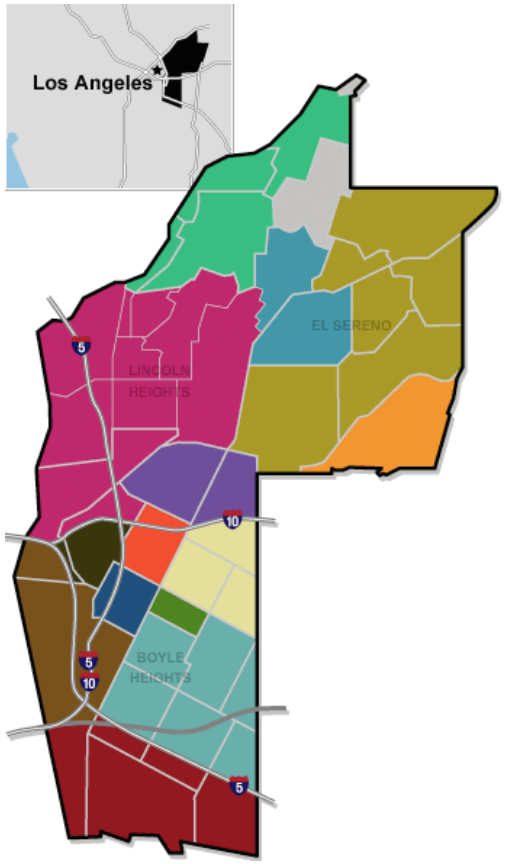
University of California Los Angeles

August 2, 2011

Outline

Hollenbeck Gang Activity

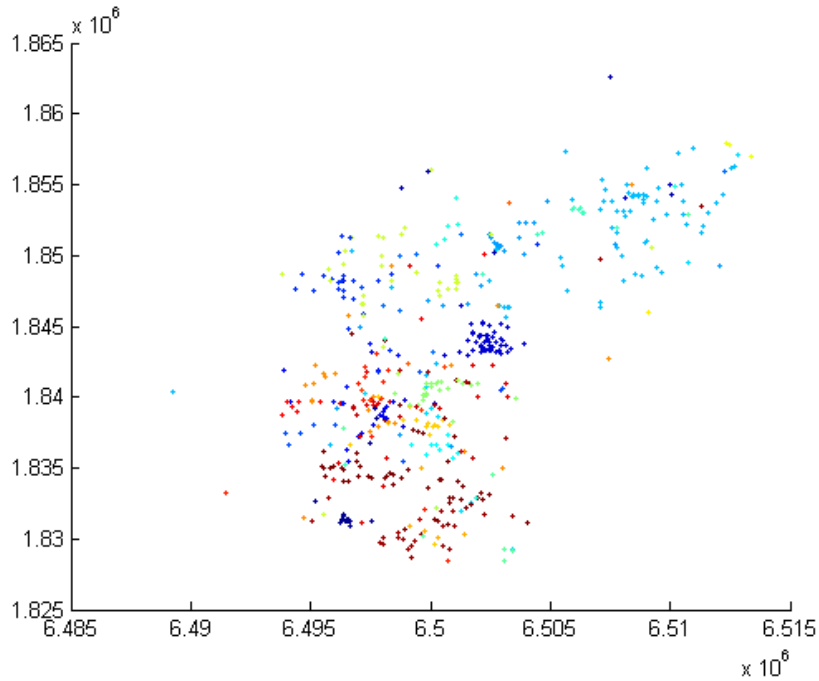
- Hollenbeck has an area of approximately 15.2 miles.
- In this area, 31 violent gangs reside.
- Hollenbeck is one of the top three most violent LA policing regions.
- Gang violence in this region has existed since before WWII.



Hollenbeck Gang Activity

The LAPD has provided an Excel database of non-criminal stops they have made in the Hollenbeck area. The data includes:

- time of stop
- location (gang territory and coordinates)
- gang affiliation
- sex
- age
- ethnicity

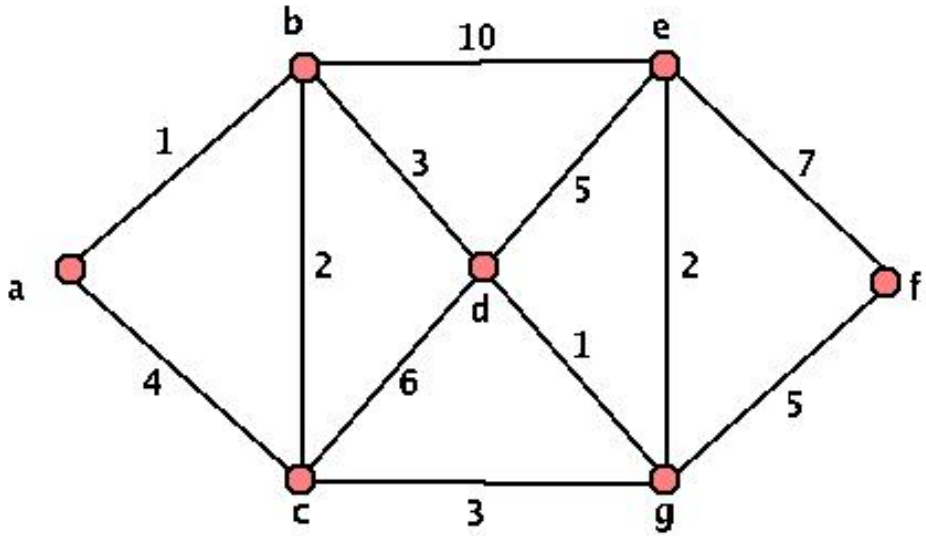


Goals

- Use clustering techniques to predict unknown gang affiliations.
- Detect other social structures that may not be captured by gang affiliation.

Graph Models

- Convert individuals into nodes.
- Edge weights indicate similarity.
- Unfortunately, data is sparse.

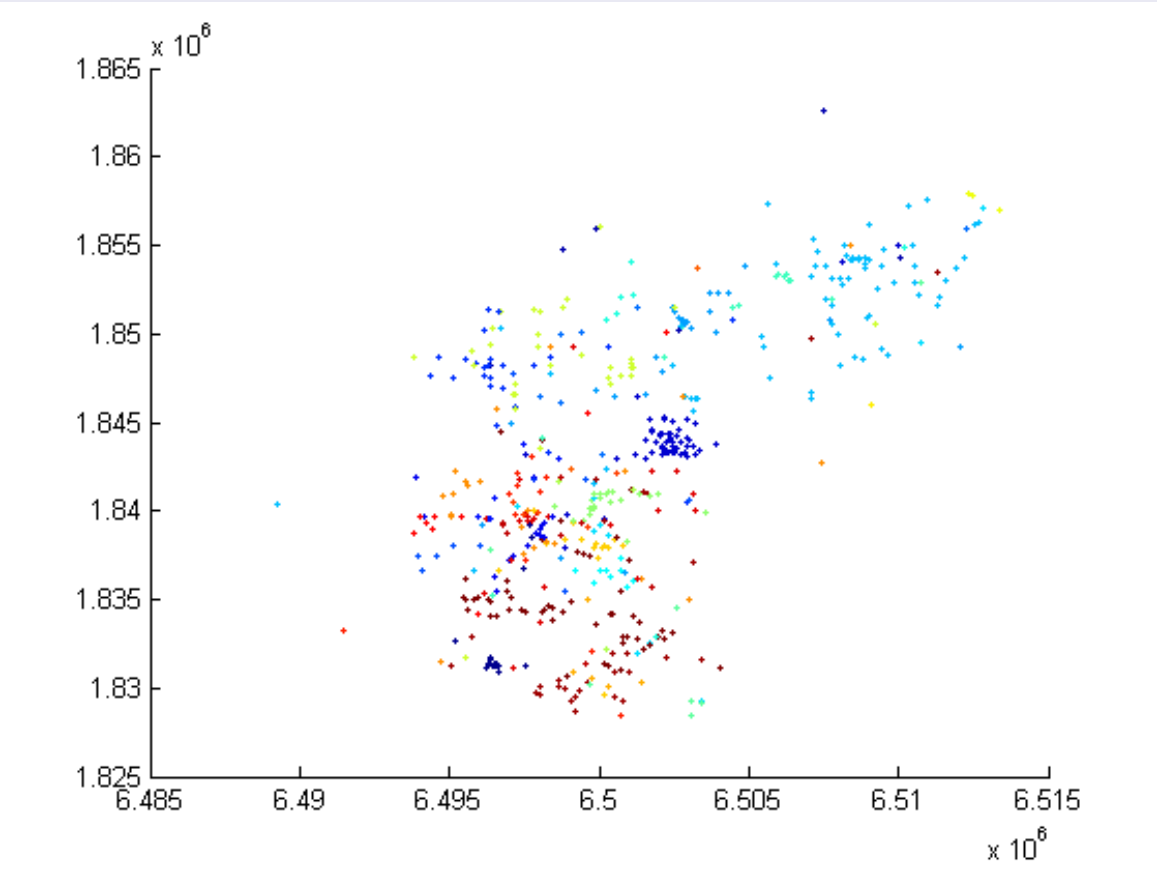


Choosing a Measure of Similarity

- A function of Euclidean distance
- Dot product of feature vector
  - Gang territory
  - Individuals and their gang associations
  - Individual to individual interactions

Results

Actual Gang Clusters



Outline

K-means algorithm

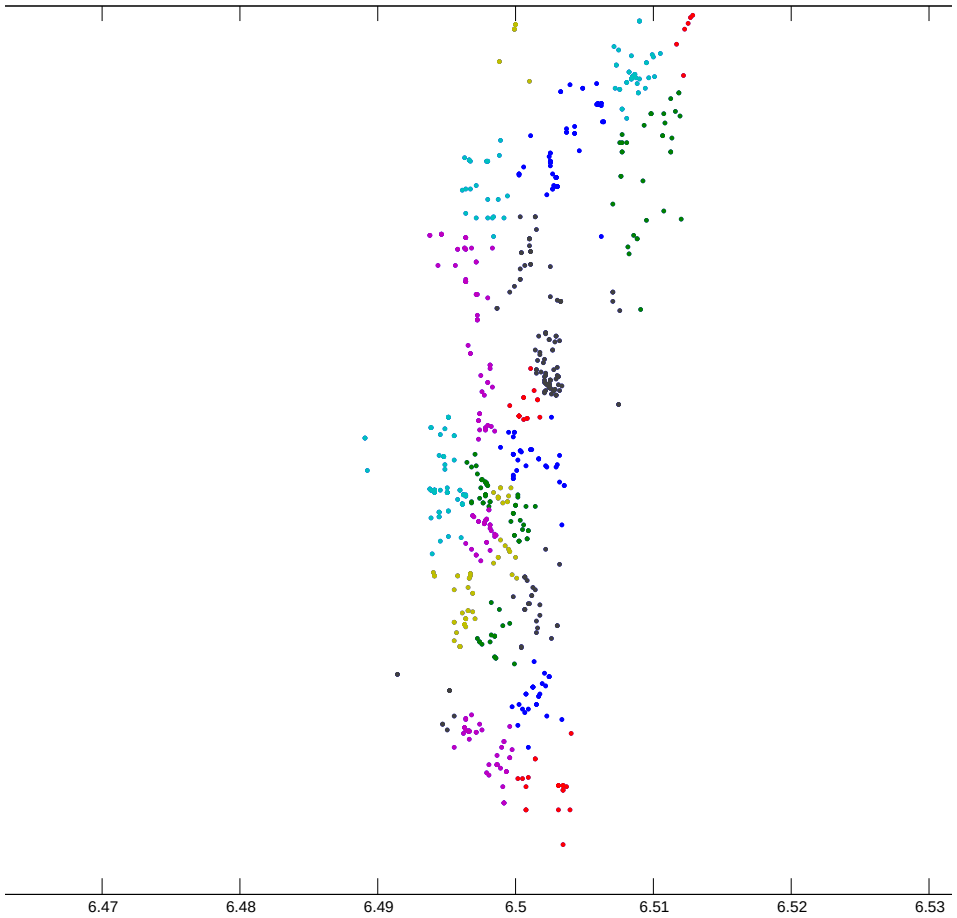
Algorithm

- Choose number of partitions.
- Compute centroids.
- Shift centers to the centroid of their affiliated points.
- Repeat until equilibrium is achieved.

Cons

The K-means algorithm only accounts for location. We hope to utilize more of the data.

Results: K-means Approach



A Metric for Cluster Evaluation

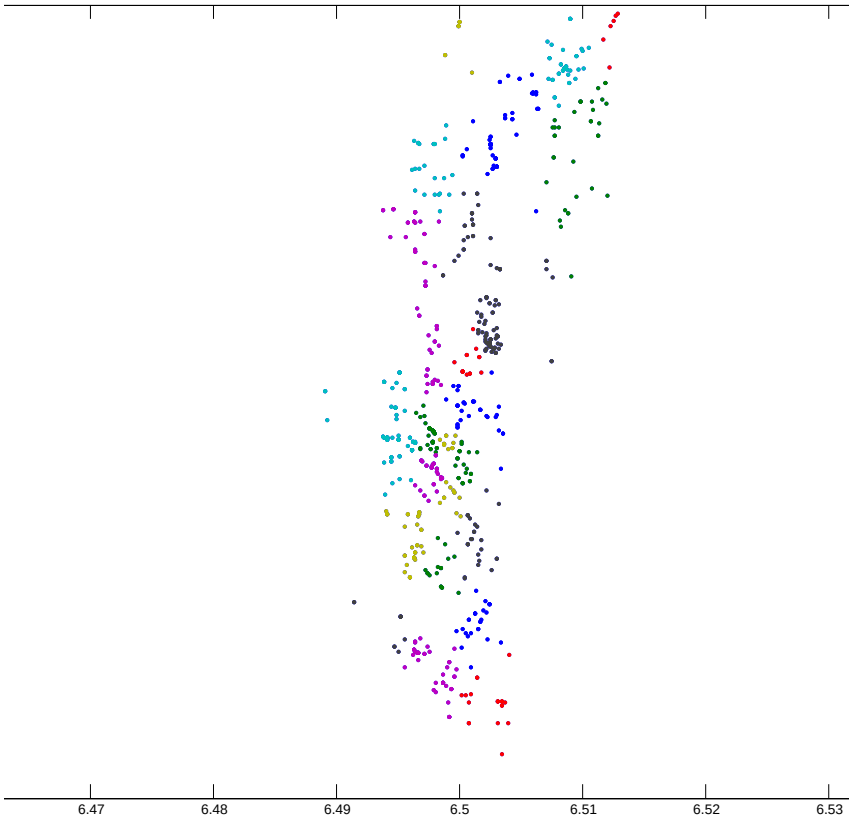
- We define **purity** to be

$$purity(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} max_j |\omega_k \cap c_j|$$

  where $\Omega = \{\omega_1, \cdots, \omega_K\}$ are the clusters and $\mathbb{C} = \{c_1, \cdots, c_j\}$ are the actual classes.
- Another measure we used was **Adjusted Mutual Information** which may be more appropriate since our gangs vary significantly in size.

Results: K-means Approach

Purity ≈ 0.4 and *AMI* ≈ 0.4

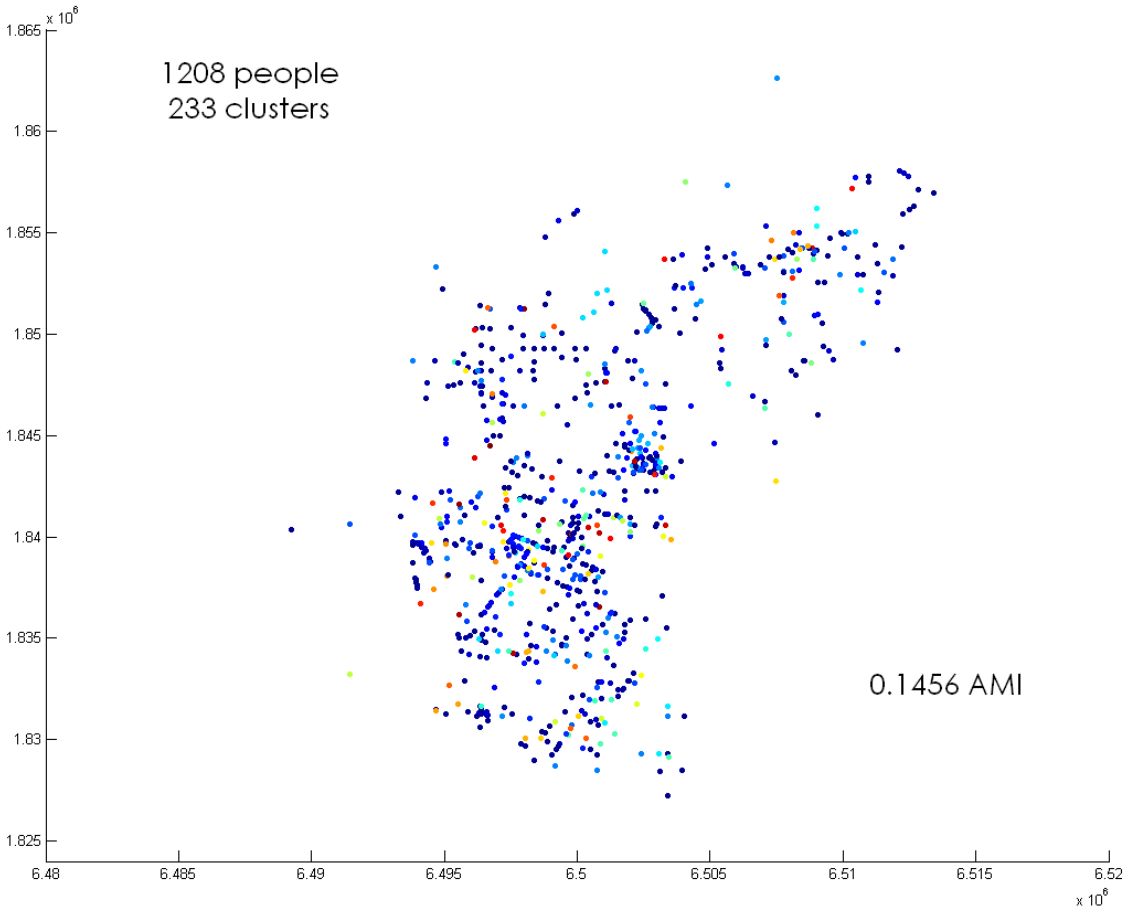


Modularity Maximization

- Modularity compares the number of edges within a cluster to the number expected
- Maximize modularity.
- We can calculate the change in modularity at each step and stop when the change is not positive

[M.J. Newman, 2006]

Results: Modularity Maximization



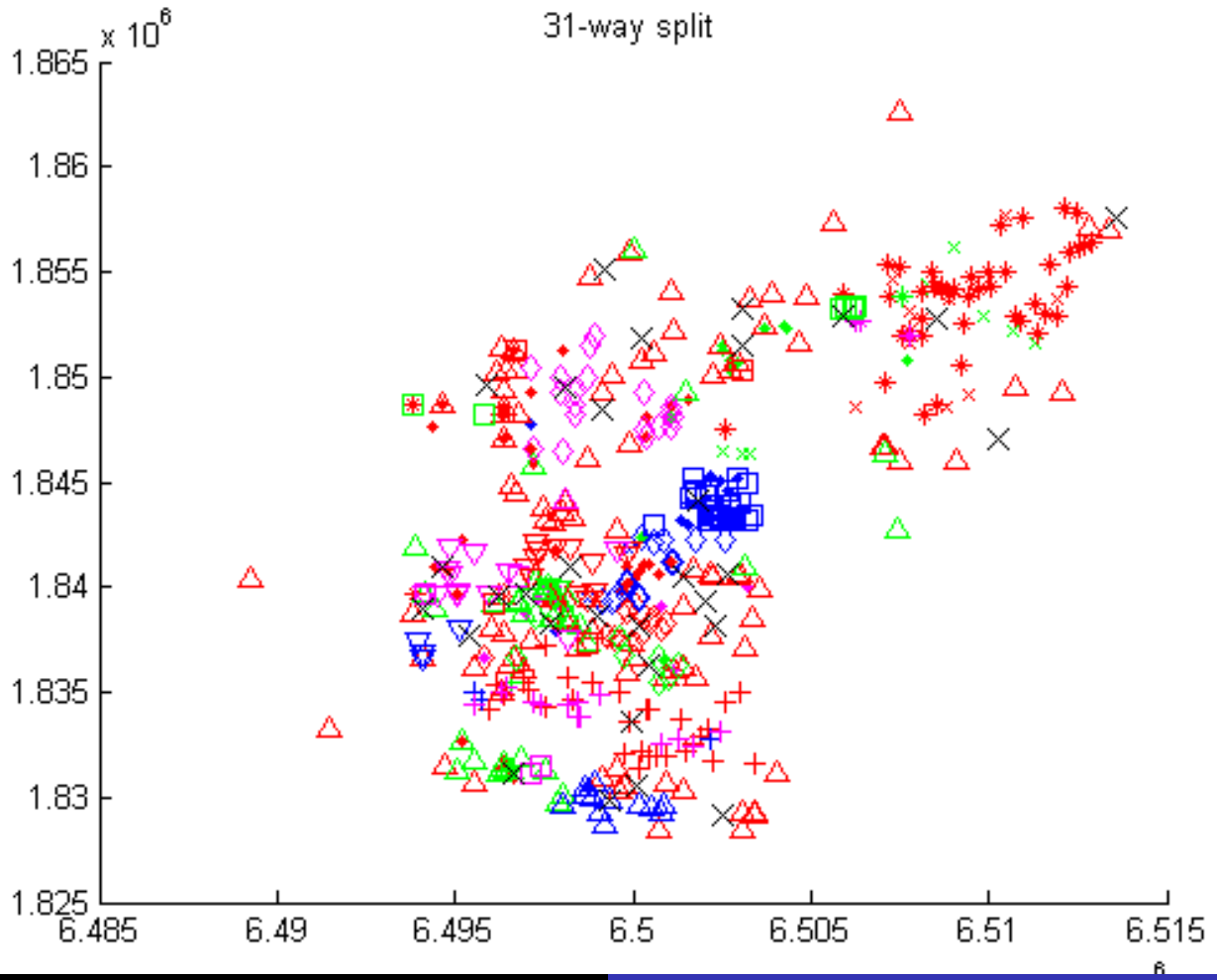
Convergence of Iterated Correlations (CONCOR)

- Compute correlations of entries to the mean of rows/columns
- Continue to calculate the correlations of the correlation matrix until we are left with $+1$ and -1.
- The method is repeated on each cluster to achieve a finer partition.

[Wasserman, 1994]

Results: CONCOR

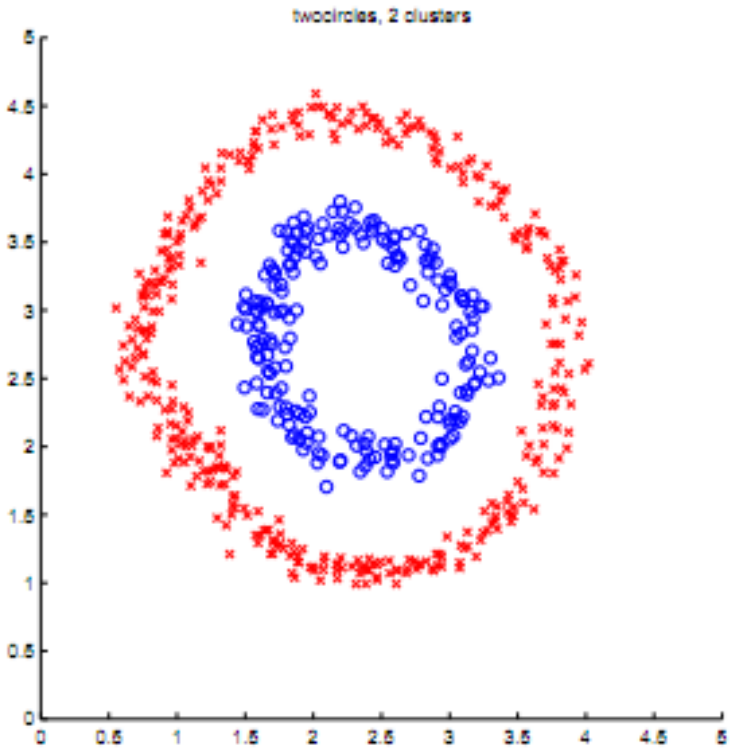
Purity ≈ 0.5 and *AMI* ≈ 0.46.

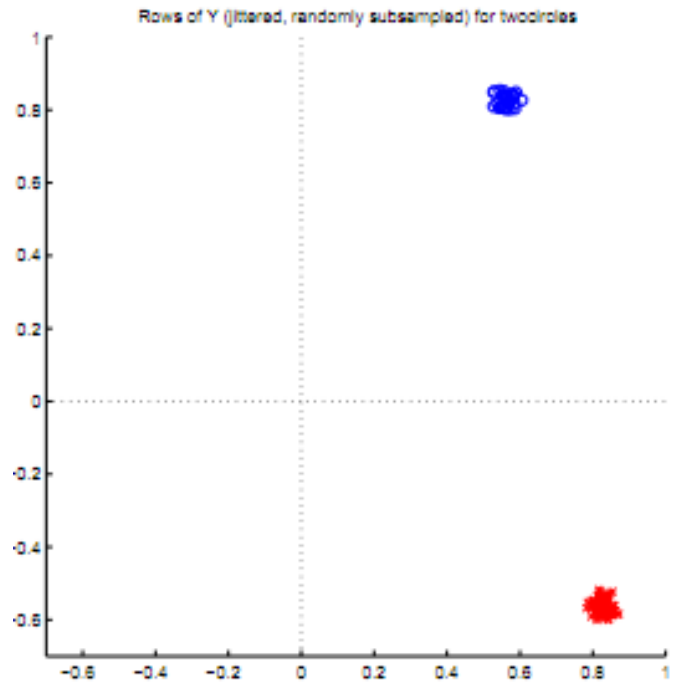


Outline

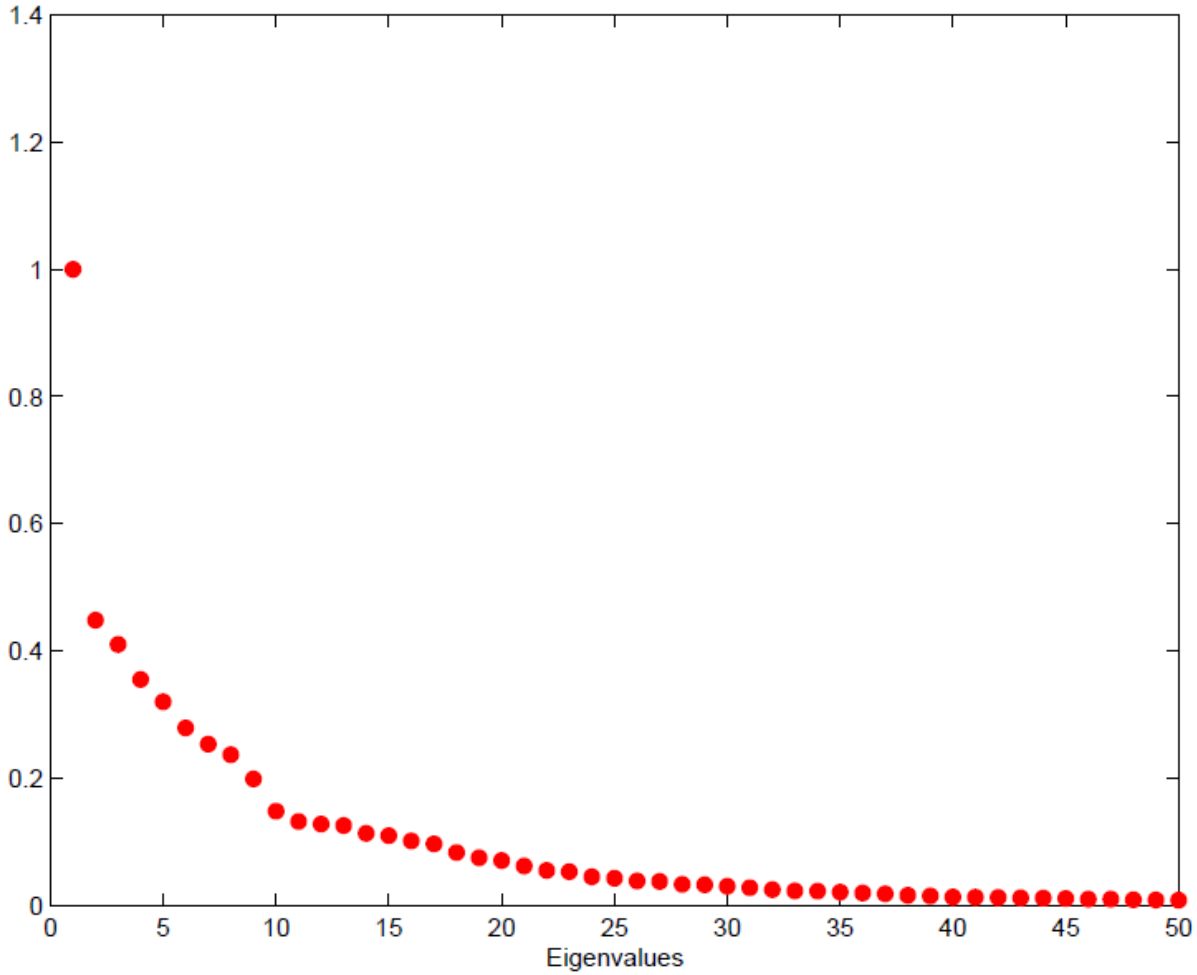
Spectral Clustering

- Create a matrix of eigenvectors of the Adjacency matrix.
- The eigenvectors capture the axes which contain the most variation in the data.
- Run k-means algorithm on new space.

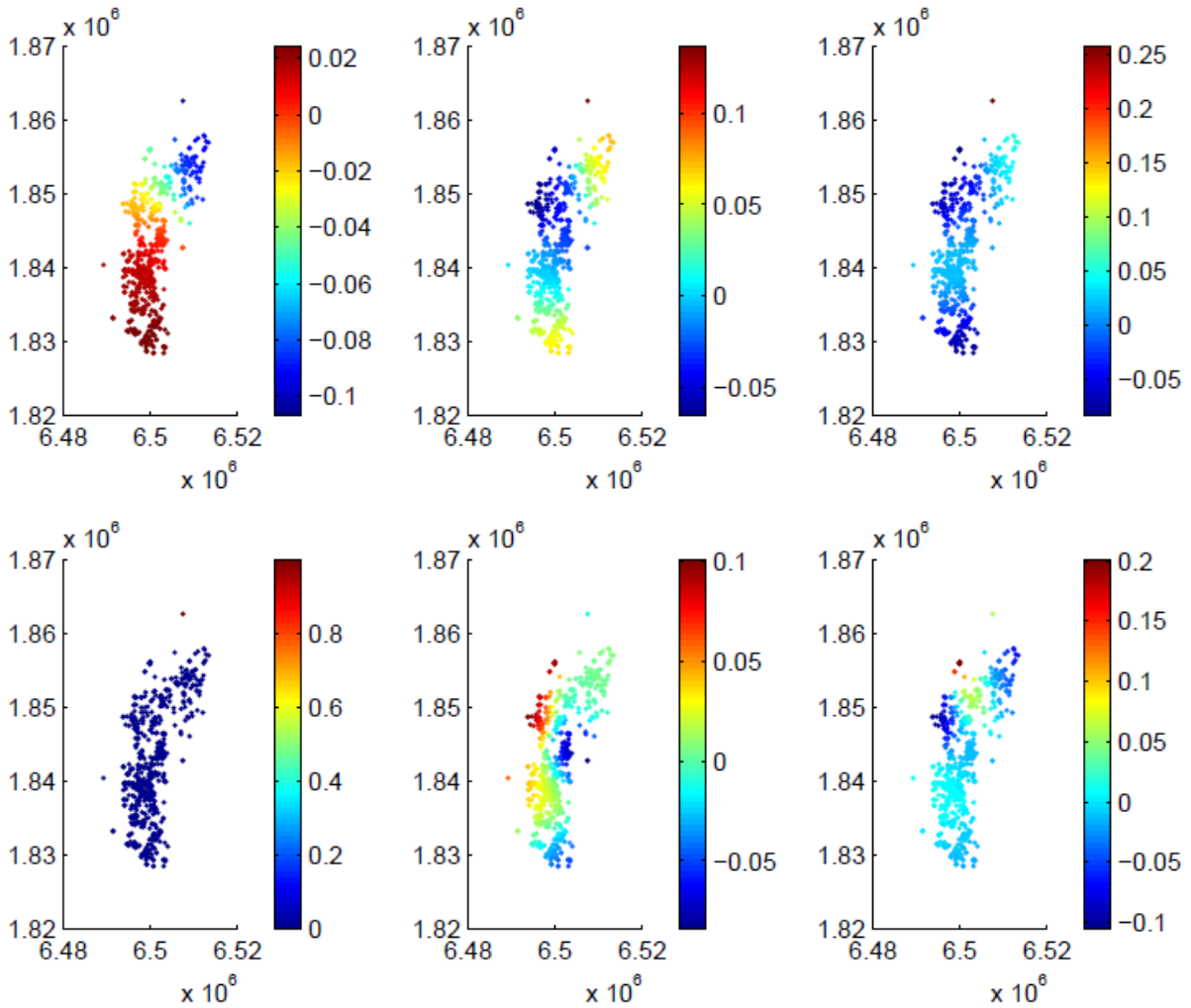




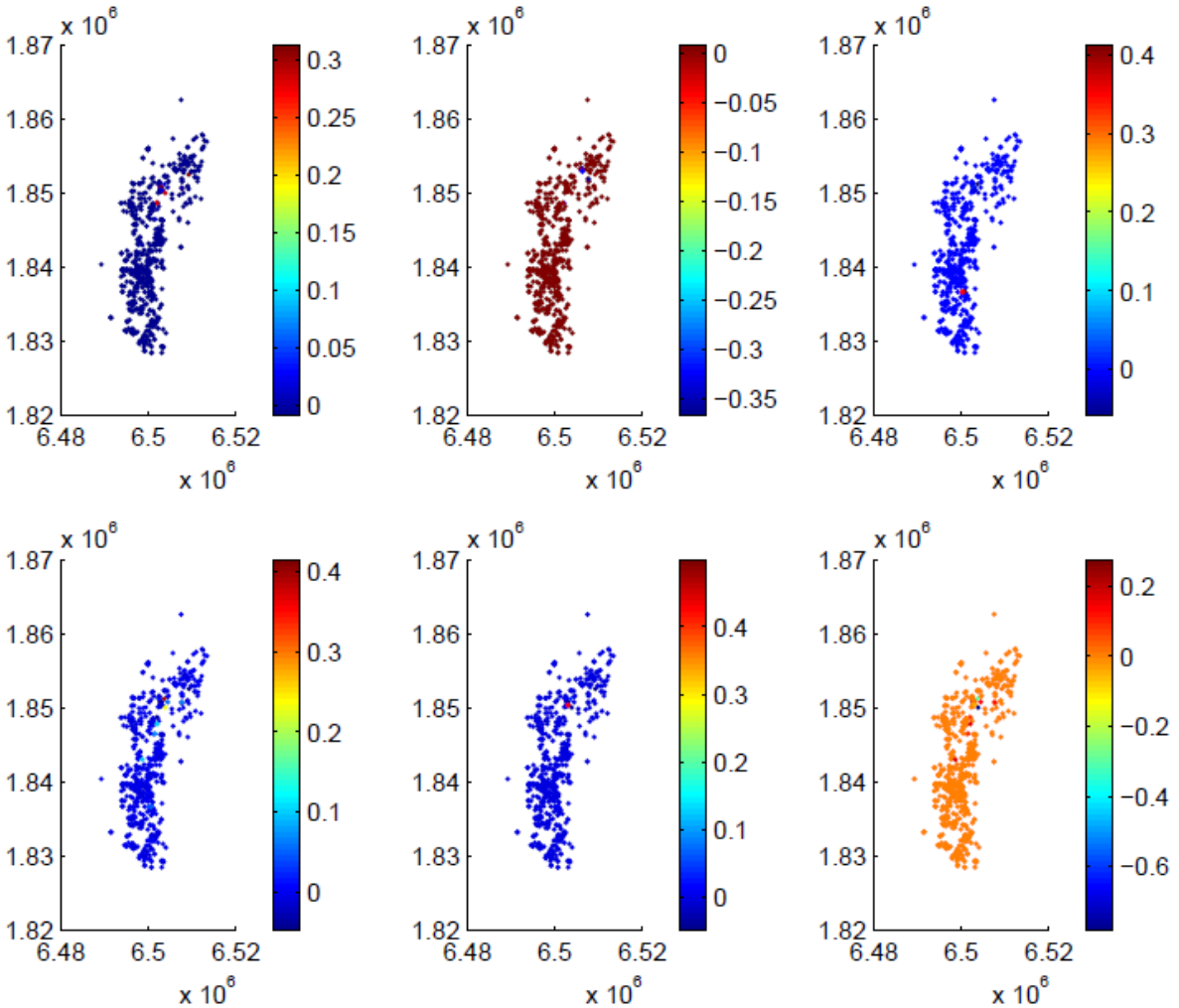
Eigenvalues



Eigenvector Plots: Distance Only



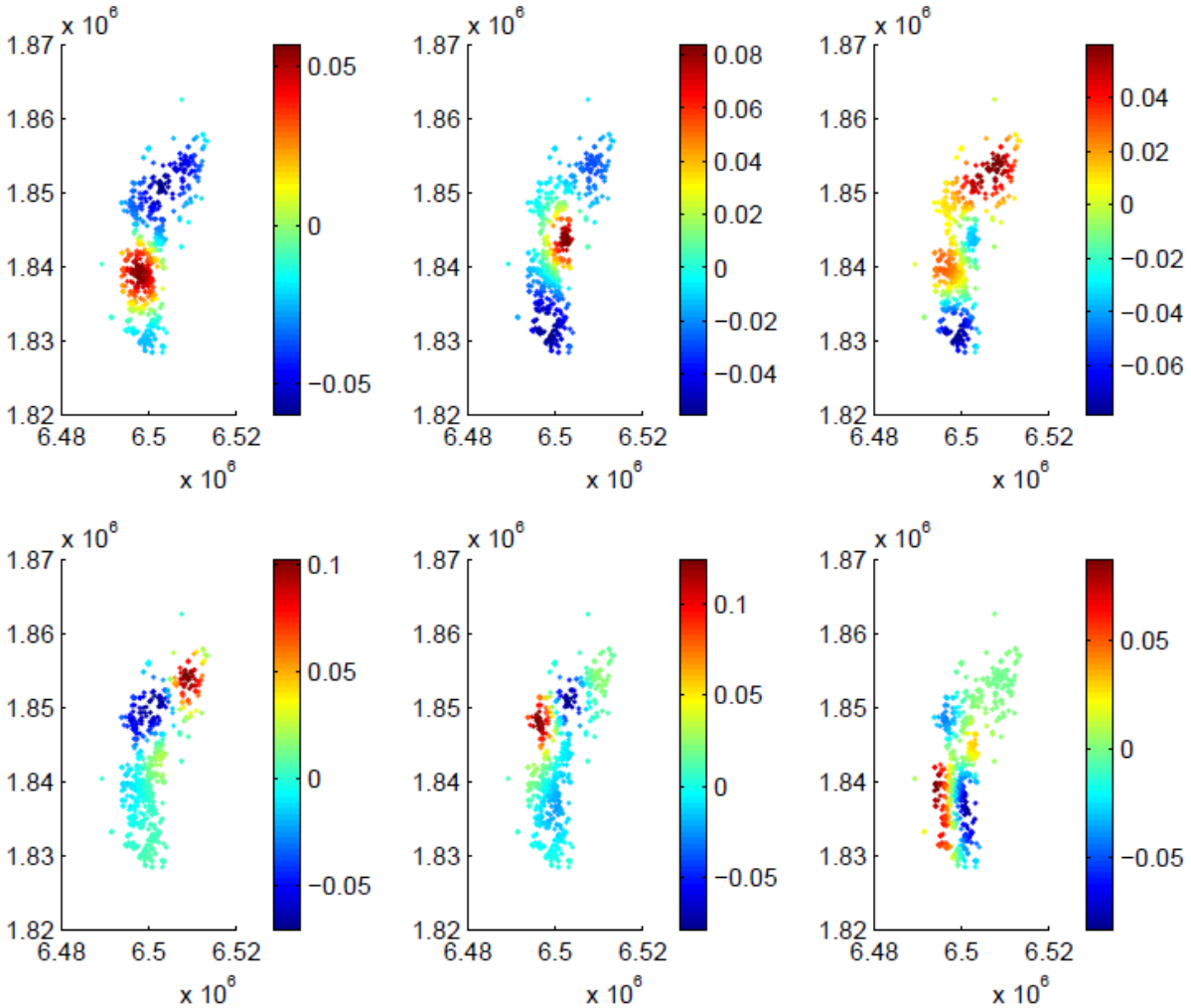
Eigenvector Plots: Social Information Only



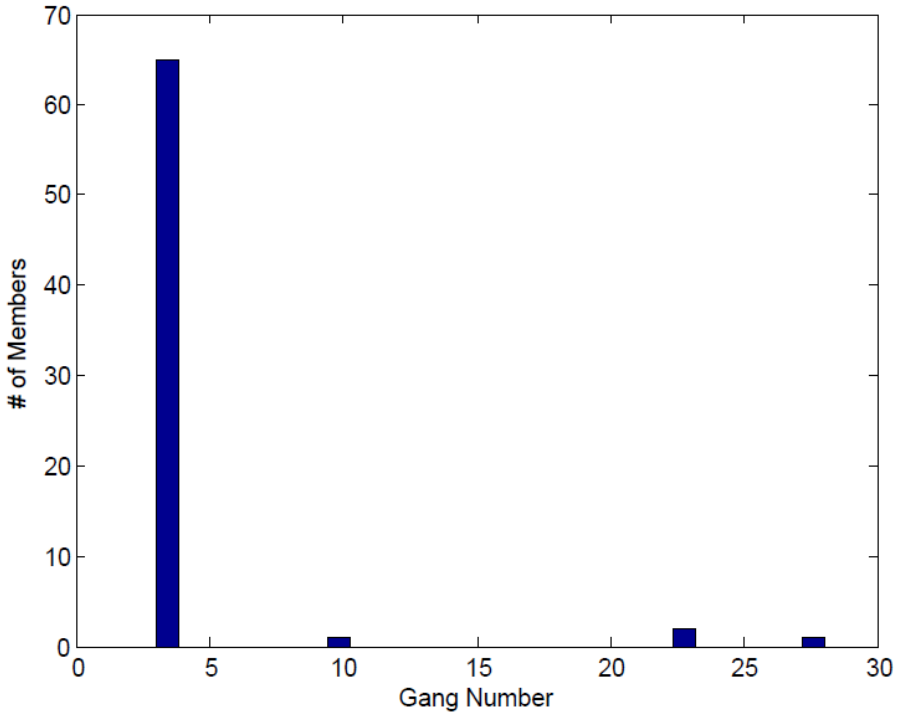
Where to go from here?

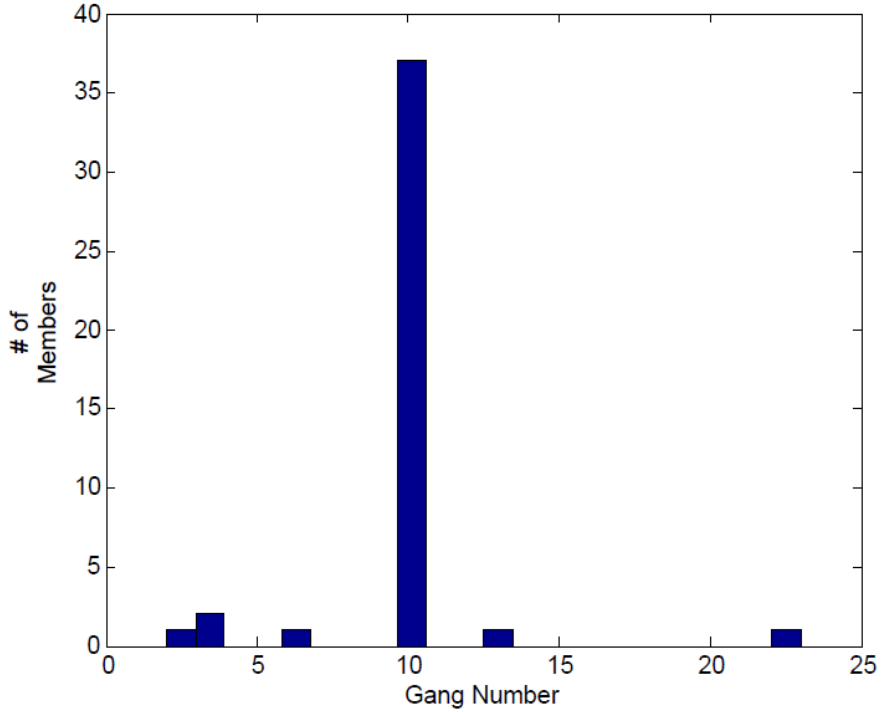
- The geographic data provides no insights.
- The social data is so sparse that its eigenvectors are completely useless alone.
- We decided to combine the two adjacency matrices, $\alpha A + (1 - \alpha)B$, where α is a weighting parameter.

Eigenvector Plots: Combined

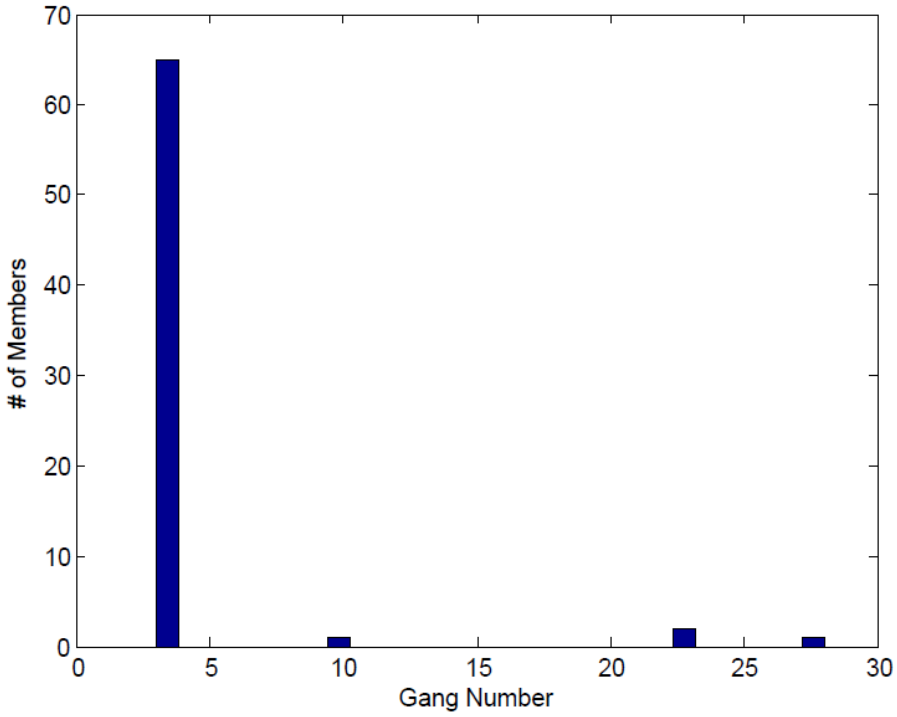


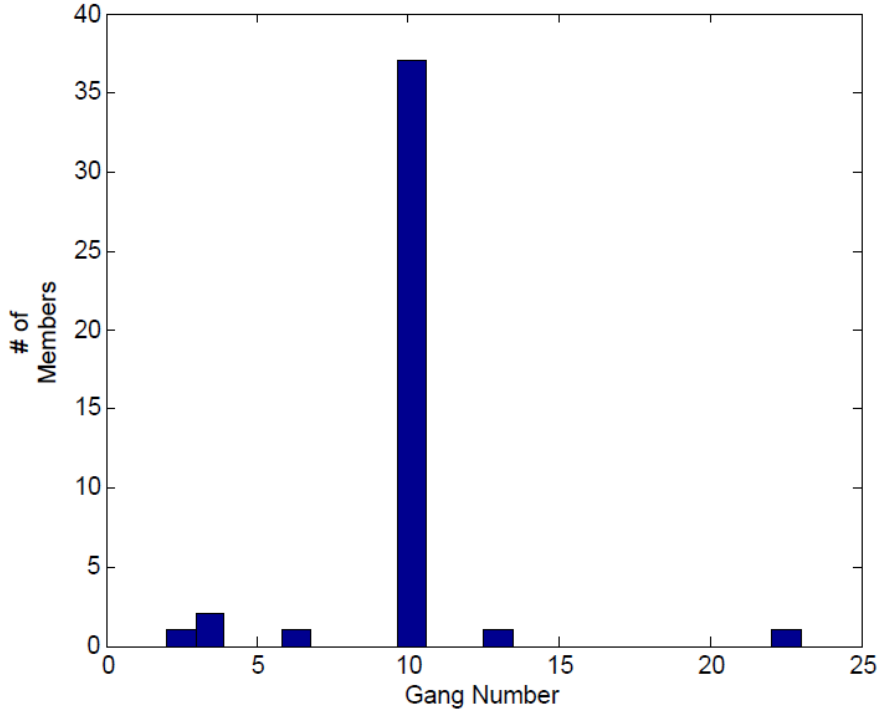
Clustering Results



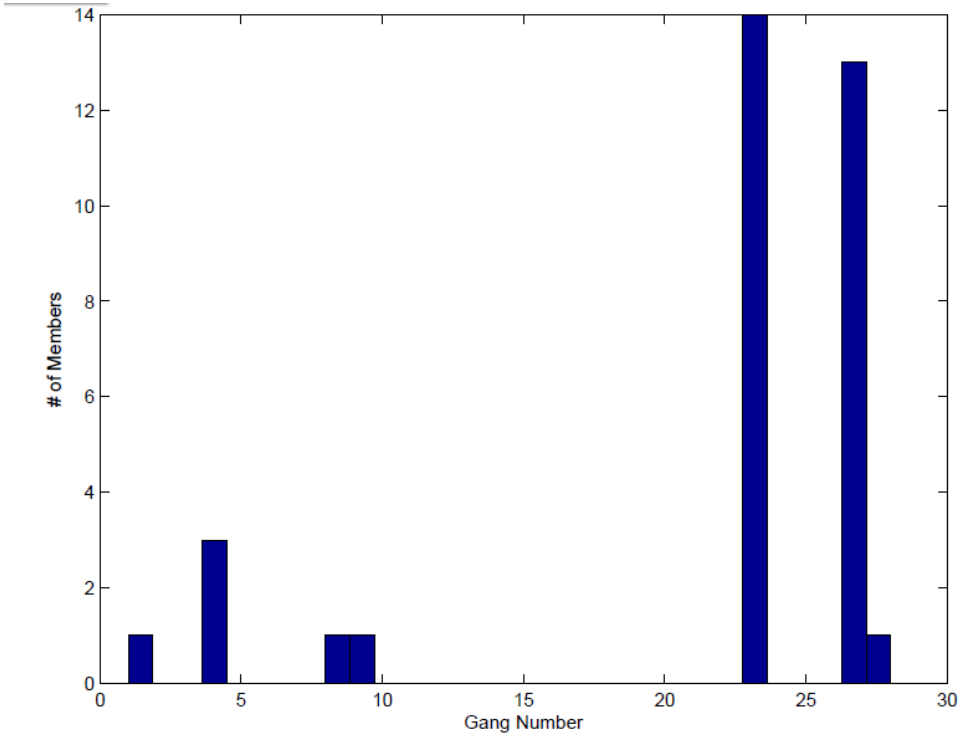


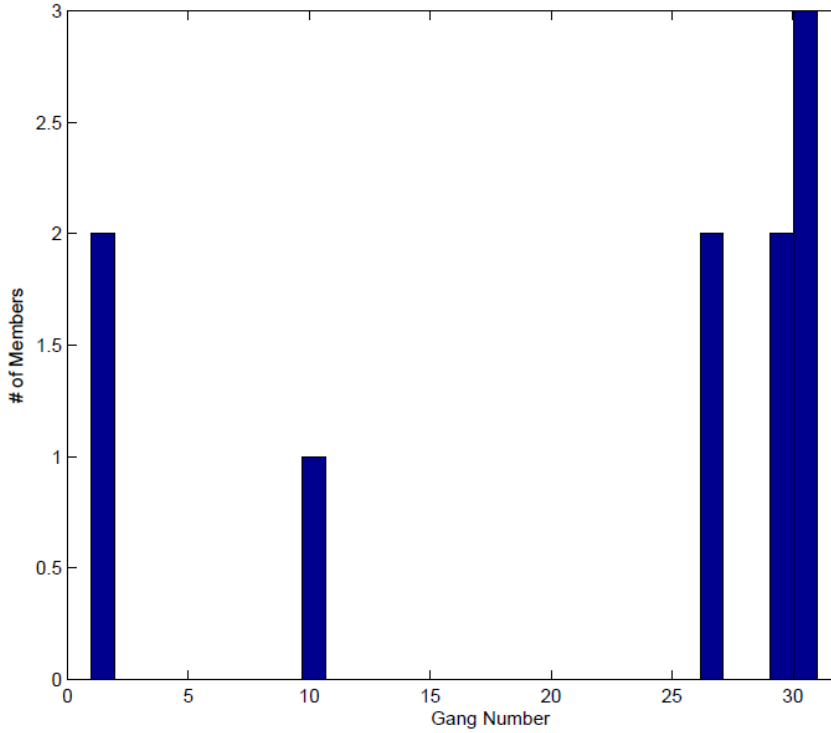
Clustering Results





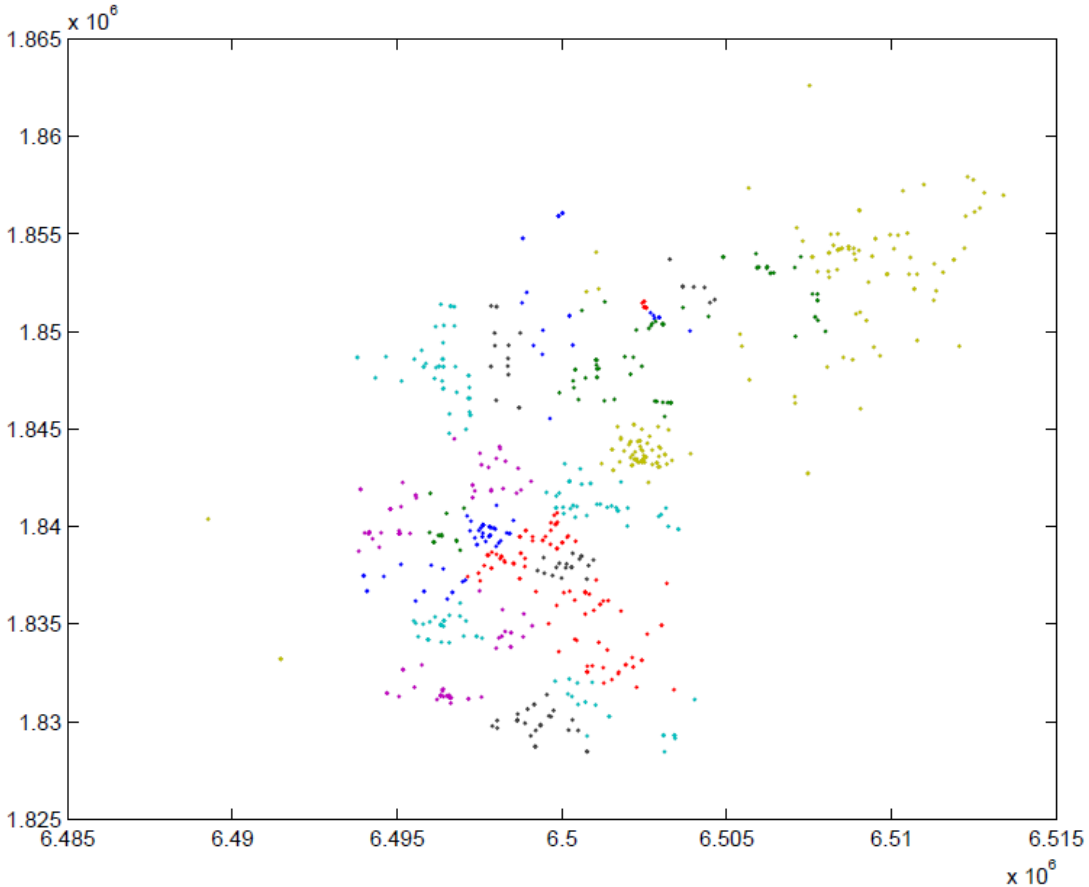
Clustering Results





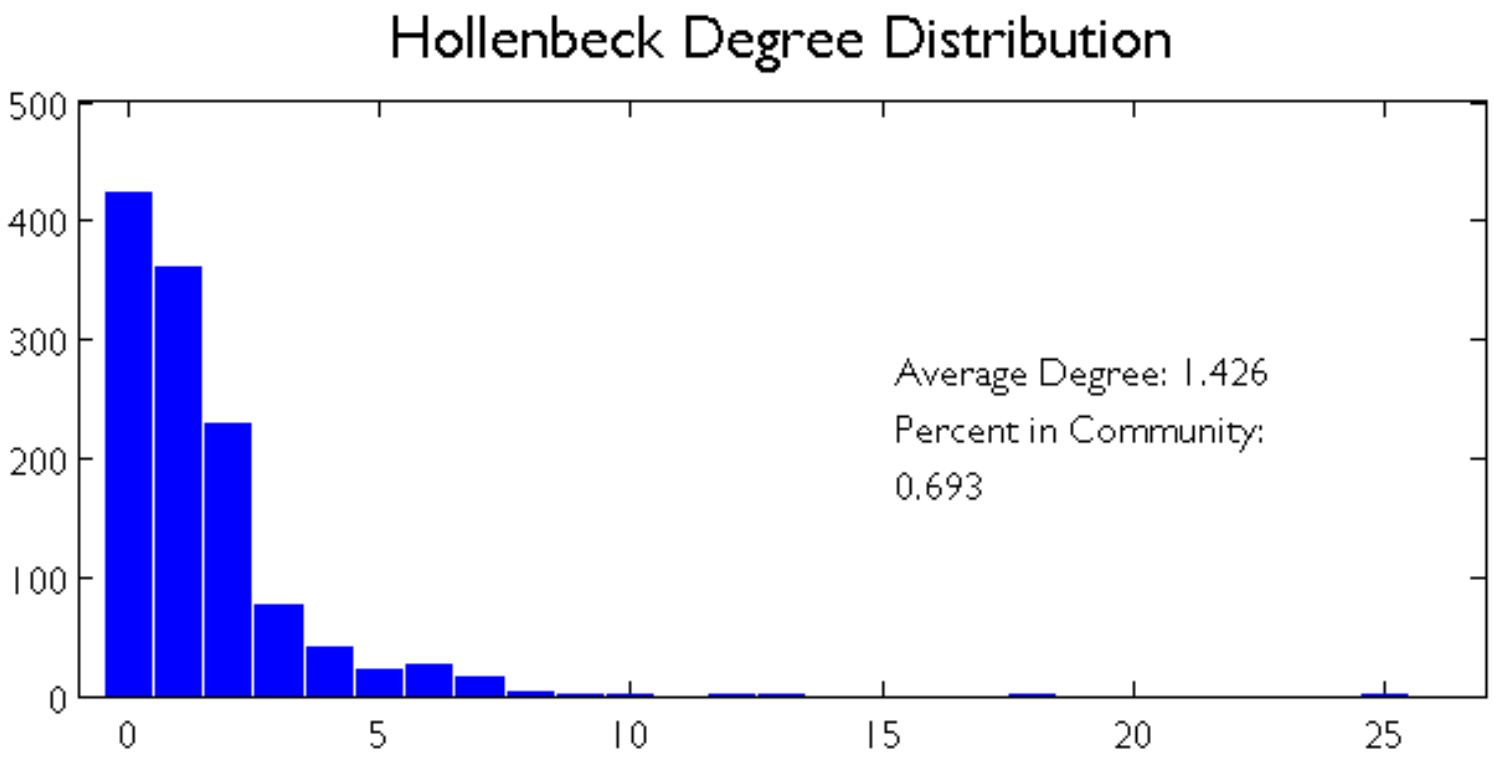
Results: Spectral Approach

Purity $\approx .7$ and *AMI* $\approx .65$



Outline

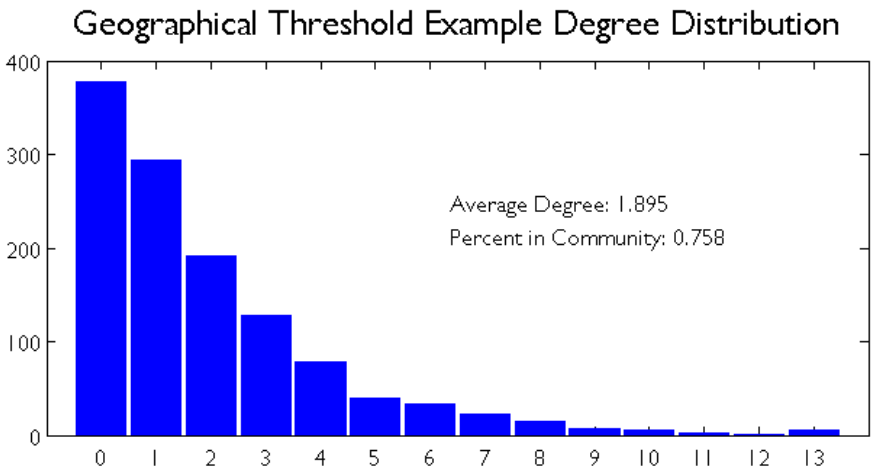
Artificial Data



Artificial Data

Inputs:

- Number of people
- Number of communities
- Gang multiplier
- Threshold
- $G(x, y) = \frac{(\eta_x + \eta_y)\sigma}{dist(x,y)} (1 + M\delta_{ij})$

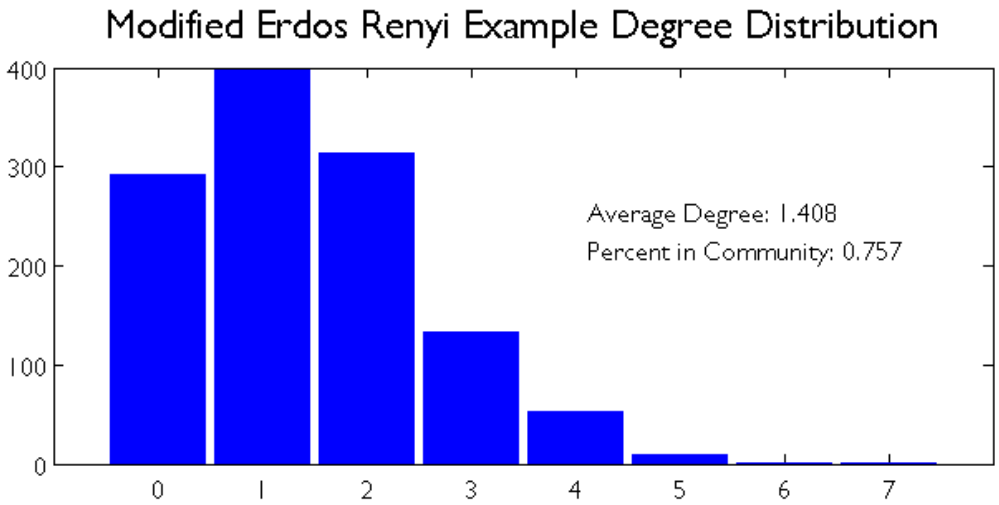


[N Masuda, 2005]

Artificial Data

Inputs:

- Number of people
- Number of gangs
- Probability within gangs
- Probability outside gangs

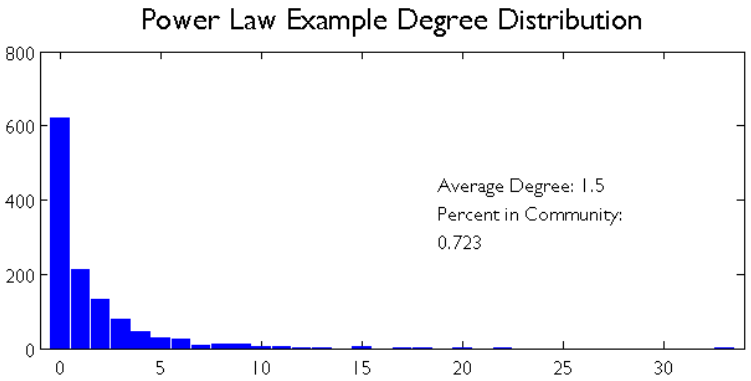


[A. Lancichinetti, 2008]

Artificial Data

Inputs:

- Number of people
- Number of gangs
- Average degree
- Mixing parameter

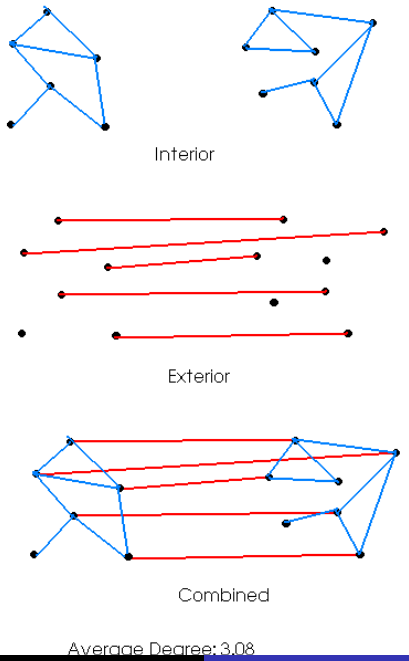


[E.N. Gilbert, 1959]

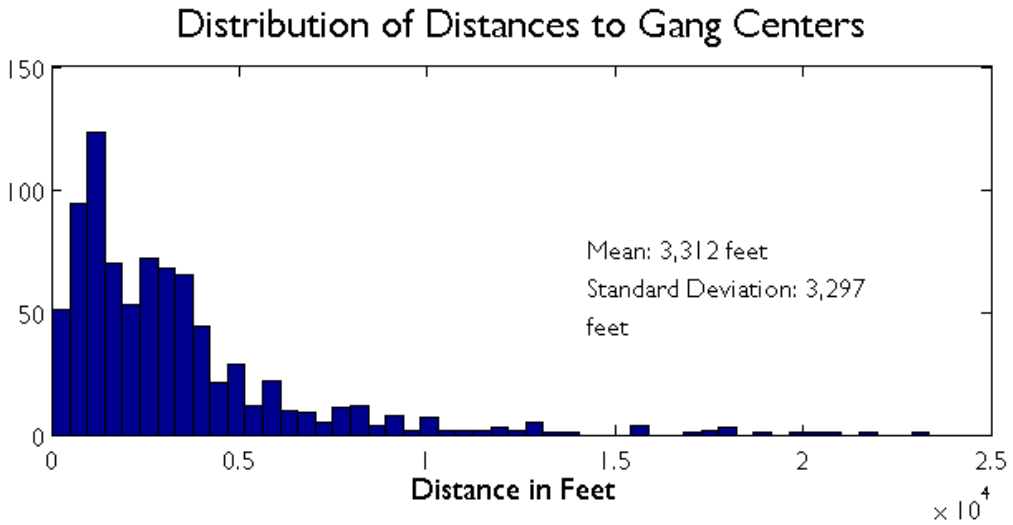
Artificial Data

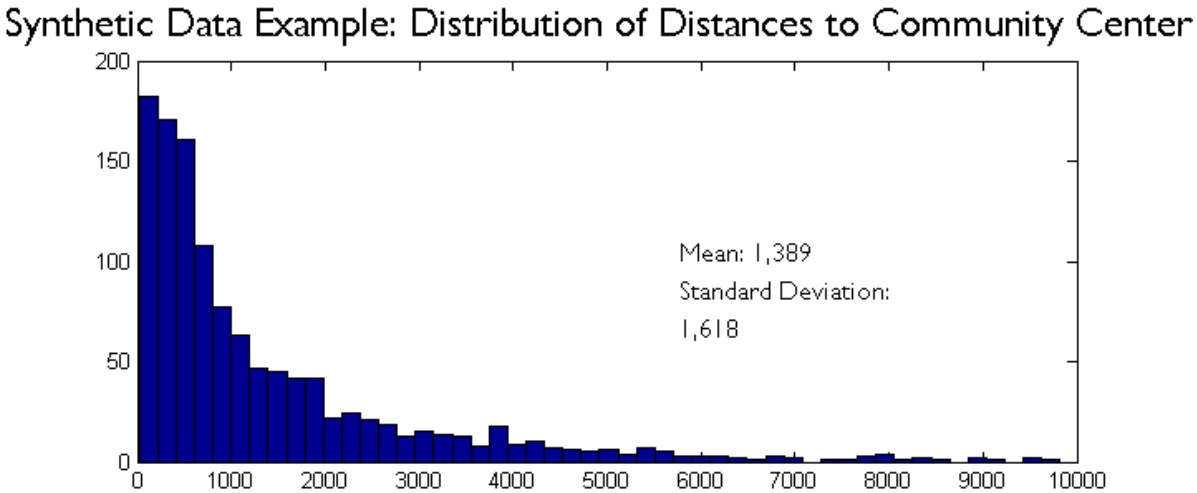
Inputs:

- Number of people
- Number of gangs
- Average degree
- Mixing parameter



Artificial Data





Outline

Future Work

- Matrix Completion and Link Prediction
- Robustness of Algorithms
- Artificial Testing Data

Acknowledgements

We would like to thank Professor Yves van Gennip for advising us throughout the project. We are also grateful to Professors Blake Hunter and Allon Percus for their helpful advice.

References


Ng et al. (2001)
On Spectral Clustering: Analysis and an Algorithm
Advances in Neural Information Processing Systems 14(2), 849 – 857.


Wasserman (1994)
Social Network Analysis
Cambridge University Press


M.E.J. Newman (2006)
Modularity and community structure in networks
Proceedings of the National Academy of Sciences in the United States of America 103 (23) 8577-8582


Gilbert (1959)
Random Graphs
The Annals of Mathematical Statistics


Lancichinetti et al. (2008)
Benchmark Graphs for Testing Community Detection Algorithms