Cleaning Up The Neighborhood: Duplicate Detection and Community Analysis of Hollenbeck Gangs

Ryan de Vera, Anna Ma, Daniel Moyer, Brendan Schneiderman

August 10, 2012

Abstract

The gang activity in Hollenbeck, a region in Los Angeles, has been both well documented by criminologists and monitored by the LAPD. The 2011 UCLA REU Social Networks group proposed that a spectral clustering method could predict gang affiliation from this police data. In the present work we continue and extend this method, applying it to a larger dataset from the same area. We further propose a modularity based method as an alternative to spectral methods. Unlike the previous work the data presented required considerable reshaping and linkage before use, and a method for such is also developed. Gang interactions, expressed as graph superstructures, are considered and discussed as well.

1 Introduction

In this report, we explain our attempt to gain a deeper understanding of the gang community in a Los Angeles policing district known as Hollenbeck. Collecting data from "Field Interview cards, filled out by police officers in the field, we investigated gang activity in several different ways. For the first part of the project,we had to clean up the data, which was ridden with duplication, misspelling and inconsistent formatting. After first identifying which members in the LAPDs database belonged to gangs, and then dividing these gang members into the 31 distinct gangs found in Hollenbeck, we then implemented a string-matching technique in an attempt to find non-identical entries which referred to the same piece of information. Once this data was cleaned up, we performed several different analyses of it. Continuing on the work done by [18], as well as a project performed by [1], part of our analysis involved predicting which subjects belonged to which gangs, and comparing our results to the ground truth (what the LAPD had on file). In [18], 748 distinct gang members in Hollenbeck (see Figure 1) were clustered into 31 clusters, in an attempt to correctly identify the 31 gangs that Hollenbeck is known to house. Our project was similar to these previous works in that we used similar data and attempted to determine which persons were members of which gangs, but there was also great disparity between the two projects. Where [18] focused on spectral clustering, we also implemented the community detection technique known as modularity, and an extension of this, the multi-slice method. Furthermore, beyond merely attempting to correctly place subjects in gangs, we also inspected gang interactions on a macroscale, by investigating intergang activity in Hollenbeck. To do this, we determined 2 types of possible gang relations (rivals, "common enemies) and determined the probability of interactions between gangs of each type of relation occurring at random. We then compared this random distribution to the observed distribution in the field. Several other questions concerning tendencies amongst gangs were posed and are currently under investigation.

1.1 Motivation

Community detection is a technique in mathematics with a wide array of applications. With uses ranging from online networking to macrobiology, the ability to cluster given data points into related groups provides several benefits.

One area in particular that can benefit from accurate community detection is criminology and law enforcement, specifically with regard to gang activity. Knowing which gang members associate with one another, where they interact, and for what purpose, enables police to better evaluate where to be and when, as well

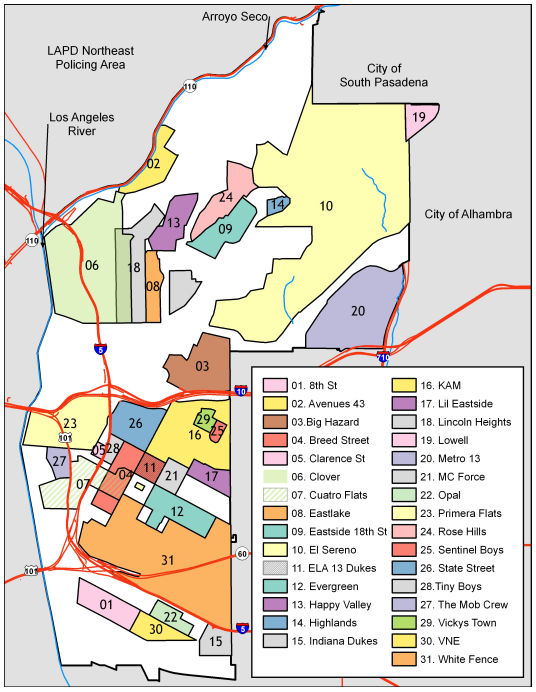

Figure 1: Map 31 Gang Territories in Hollenbeck.[14]

as to whom they should pay special attention. Any extra piece of knowledge can help to protect citizens, reduce violence, and solve or even stop crime. For criminologists, deeper understanding of gang interactions and tendencies enables better analysis of the gang-member psyche, leading to better prevention, reduction and response to criminal activity.

For many urban districts, where gang activity is most prevalent, the sheer volume of residents, along with obvious limits on the intrusiveness of police, makes effective investigation difficult. While methods such as interrogation and undercover work can provide accurate, meaningful information to those interested, such methods require extensive funding, resources and time. Fortunately, alternative methods, the ones we explore in this paper, exist for analyzing gang networks involving large amounts of data. If these methods can shown to be as, if not more, accurate as methods currently used by the LAPD, the time and cost-saving benefits would have extraordinary implications. In addition, the vast amount of data that the LAPD collects over time, which, under current practices, is stored and forgotten, could be instead utilized to provide more intelligence. Furthermore, such methods can also be applied for social networks where status quo strategies (i.e. interrogation and undercover work) are not as feasible, such as with terrorist networks.

One final motivation for this project was to perform a rigorous, successful applied mathematics investigation as members of UCLAs Applied Mathematics Summer REU.

1.2 Hollenbeck

Located east of downtown Los Angeles, the neighborhood of Hollenbeck covers 15.2 square miles and consists of approximately 200,000 residents. Home to 31 distinct gangs 1, Hollenbeck is unique geographically in that some of its borders serve as barriers from other neighborhoods. To the west is the Los Angeles river, to the northeast is Route 710, and to the south is the barely inhabited industrial city of Vernon. Barriers exist within the borders of Hollenbeck, as well: several freeways, including Interstates 5 and 10, as well as Routes 101 and 60, intersect the area in various places. The result is a metropolitan sub-community of particular interest to sociologists and criminologists, as its residents tend to remain constant and are disinclined to socialize with residents of other neighborhoods. Furthermore, the freeways and sharply defined borders of Hollenbeck serve to strictly define many gang territories.

1.3 The Data

Every time an LAPD officer stops a subject, or group of subjects, in the Hollenbeck area, they are required to fill out a Field Incident (FI) card. Essentially index cards, these FI cards contain 61 field entries for information. Some of these entries include first name, last name, suffix, date of incident, location of incident, gang affiliation of subject (if applicable), gang moniker of subject (if applicable), social security number and residential address. Details of gang affiliation and moniker are often trusted at the subject's word, under the assumption that there is social pressure within his or her gang to proudly represent the gang he/she belongs to.

In addition to filling out as much information about the primary subject as they can, police are also expected to fill out basic information about anyone else present at the incident. These "associates" provide first and last name, birthdate, gang affiliation and moniker. Ideally, if there are n individuals at a given incident, police will fill out n FI cards, with each one listing a different individual as the primary subject, with $n-1$ associate subjects also listed. Though this was the case the majority of the time, on several occasions, multiple incident ID numbers were used for the same incident.

In the initial raw data set used for this paper, there were 34,303 entries, ranging in date from 2000 to 2011, though the majority of the incidents occur between 2006 and 2011. However, unlike the data set used by [18], this data set was "dirty": it was comprised of duplicates of individuals, inconsistently formatted entries, typos and missing fields altogether. There are many explanations for the data's being so problematic. First, miscommunication can easily occur in the field: whether an officer incorrectly hears a subject describe their name or gang, or a transcriber, responsible for digitizing the handwritten FI cards, misinterprets an officer's description. Additionally, there are other kinds of mistakes that can be made. For example, officers may spell a name incorrectly. Even if it is spelled correctly, it may be written legibly and transcribed improperly. Finally, even if no mistakes are made, inconsistencies can occur. Officers may use two different phrases to describe the same gang, e.g. "8th Street" and "Eighth St.". Furthermore, a subject may describe himself as Juan Ortiz on one stop, and J. Ortiz on another. Lastly, frequent moving of residence is not uncommon among residents of Hollenbeck. As a result, one person may have several different listed residential addresses, and, thus, may appear as several different people. Because we were looking specifically at the geographic locations of *individuals* as well as the social connections between them, we wanted to minimize these errors as best as possible. Assume persons P and $P\prime$ are one and the same, and while P has been seen with person A, $P\prime$ has been seen with person B. Failure to recognize persons P and $P\prime$ as the same person would cause a failure to relate persons A and B by way of a relationship with $P/P\prime$.

2 Data Cleaning

As described in the previous section, the data were originally collected by the police on FI Cards, then digitized and entered by hand into a large spreadsheet. This spreadsheet holds every FI Card entry, so each event recorded by the police has at least as many entries as persons at the event, and each unique person has at least as many entries as the number of recorded events at which he or she has been observed. Obviously, such a format is not optimal for any analysis.

Our final outputs (so far as cleaning data is concerned) were two data sets, one of distinct events and one of unique persons. The unique persons are linked socially by the events they've mutually attended, and each person's geographic centroid is the mean of coordinates of all their attended events. The following process produces these two data sets, upon which the remainder of our work is based.

2.1 String Comparison

In order to find unique persons, entries in the original data set were compared to each other across several fields, and similarity scores produced for each pairing. However, due to errors and inconsistencies in the data as well as the large observation period during which entries were logged, a unique person FI'd by police several times may have several non-matching entries, such as variations in name (e.g. William and Will) or variations in spelling. To account for this, string similarity metrics were used instead of only exact match comparisons. For two fields of alpha-numeric identification numbers (SSN and Drivers Licence Number), the Jaro-Winkler distance[**?**] was used. For all other fields compared, (first, last, and middle name, suffix, and

gang moniker) the tokenizing scheme softTFIDF [3] was applied, injected with the Jaro-Winkler distance. The method tokenizes the entries, reweights them, and then compares the weighted tokens. Both scores range from zero, 'dissimilar', to one, identical.

Once all fields for all entries in question had been compared using the appropriate method, scores were summed for each entry, then thresholded, where remaining groups of connected pairings (those scoring above the threshold or having some mutual connection above the threshold) were taken as preliminarily correctly identified persons. For this pass, a threshold of 4 out of a possible 7 was used.

A second pass was undertaken if a preliminarily correctly identified person was found to break into several people under slight increases to the threshold. This second pass incorporated an 8th category, birthdate, compared by softTFIDF, and reweighted the addition of the fields, using 65% of the new, 8th category, and 5% of every other comparison category. The threshold for the reweighed comparison was 6.4 out of a possible 8. After this second pass, new identification numbers (so called 'REU numbers') were assigned to each entry, one number for each distinct person, creating the finalized listing of unique persons.

All of these schemes were implemented in MATLAB, and, besides reordering the entries by self identified gang alphabetically and removing any commas or quotations to aid the comma separated values format, all work done to the dataset by the group was done in MATLAB. Our tokenizing scheme builds tokens by splitting strings by whitespace, backslashes, and hyphens, though the softTFIDF algorithm itself does not specify exactly which method and delineation to use. For comparisons between entries with empty fields, the scores were reweighed to only count comparisons between filled fields.

Because of the computational costs of making comparisons between all possible pairings in large groups, entries were first separated into groups by their self identified gang. While this allows for analysis of the most critical parts of the data, a large number of people who did not self identify were excluded. The authors believe that their observed social density (average degree) is lower than that of the gang members, but this cannot be shown without running all their entries through the aforementioned process. If the non-self-identified entries are can be paired down into smaller groups with some confidence such as by the first letter of the last name, or if another implementation of the comparison methods is found to be orders of magnitude faster, then these entries may be cleaned and used.

2.1.1 Jaro-Winkler String Comparison Metric

The Jaro string metric [3] is a measure from zero to one of the similarity of two strings. If two strings are identical, they receive a score of one; if two strings share no common characters (defined below), they receive a score of zero. For strings that are intuitively close but not exact, the metric assesses the 'closeness of such a pair by counting common characters and then, as a subset of that, characters that are common but not in the same position in both strings. More formally, given the strings S_1, S_2, where $S_1 = a_1 a_2 \ldots a_m$ and $S_2 = b_1 b_2 \ldots a_n$, we define the search radius $\mathbf{h} = \frac{\min(|S_1|,|S_2|)}{2}$. a_i contained within S_1 is said to be *in common* with S_2 if there exists b_j contained within S_2 such that $b_j = a_i$ and $i - \mathbf{h} \leq j \leq i + \mathbf{h}$ (i.e. within the search radius of the current position). If we then define λ to be equal to the number of characters *in common* between S_1 and S_2 and t to be defined as the number of *in common* characters a_k where $a_k \neq b_k$ (the number of transpositions) then we can calculate the $JaroDist_{1,2}$ which is defined as $\frac{1}{3}(\frac{\lambda}{|S_1|} + \frac{\lambda}{|S_2|} + \frac{\lambda - t}{\lambda})$

Jaro-Winkler is an extension of this that uses $P =$ the largest common prefix between S_1 and S_2 and $P' = \max(\text{P},4)$. $JaroWinkler_{1,2}$ is defined as $JaroDist_{1,2} + \frac{P'}{10} * (1 - JaroDist_{1,2})$

2.1.2 softTFIDF

To account for inconsistencies in our records, we use the softTFIDF, a token-based similarity function that weights in the rarity of a token in its phrase and in the document as whole. The reason we want to do this is because we find there are some words that occur very often. These records are more likely to be linked together even if they deviate slightly. For example, if we have a corpus of last names with a very common last name, we would like to more heavily weigh the last names that are not common. softTFIDF has two general steps. In the first step, we weigh tokens of phrases being compared using the tfidf weights. The second step is to use a string metric, Jaro Winkler in our case, to decide whether a string is considered a match or not. Whether we consider two tokens to be matching depends on a threshold that we choose.

We considered each category of data as a corpus, which served as the "document" and each entry of a category as a "phrase". First, the phrases were broken up into tokens. These tokens are defined by single words in the phrase. For example "Happy Valley" would contain two tokens, 'Happy' and 'Valley'. After tokenizing the entire corpus, can weigh the distinct tokens in our corpus. This is done using the following equation [5]

$$tfidf_{s,t} = log(tf_{t,c} + 1) \times log(idf_{t,c}) \tag{1}$$

where we are comparing the tokens of phrase s and phrase t. $tf_{t,c}$ is the term-frequency of the token within the two phrases, or the number of times the term occurs within both phrases. $idf_{t,c}$ is the inverse document frequency, of the number of times of the token occurs in the corpus.

Once we have weighed each token, we can compare each token of phrase p to each token with phrase q. We used the Jaro Winkler metric with a threshold of .8. In other words, if two tokens had a Jaro-Winkler similarity score of .8 or higher, the tokens are considered similar. If token i of phrase p and token q of phrase t has a similarity score higher than our threshold,

$$softTFIDF = \sum_{similartokens} (\frac{tfidf(t_p i}{\text{number of tokens in p}} \times \frac{tfidf(t_q j}{\text{number of tokens in q}}) \tag{2}$$

After computing the softTFIDF similarity scores between each phrase within each category, we threshold the sum of the similarity scores.

2.2 Event Consolidation

Consolidating entries into distinct events is much easier than consolidating entries into distinct people. The process is greatly aided by the LAPD's use of FI numbers which link entries to a single event. These are compared for exact matches against each other, and collapsed into a single entry on the listing of unique events. Coordinates and attendees' REU numbers are extracted, and an adjacency matrix constructed by traversing the listing of events and linking any distinct persons seen at the same event. A number of REU numbers and events at this point were removed due to the inclusion of bad coordinates (those obviously outside the United States or not over land), the lack of coordinates, or coordinates quite far outside of Hollenbeck.

2.3 Results

Out of 34303 entries, 8834 self reported gang affiliation. After both passes of softTFIDF/Jaro Winkler and the removal of REU numbers with bad coordinates, 3163 probable unique persons emerged, on average appearing 2.63 $\pm$ 2.94 times with an mean degree of 1.65 $\pm$ 3.17. There were 1632 singletons with no social connections. 22610 distinct FI card number included 2987 events with at least one gang member. After cleaning, a sample of 40 entries with assigned REU numbers were shown to a criminology expert, who found no errors.

2.3.1 Data Sparsity

The data set used for this year's REU includes 5 years of FI card data as opposed to the 1 year of FI card data used in Van Gennip et als [18] paper. We were hoping that the larger amount of data will relieve the problems with sparsity that the 2011 REU encountered. Unfortunately, this is not the case. The data that we have this year is actually more sparse than the data used by last years group (data from only 2009).

In order to investigate the sparsity of our data, we consider sets of all pairs such that an intra-gang pair is a pair that involves member of the same gang and an inter-gang pair involves members of two different gangs. A is the set of all pairs that appear in our data and G, our ground truth set is the set of all possible intra-gang pairs. Between the 3163 individuals, there are 5,000,703 total possible pairs, 287,754 of these pairs are intra-gang pairs and 4,712,942 of these pairs are inter-gang pairs. The sets are represented by the following contingency table.

	G	$\sim$ G
A	1540	106
$\sim$**A**	286214	4712843

The percentage of true positives is represented by the number of intra-gang pairs in A divided by the total number of possible in gang pairs, i.e. $TP = \frac{|S \cap G|}{|G|}$. As it turns out, the percentage of true positives in our data set is 0.535%, substantially less than the 2.66% of true positives found in last year's REU data set.[18] The percentage of false positives in our data set is defined as the number of inter-gang pairs in A divided by the number of pairs in A, i.e. $FP = \frac{|A \cap \sim G|}{|A|}$. 6.43% of the pairs in A are false positives, which is less than that of last years. We also look at the percentage of pairs that are true negatives, or the fraction of possible inter-gang pairs that are not in A, $TN = \frac{|\ A \cap\ G|}{|\ G|}$. It is not surprising to see that 99.99% of the true negatives are captured by our data set. 99.98% of true negatives were found in last years data. The percentage of false negatives is the number of intra-gang pairs that are not in A, ie $FN = \frac{|\sim A \cap G|}{|\sim A|}$. 5.72% of our data is false negatives, which is comparable to the percentage of false negatives found in last years data, 5.56%.

Another way we can measure the sparsity of our data is by comparing the average degree of a node to that of last years. Last year, the average node degree was 1.275 ± 1.894. This year, the average node degree is 1.65 ± 3.18. Of the 3,163 individuals, 1,633 of them are singletons, or people who were only seen by themselves. In other words, more than half of our individuals do not have social connections according to our current social adjacency matrix.

If we suppose in an ideal situation a gang member should appear with every other gang member of his or her, in our real data we observe 0.5% of these connections, compared to 2.66% in the previous year's data.

3 Data Analysis

3.1 Spectral Clustering

The first method that we use to analyze the data is a technique known as spectral clustering. Spectral clustering has many fundamental advantages. Some of these advantages are, spectral clustering is very simple to implement, and can be solved efficiently by fast linear algebra solvers. Clustering is one of the most widely used techniques for data analysis, with applications ranging from statistics, computer science, biology, and the social sciences. In many fields dealing with empirical data, people attempt to get a first impression on their data by trying to identify groups of "similar behavior" in their data [19]. In our data set we are interested in the behavior of the gangs in the Hollenbeck area. Determining the groups into which people organize themselves is a first attempt to understanding their behavior. The social group in which someone belongs to can reveal crucial information [18].

The idea behind spectral clustering is that given a set of data points $x_1 \dots x_n$ and some form of similarity $w_{ij} \geq 0$ between all pairs of data points x_i and x_j, the goal is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. Clustering can also be thought of in terms of graph theory, where we want to find a partition of the graph such that the edges between different groups have very low weights (which means that the points in different clusters are dissimilar from each other) and the edges within a group have high weights (which means that points within the same cluster are similar to each other). In the next section some basic graph notation will be introduced as well as the type of graph that we use in our clustering method.

3.1.1 Graph Notation and the Fully Connected Graph

Let $G = (V, E)$ be an undirected graph where V is the vertex set and E are the edges of the graph. We take G to be weighted, that is each edge between two vertices v_i and v_j carries a non-negative weight $w_{ij} \geq 0$. The weighted adjacency matrix of the graph is the matrix $W = w_{ij}$, where $i, j = 1 \dots n$. If $w_{ij} = 0$ this means that the vertices v_i and v_j are not connected by an edge. Since G is undirected then we require the graph to be symmetric, that is $w_{ij} = w_{ji}$.

The degree of a vertex $v_i \in V$ is defined as

$$d_i = \sum_{j=1}^{n} w_{ij}.$$

The degree matrix, D, is defined as the diagonal matrix with degrees $d_1 \dots d_n$ on the diagonal.
The type of graph that we are interested in is the fully connected graph. For the fully connected graph we connect all points with a positive similarity with each other, and we weight all edges by w_{ij}. The graph should represent local neighborhood relationships. The creation of this graph is only useful if the similarity function models local neighborhoods. The gaussian similarity function is a good example of a similarity function that should be used. This particular graph model is what is used in the calculation of the similarities between points in the geographic component of the weighted affinity matrix.

3.1.2 Method: Self-Tuning Spectral Clustering

For our purposes we construct a graph where each node represents a distinct individual in the Hollenbeck area. From our data set there are 3,163 distinct individuals. Every pair of nodes i and j are connected by an edge and have weight,

$$W_{i,j} = \alpha S_{ij} + (1-\alpha) e^{-d_{ij}^2/\sigma^2},$$

where $\alpha \in [0,1]$, d_{ij} is the regular Euclidean distance between the average stop locations of individuals i and j, and the parameter σ controls the width of the similarity neighborhoods. Choosing the Gaussian similarity function is natural for the geographic measure component of the weighted affinity matrix because we want to look for local neighborhoods into which these individuals organize and we can set the width of the neighborhood to be the length scale where a lot social events take place. We set the social adjacency matrix to be labeled S_{ij}, where S_{ij} is defined by,

$$S_{ij} = \begin{cases} 1 & \text{if i has met j} \\ 0 & \text{otherwise} \end{cases}.$$

One aspect to spectral clustering that is paramount to correct clustering with the geographical aspect is the scaling parameter σ in the Gaussian similarity function. Instead of selecting a single scaling parameter for all sets of data points, we calculate a local scaling parameter σ_i for each data point x_i. The distance from x_i to x_j as seen by x_i is $d(x_i, x_j)/\sigma_i$ while the converse is $d(x_i, x_j)/\sigma_j$. Therefore the square distance d^2 seen earlier may be generalized as $d(x_i, x_j)^2/\sigma_i\sigma_j$. The affinity between a pair of points can now be expressed as,

$$\hat{A}_{ij} = \exp(-d^2(x_i, x_j)/\sigma_i\sigma_j).$$

Using a specific scaling parameter for each of our data points allows self-tuning of the point-to-point distances according to the local statistics of the neighborhoods surrounding points i and j. The selection of the local scale σ_i can be done by studying the local statistics of the neighborhood of data point x_i. An easy choice is:

$$\sigma_i = d(x_i, x_K)$$

where x_K is the K^{th} nearest neighbor of the point x_i. The selection of K is independent of the scale and is a function of the data dimension of the embedding space [**?**]. The K that was chosen was $K = 300$. The weighted affinity matrix can now be expressed as

$$W_{ij} = \alpha S_{ij} + (1-\alpha)\hat{A}_{ij}.$$

The parameter α can be adjusted to set the importance between the social and geographic information that is being used in the weighted matrix W. If $\alpha = 0$ then only the geographic component is being used and if $\alpha = 1$ then only the social component is being used. The social and geographic matrices are 3163×3163 in size. Displayed below are the images of the social and geographical adjacency matrices, in figures 2 and 3, respectively. From the social adjacency matrix and the geographic locations of the average stop locations we can also construct a graph plot of all known social connections displayed in figure 4.

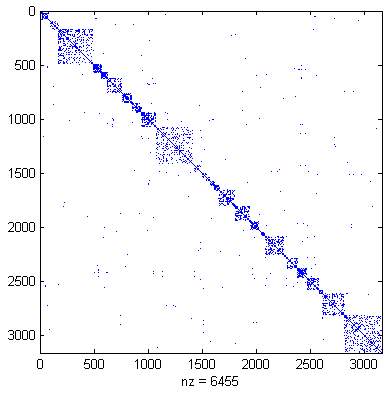

Figure 2: Image of social adjacency matrix. Each blue pixel represents a social connection found in the data. Each block on the diagonal represents a different gang.

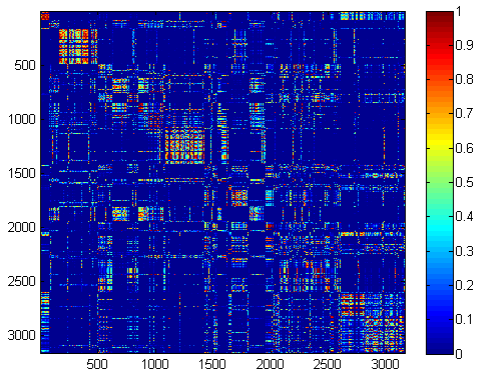

Figure 3: Image of Gaussian similarity function. Similarity between points in based on a color where blue is no similarity and red is full similarity.

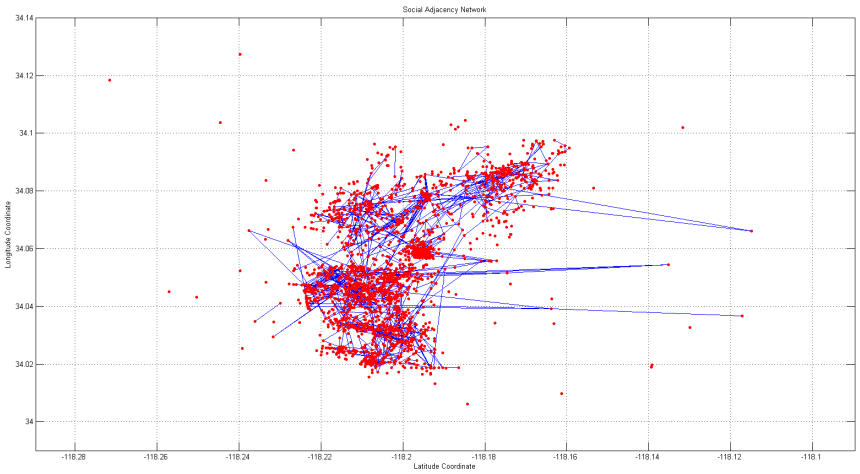

Figure 4: LAPD FI card data showing the average stop location of each individual with social links of whom was stopped with whom.

Using spectral clustering we place the individuals into 31 distinct clusters. The decision for the number of output clusters is 31 and the motivation for this number was determined by observations from the LAPD that there were 31 active gangs in the Hollenbeck at the time that the data was collected.
We use a spectral clustering algorithm for its simplicity and transparency in making non-linearly separable data points separable.

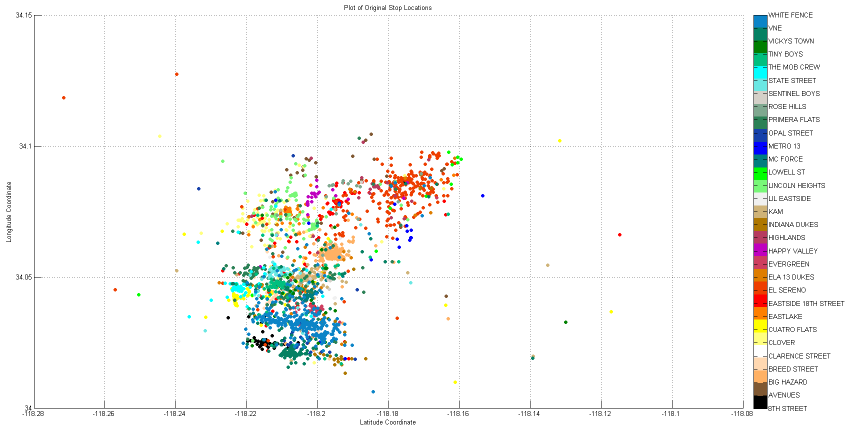

Figure 5: Image of each average individual stop location. Each individual is associated with a gang represented by a distinct color. From this plot it is easy to see why spectral clustering would be beneficial. Many of the data points are overlapping or mathematically speaking they are non-linearly separable.

For calculating the eigenvalues and eigenvectors we use Matlab's *eig* function. The first eigenvalue calculated from the affinity matrix is one and the corresponding eigenvector is a constant vector. A cluster based only on this eigenvector place all of the vertices in one cluster. To avoid this we look into the information of the other non-trivial eigenvectors.

We compute the first 31 eigenvectors (ordered according to decreasing eigenvalues) of the normalized affinity matrix $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, where D is the degree matrix that we discussed earlier. Shown below is a plot of the second, third, fourth, and fifth eigenvectors, respectively.

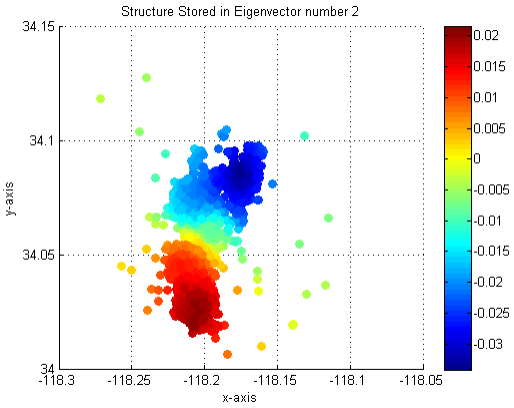

Figure 6: Second eigenvector calculated from the normalized weighted affinity matrix. $\alpha = 0.7$

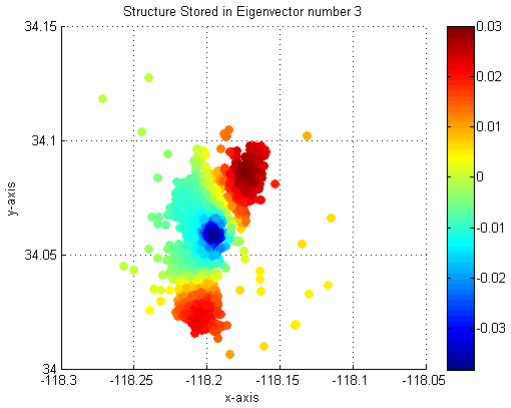

Figure 7: Third eigenvector calculated from the normalized weighted affinity matrix. $\alpha = 0.7$

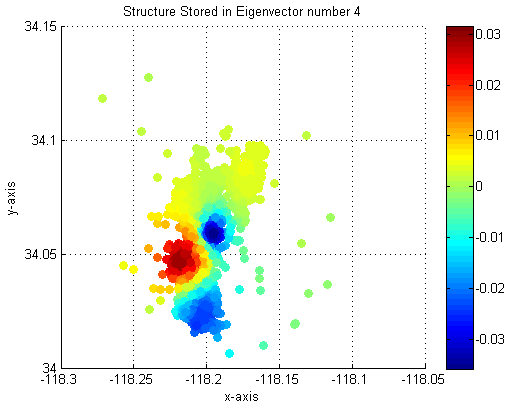

Figure 8: Fourth eigenvector calculated from the normalized weighted affinity matrix. $\alpha = 0.7$

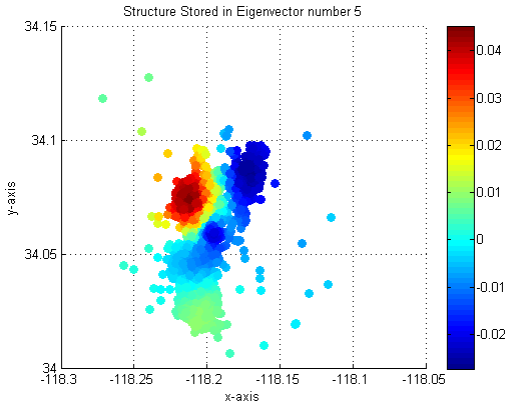

Figure 9: Fifth eigenvector calculated from the normalized weighted affinity matrix. $\alpha = 0.7$

Figure 10: The data from the LAPD FI cards are colored by the eigenvectors. The x-axis represents the latitudes of the data points and the y-axis represents the longitudes of the data points.

As stated before, the first eigenvector is constant and therefore reveals no crucial information about the structure of the data set. However, from the scatter plots above we can see that other non-trivial eigenvectors capture the most prominent structures of our data set. The second eigenvector separates Hollenbeck into three distinct regions from North to South. If one looks into the third eigenvector, it captures the structure of Hollenbeck's three largest gangs, which are El Sereno (represented by red in the North), Big Hazard (represented by the large blue structure in the middle), and White Fence (represented by the red in the South). Observing the fourth eigenvector, the immediate structure that is recognizable again is Big Hazard. The fourth eigenvector also picks up a structure located in West Hollenbeck, which includes gangs, Primera Flats, The Mob Crew, Cuatro Flats, Clarence St., and Tiny Boys. Another structure in the southeast corner of Hollenbeck is captured and this structure contains gangs, White Fence, VNE, Opal, and Indiana Dukes. Finally, from the the fifth eigenvector we can see that again Big Hazard was detected as well as El Sereno. A new structure that is detected in the fifth eigenvector is in the Northwestern half of Hollenbeck, which includes the gangs Clover, Lincoln Heights, Avenues 43, Eastlake, and Happy Valley.

These eigenvectors of the normalized affinity matrix are known to solve a relaxation of the normalized cut (Ncut) problem turn them into binary approximations using the k-means algorithm which iteratively assigns individuals to their nearest centroids in the space spanned by the eigenvectors and updates the centroids after each step [9]. In other words, the data that is embedded in the eigenspace is desirable because the representation of the abstract data points (non-linearly separable points) has a new representation which enhances the cluster-properties in the data so now clusters that were once hard to detect can be detected trivially. In particular, the k-means clustering algorithm has no difficulties to detect the clusters in this new representation. Because k-means uses a random initial drop of the centroids, in the computation of the metrics to determine the accuracy of our clusters, we average over 10 spectral clustering runs.

We want to look at two different questions. First we want to know if we can classify different social structures with information given in the LAPD FI card data set which includes information such as, average stop locations of each individual and their known associates. In particular, do we benefit from adding geographic information to social information? We also look at how well our specific FI card data set performs compared to data from 2009, considering we have much more information spread over many years. The second question we want to answer is how much social information should be used to get the optimal information out of our data set, given that our goal is to identify gang affiliations of the individuals in our data set? [18]

We use a set of metrics that will compare the quality of our clustering results against a set of known gang affiliations.

3.1.3 Metrics: Evaluation of Clustering

The first metric that we will use to compare the quality of our clusters against a ground truth is called purity. Purity is a simple and transparent evaluation measure and is an often used clustering metric [**?**]. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned points and dividing by N, where N is the cardinality of the whole set of clusters. For the context of this document, it is the percentage of correctly classified individuals, when classifying each cluster as the gang in the majority (in that cluster). The equation for purity is given by,

$$Purity(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j|,$$

where $\Omega = \{\omega_1, \ldots, \omega_k\}$ is the set of clusters which represent ground truth and $C = \{c_1, \ldots, c_j\}$ are the clustering results [**?**]. For purity, one should also be aware that purity has bias for a large number of clusters. The next metric that will be discussed is the z-Rand score.

For purity one of the clusterings has to be assigned as the true clustering, this is not the case for the z-Rand score. To define the z-Rand score we first need to introduce the pair counting quantity w_{11}, which the the number of pairs which belong both to the same cluster in our spectral clustering and to the same gang. The z-Rand score z_R, is the number of standard deviations which w_{11} is removed from its mean value under a hypergeometric distribution of equally likely assignments subject to the same numbers and sizes of clusters [18][**?**]. A more intuitive way to think about this is that the z-Rand score is a comparison of our clustering results with a random clustering. If the number of standard deviations is closer to the mean of the random clustering then there is more of chance that our clustering is happening at random. However, if the number of standard deviations is far away from the mean that explains to us that the clustering is not happening randomly and our clustering is an implication of the social and geographic data. The z-Rand score for clustering into true gangs is 1030.2755.

3.1.4 Evaluation of Self-Tuning Spectral Clustering on FI card data set

In the table shown below we show the purity and the z-Rand score using the social adjacency matrix $S = A$ for different α (for each α we give the average over ten runs of the spectral clustering algorithm). It is clear from Table 1, for social information, that is $\alpha = 1$, does not perform well on its own.

On the other hand when we use just the geographic information, $\alpha = 0$, the spectral clustering algorithm performs better but still does not give an optimal value. The maximum purity that is obtained is a mixture of both social and geographic information with $\alpha = 0.9$. The maximum z-Rand score that is obtained is 495.1689 and this is with an $\alpha = 0.7$. Also by the table one can see that a mixture of both the social and geographic information is preferred.

α	Purity	z-Rand score
0	0.4147± 0.0076	486.8918± 158.6175
0.1	0.4185± 0.0174	475.9758± 71.9517
0.2	0.4246± 0.0027	476.940 ± 57.3643
0.3	0.4247± 0.0098	486.7086± 159.7504
0.4	0.4241± 0.0154	487.7757± 199.1029
0.5	0.4255± 0.0221	484.1461± 136.4393
0.6	0.4208± 0.0425	478.7644± 109.4330
0.7	0.4285± 0.0244	495.1689± 61.7816
0.8	0.4279± 0.0156	490.1525± 27.1804
0.9	0.4328± 0.0172	457.7581± 13.8751
1	0.3417± 0.0013	253.3737± 81.9061

Table 1: A list of the mean and standard deviation over ten spectral clustering runs of the purity and the z-Rand score, using S=A. The highest z-Rand score that can be obtained in 1030.2755, for a perfect clustering, and purity is a number between 0 and 1 representing a certain percentage.

One can also observe that the purity ranges from 30 - 40% and the maximum z-Rand score ranges from 200 - 500. This is not unexpected because although we have more data in general to work with compared to the REU last year, the data set used in these calculations is in fact, more sparse [18].
The next figure that is displayed below shows a pie chart of one run of the spectral clustering algorithm, using $S = A$ and $\alpha = 0.7$. In this representation of the gangs we see that there are no clusters that are homogeneous. All of the clusters depicted in this information are in fact heterogeneous. This suggest that the internal organizations of the gangs may be more complex than we had expected. However, recall that in this paper we prescribe the number of clusters to be exactly 31, so gang members may be forced to cluster in ways that may not represent true gang organization. Compared with the result from a previous paper, [18], the fact that social data in the FI card data set is to sparse to stand on its own remains consistent. We also see that adding some geographic information improves the result. Geographic information typically does well on its own but it can also be improved by adding some social data to it.

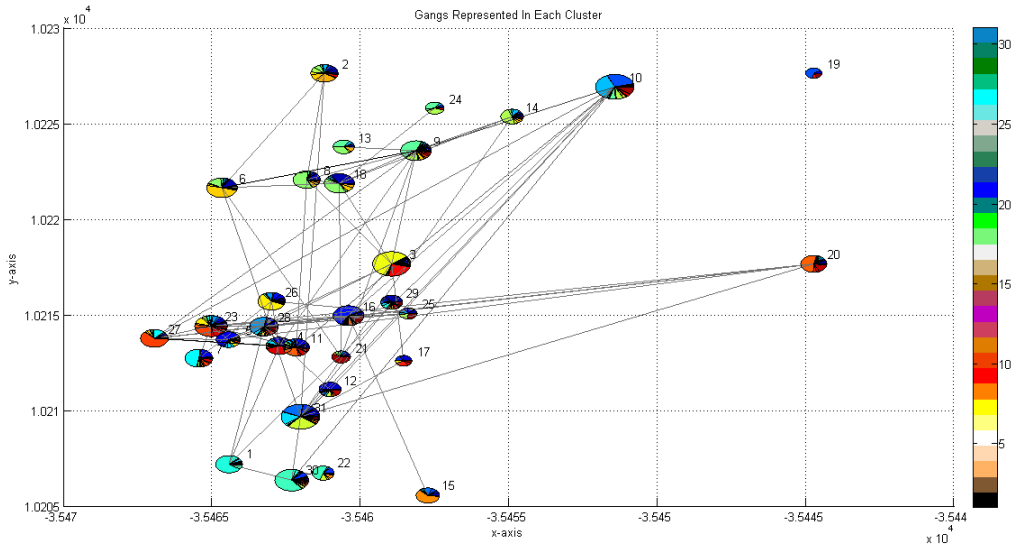

Figure 11: Pie charts made with the code from (reference) for a spectral clustering run with $\alpha = 0.7$. The size of each pie represents the cluster size and each pie is centered as the cluster's centroid. The coloring indicates the gang make-up of the cluster and agrees with the color legend shown above. There are 31 distinct colors that represent each gang. The axes are a multiple of the gang centroids (geographic placement of the clusters in latitude and longitude) on the order of $3 * 10^3$ to avoid overlapping of the pie charts. The connections between pie charts indicate inter-cluster social connections (or non-zero elements of A).

3.1.5 Eigenvector Reweight

To improve both Z-Rand scores and Purity, eigenvectors were normalized and reweighted by their eigenvalues prior to spectral clustering. The authors are of the opinion that this should increase separation between a number of groups. The results presented below show a marked change between the first two rows which are unmodified and normalized eigenvectors' results, respectively, and the reweighted eigenvectors' results in the third row. These results used a fixed sigma of the mean distance between all points.

Z-Rand (top) Purity(bottom) 100 Trials	$alpha = 0.3$	$alpha = 0.5$	$alpha = 0.7$
Unmodified	311.21 ± 7.97 0.0482 ± 0.003	324.00 ± 7.30 0.0488 ± 0.002	310.92 ± 7.70 0.0477 ± 0.003
Normalized	308.90 ± 9.03 0.0489 ± 0.003	300.86 ± 7.72 0.0483 ± 0.003	302.06 ± 7.81 0.0474 ± 0.003
Reweighted	429.08 ± 3.77 0.0496 ± 0.002	431.66 ± 3.96 0.0495 ± 0.002	435.33 ± 3.90 0.0492 ± 0.002

3.1.6 Composite Assignments Scores

The nondeterministic nature of the k-means clustering algorithm, used within the spectral method presented above, is generally averaged across several runs due to the possibility of local minima. However, when k ¿ 2, averaging across several runs does not produce an averaged assignment of each node to a group. Composite assignments, however, were intended to generate such a node by node assignment.

Composite assignments are generated by comparing the number of times each possible pairing is observed, then dividing by total number of observations to produce another similarity matrix with scores between 0 and 1. This is then thresholded and remaining connected graphs taken to be groups. This does not necessarily produce k clusters; in fact, as the stability of k-means decreases (either by the spectral method or in regular space), the composite assignment will be no more accurate than the individual assignments, sometimes much less. In a perfectly random assignment of groups, the composite assignment across several iterations would be composed of mostly singletons; in the opposite case with no local maxima besides the global maxima and (if given sufficient iterations of the method) only one possible k-means assignment, the composite assignment would be identical to the individual assignments, which themselves would be identical. In some intermediate case between the two, those pairings that repeatedly appear would appear in the composite assignment.

For larger groups, even though there is no direct community detection it can be assumed given a large number of iterations of k-means and a sufficiently large threshold, nodes i and j within the same composite group must have been clustered with one another with frequency greater or equal to $(1 - threshold) * DS_{i,j}$, where DS is the degree of separation between i and j.

As seen in Figures INSERT NUMBERS HERE, in general the Z Rand scores for composite assignment increase under normal conditions, but for the reweighted eigenvectors, composite assignments are well below the average Z Rand score of their set. This shows that while increasing the Z-Rand score on average as shown in the last section, the reweighting decreases the stability of the assignments significantly. Even though as a whole most iterations are scoring around the average Z Rand score with the same if not less variance (c.f. INSERT TABLE NUMBER HERE), the groups elicited from the data are not the same across multiple iterations.

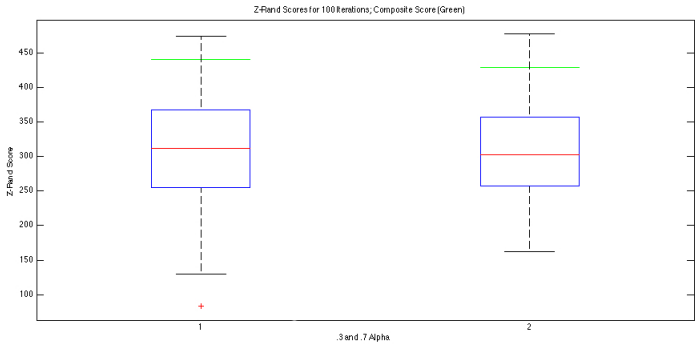

Figure 12: Box plots of the ZRand distribution of results from spectral clustering for $\alpha = .3$ and .7 using a fixed σ equal to the square root of the mean distance between all points. The composite assignment is in green.

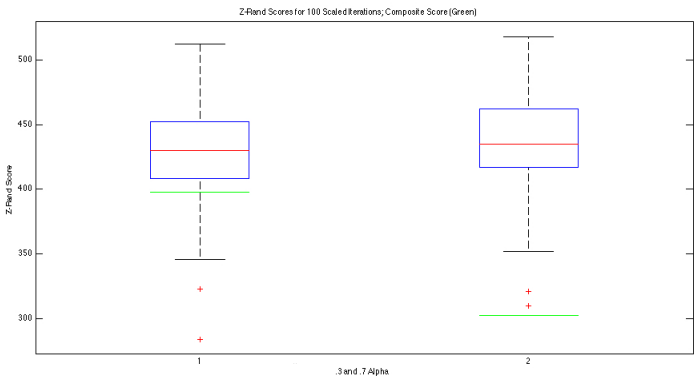

Figure 13: Box plots of the ZRand distribution of results from spectral clustering using reweighted eigenvectors for $\alpha = .3$ and .7 using a fixed σ equal to the square root of the mean distance between all points. The composite assignment is in green.

3.2 Modularity Clustering

Spectral clustering requires the number of desired clusters as an input. Modularity on the other hand, does not have this constraint. Modularity optimization is a community detection method that optimizes the intra-community density of a network. When using modularity to detect communities, it is possible to find that no communities form, which can be an important result that spectral clustering cannot obtain.

Modularity is a graph quality defined as the total weight of edges falling within a group minus the weight of edges in an equivalent network with weighted edges placed at random. [12] The quality function for modularity Q is defined as follows

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(i, j) \tag{3}$$

where $k_i = \sum_j A_{ij}$ and $k_j = \sum_i A_{ij}$ are the degrees of node i and j respectively, A_{ij} is the weight of the edge between node i and j, and $2m = \sum_{ij} A_{ij}$ is the total degree of the graph. [2] For our adjacency matrix, A_{ij} we use **??**.

We want to optimize modularity because the modularity quality function describes the density of a community as compared to the density of the links outside the community. In order to optimize modularity, we use the Generalized Louvain Algorithm [2], which will be described in the next section.

3.2.1 Generalized Louvain Algorithm

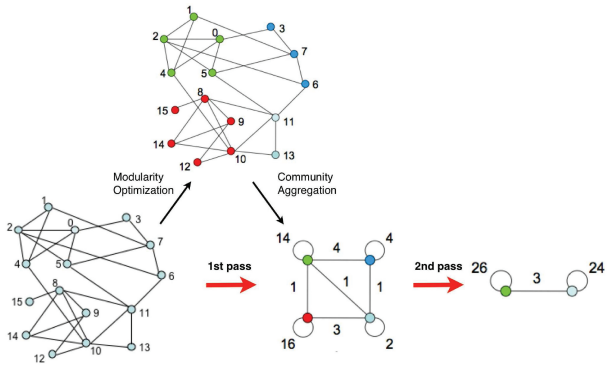

Figure 14: This is a graphic that describes the phase and pass system of the Generalized Louvain Algorithm

The Generalized Louvain Algorithm [2] is an iterative process that is broken up into two phases, which is depicted in 14 First, we start with a network of n nodes so that each node is in its own community.

In phase 1, for each node i we consider node j and calculate the gain in modularity the network would achieve if we were to place node i into j's community. We do this for each node until no positive gain in modularity is possible and then enter into phase 2.

In the second phase, we build a new network using the communities we have created from phase 1. Each community will become a single node and edges will be reweighed. The weight of the of community p to community q is the total weight of edges linking community p to community q. As for the nodes inside the community, a self-loop is created with the total weight of edges within that community. We do not remove the weight of the inter-community edges because our quality function depends on the total weight of the network. [2]

Each iteration is called a pass and passes are made until no positive gain in modularity can be achieved. In other words, every time we leave the first phase, we obtain the local maximum of modularity. When we continue passes until no gain in modularity can be made, we obtain the global maximum of modularity [12].

3.2.2 Performance of Data in Modularity

We used code from Mason Porter et al.'s research group [6] to run modularity on our data set. We use different values of α to evaluate the influence the balance of social and geographical data on our clusters. After averaging 10 runs of the Louvain Algorithm, we averaged our metrics and investigated the cluster created from the highest Z-Rand score.

α	**Purity**	**z-score**
0.0	0.3598 ± 0.0091	224.711 ± 19.552
0.1	0.3623 ± 0.0065	221.524 ± 10.668
0.2	0.3634 ± 0.0065	215.633 ± 7.3484
0.3	0.3576 ± 0.0101	217.275 ± 25.556
0.4	0.3633 ± 0.0060	215.511 ± 10.961
0.5	0.3638 ± 0.0058	219.003 ± 7.7502
0.6	0.3619 ± 0.0090	211.613 ± 22.842
0.7	0.3648 ± 0.0060	217.531 ± 11.512
0.8	0.3675 ± 0.0011	213.923 ± 7.1868
0.9	0.3608 ± 0.0035	224.491 ± 14.189
1.0	0.9712 ± 0.0000	255.309 ± 0.9148

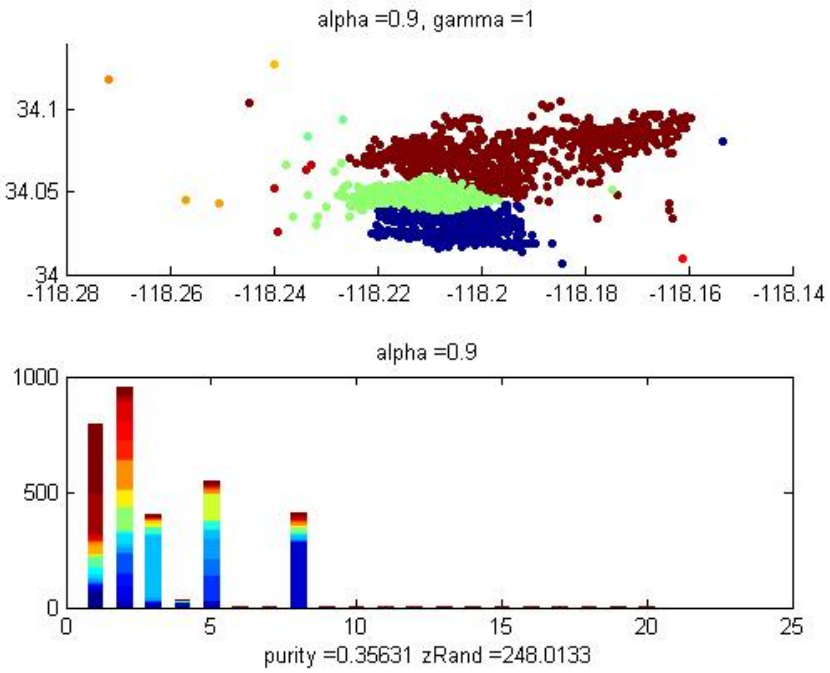

Figure 15: The above scatter plot represents the clusters created using a Generalized Louvain Algorithm on our data set with 90% social data and 10% geographic data contributing to the adjacency matrix A_{ij} The histogram at the bottom of this image represents the gang composition of each cluster. Each color represents one of the 31 gangs. Note that the colors on the histogram do to correspond to colors on the above scatter plot. This was the optimal result. Also note that when $\alpha = 1$, the clusters are small enough such that almost every cluster create is pure, only because every cluster only has one or two gang members.

As we can see in 3.2.2, three main clusters form. These three clusters take up the north, central, and southern sections of Hollenbeck. We attribute the formation of these clusters to the freeways that intersect Hollenbeck, namely the Interstate 10 and Interstate 60. We can also see that looking at 21 the rivalry map of Hollenbeck gangs from [14], these are three sections of Hollenbeck where gangs are most active with each other. According an expert in criminology, it is unlikely for a gang in the northern part of Hollenbeck to have interactions with a gang from the southern section of Hollenbeck.

3.2.3 Improving Modularity Clusters

We removed singletons from our adjacency matrix in an attempt to reduce the sparsity of our data but found that this did not create more desirable clusters. After calculating the percentage of true positives, we see the reason for this is because even without the singletons, our data is still relatively sparse with only 2.24% of the possible intra-gang pairs found, and 6.44% of the pairs were false positives.

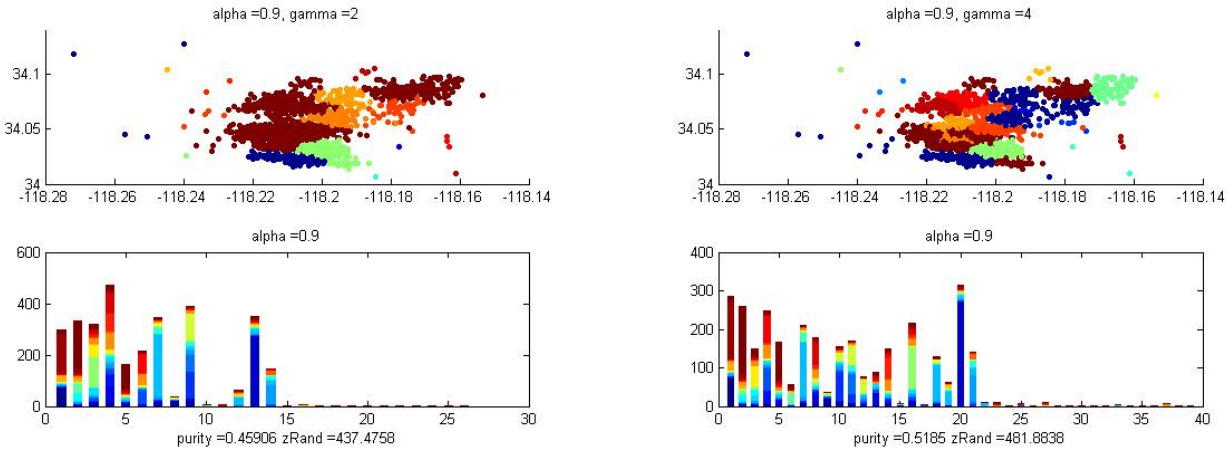

Figure 16: These graphs are the result of modularity when looking over different values of γ, our resolution parameter and $\alpha = 0.9$. We see that the clusters have broken down into more than the three main clusters when the resolution parameter increases.

In order to improve the clusters created using modularity, we use a resolution parameter γ in front of the expected number of edges term. This resolution parameter will allow the algorithm to detect smaller clusters. 16 is the result obtained from the Generalized Louvain Algorithm. We found that as we increased the resolution parameter, we obtained more clusters.

3.3 Multislice Method

The multislice method utilizes modularity for networks with different types of connections by coupling multiple adjacency matrices together. [11] It is of our interest to investigate the results of this method because it will allow us to look at our network over an extra type of connection. In our case, we can look at how our network clusters over each year and over different resolution parameters.

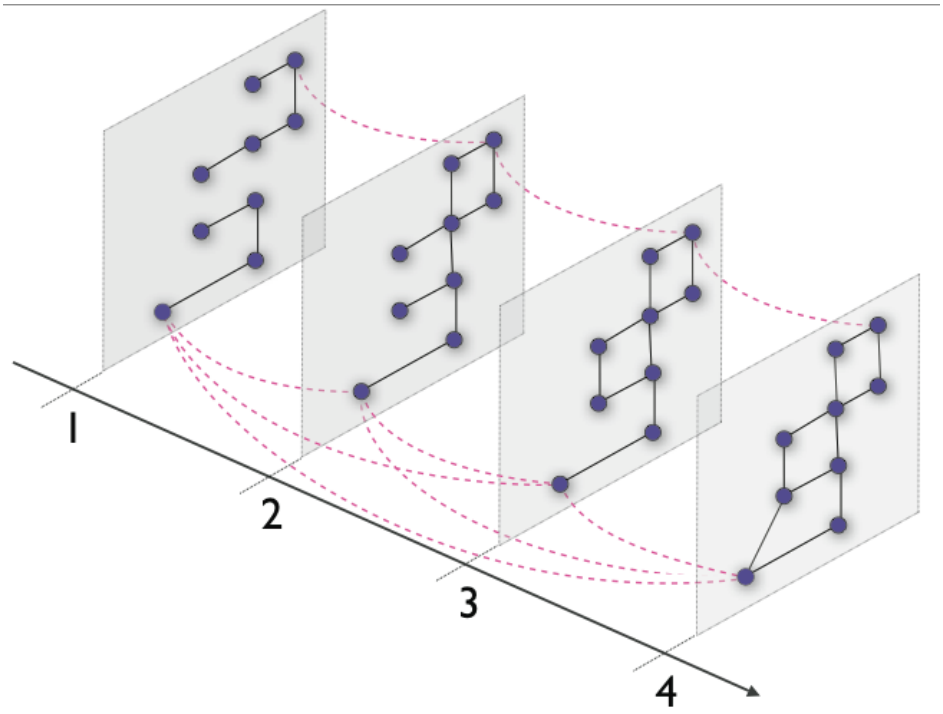

Figure 17: This is a graphic of multiple graphs being connection by between slice nodes in order to illustrate the idea of multislice. Notice that all nodes are present in every network and there only exists between slice relationships between nodes that are the same. [2]

For a multi-slice network, we look at individual networks that are linked together by the same node in each network. In other words, each network in a slice of the multi-slice method must have the exact same nodes. The quality function for multislice networks is very similar to that of modularity:

$$Q_{ms} = \frac{1}{2m}\sum_{ijrs}[(A_{ijs} - \gamma_s \frac{k_{is}k_{js}}{2m})\delta_{sr} + \delta_{ij}C_{jsr}]\delta(g_{is}, g_{jr}) \tag{4}$$

[11] The term $\delta(g_{is}, g_{jr})$ is 1 if node i in slice s and node j in slice r are in the same cluster. The terms inside the summation take are both intra-slice relationship of nodes as well as the inter-slice relationships of nodes. $\delta(s, r)$ is 1 if slice s and slice r are the same slice and 0 if slice s and slice r are not the same slice. A_{ijs} is the weight between node i and j of slice s and $k_{is} = \sum_j A_{ijs}$ is the degree of node i in slice s and $k_{js} = \sum_i A_{ijs}$ is the degree of node j in slice s. This difference only contributes to modularity when we are looking at nodes within the same slice. The inter-slice relationship of nodes in multislice networks are represented by $C_{jsr} = \{0, \omega\}$ such that ω is a constant. $\delta_i j$ is 0 if i is not the same node as j and 1 if i is the same node as j, therefore there only exists inter-slice relationship for the same nodes. The resolution parameter γ has the same purpose as in the previous discussion.

In order to measure the accuracy of the clusters created by the multislice method, we compute the purity and Z-Rand of each slice. Using the multislice method is different from looking at slices individually because the inter-slice relationship imposes consistencies between slices. If a node is in a cluster in the previous slice, it makes the node more likely to stay in the same cluster.

3.3.1 Performance of Multislice Method

In 2007, a gang injunction was imposed on Eastlake, Clover, and Lincoln Heights, all gangs of Hollenbeck [**?**]. A gang injunction is a court mandated restriction on gangs. This injunction prevented these three gangs from hanging in a set area in Hollenbeck. We want see if we can detect any changes in clusters after the implementation of the gang injunction. To do this, we comparing clusters created before 2007 to clusters

created after 2007. We create an adjacency matrix for each year of data that we have and let each of these adjacency matrices be a slice in our multislice network.

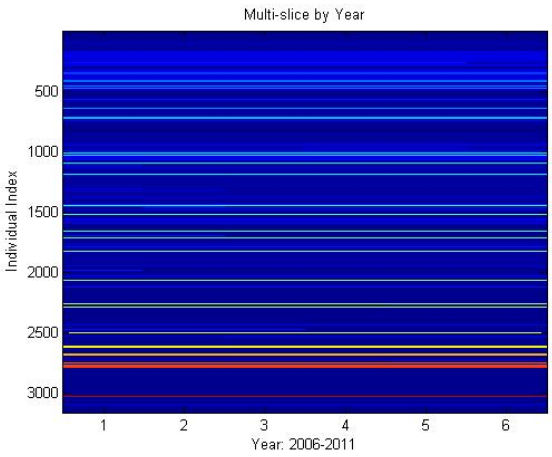

Figure 18: This is a graphic of what happens to each node in our data set over different years. Each color represents a cluster. The y axis is the index of the node and the x-axis is the slice. We are looking at the years 2006-2011. There are very little changes that occur throughout the network under the current slices.

Looking over different values of the clusters formed from the year to year multi-slice was not what we desired 18. The reason for this is because we have still have sparse data year to year. Each year, more than half of our gang members are singletons. Also, each year the same nodes do not appear so in every slice. These nodes do not have any social connections. Due of the lack of social connections, we find that many people are put in their own cluster. Because our nodes were clustered into the same cluster each year, we cannot make a conclusion about the any changes that a gang injunction may have caused.

Next, we investigate our network as a multiscale network. Although it is desirable to use as many slices as possible, we are only able to use four slices at a time due to issues with memory. We also had to threshold our adjacency matrix A_{ij} in order to save memory. We created an arbitrary threshold of 10^{-3}. In future studies, this thresholding value should be further investigated if thresholding is still a necessity. Another option is to only use the weights of the k nearest nodes, which we did not have time to implement.

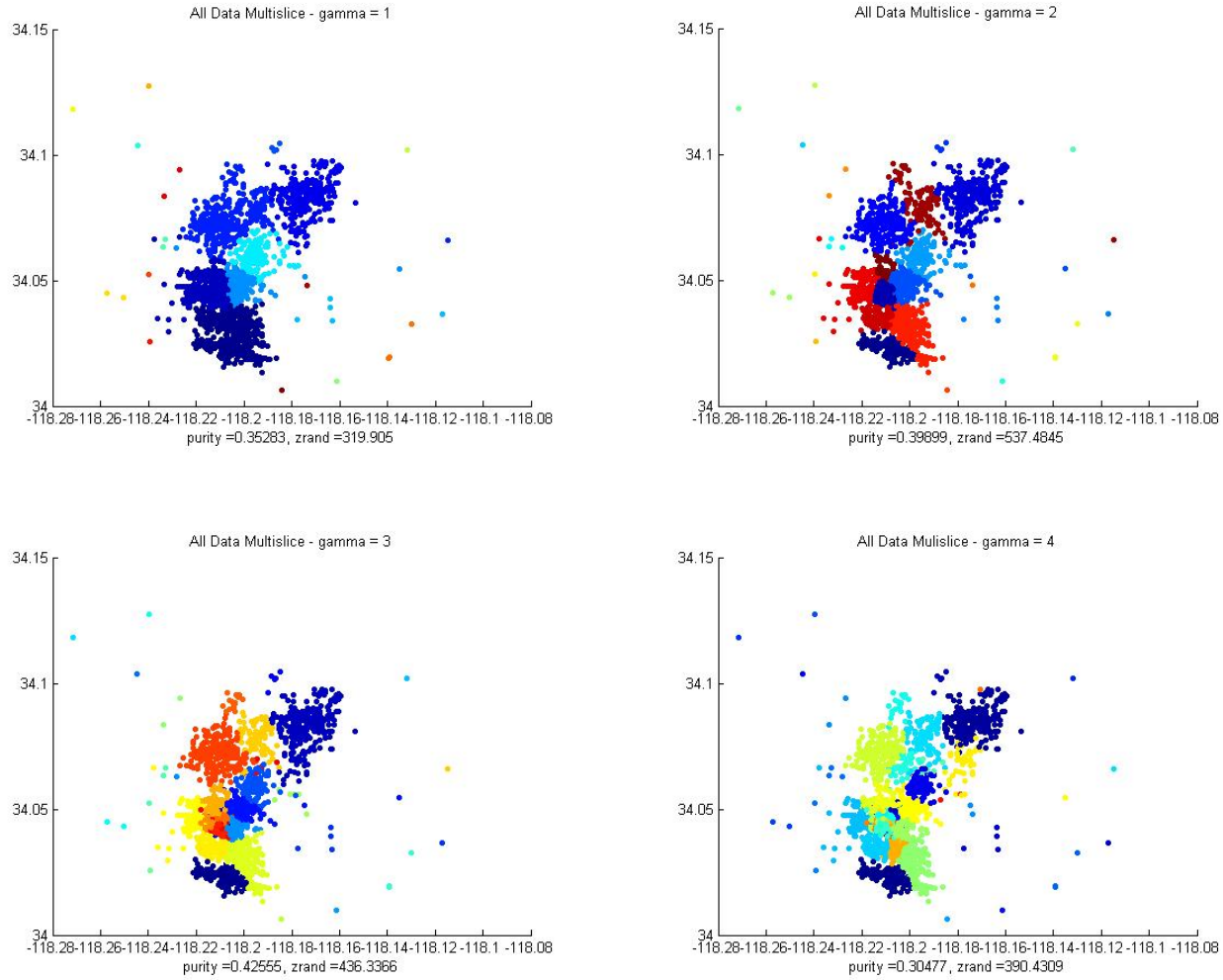

Figure 19: These are the four slices that we investigate in our multislice method. As one can see, there is a big difference between the clusters created with a resolution parameter of 1 (in the top left) and a resolution parameter of 4 (in the bottom right).

ω	**Purity**	**Z-Rand**
0.03	0.4015	533.004
0.04	0.3965	528.562
0.05	0.4015	522.221

Our multislice method over different resolution parameters showed more promising results 19. The Z-Rand score is an improvement over single slices with different resolution parameters 3.3.1. This table displays the results from our optimal result with resolution parameter, $\gamma = 2$. We used four slices, each slice had the same graph but the resolution parameter varies from 1-4. We used trial and error to find our values of ω, which can be improved upon in the future. The number of clusters more closely represented the number of clusters we were looking for. If we look at the number of clusters and Z-Rand score over different resolution parameters, we see that as the resolution parameter increases, the number of clusters increase. We also see that the Z-Rand score reaches its maximum around the right number of clusters ($\gamma = 2$, 28 clusters) 20.

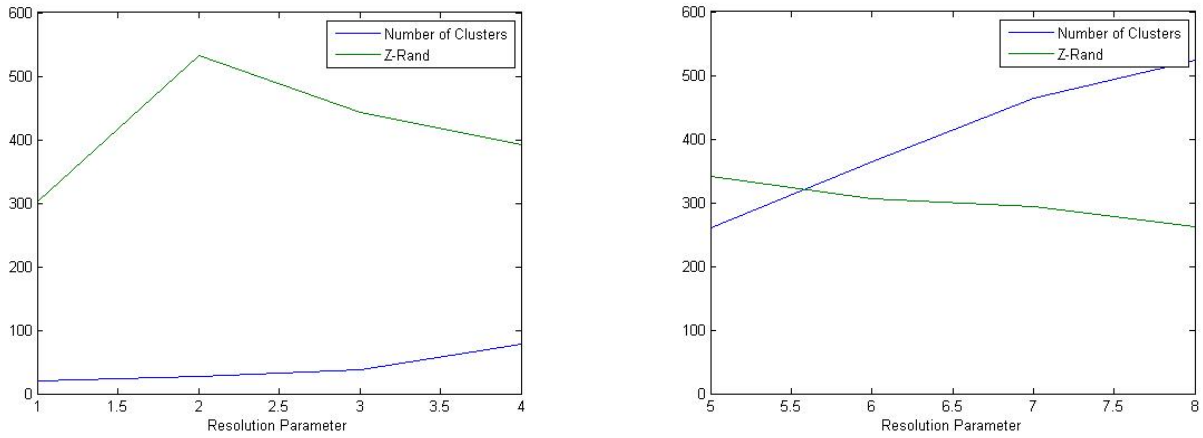

Figure 20: The left plot is a multislice over the resolution parameters from 1-4 and the right plot is a multislice over the resolution parameters from 5-8. The reason they are separated is because they were not run all together in one run, which would create different number of clusters and Z-Rand score. They were not run together due to memory issues. We were only able to run four slices at once. Still, we can see the relationship between resolution parameter, number of clusters, and Z-Rand score.

3.4 Intergang Community Detection

Criminologists and police officers alike are concerned with details beyond who is a member of which gang. Other matters, such as which gangs are rivals, which gangs tolerate each other, and which gangs work cooperatively can also prove to be important information. As the majority of gang violence stems from intergang activity, this knowledge can help police to better anticipate, respond to and prevent gang violence. Several essays have been written specifically studying the rivalries present in Hollenbeck, e.g. [15].

One of the goals of this project was to use the data given by the LAPD to detect and analyze trends of social interaction amongst *multiple* gangs, and try to find answers to some sociological questions: Do members of rival gangs ever spend time together? If they do, are their interactions always violent? If two gangs are not rivalrous, is their relationship one merely of tolerance, or do some gangs work collaboratively? What determines if two gangs partake in a merely tolerant relationship, versus a collaborative one? Answers to these questions serve different uses: criminologists learn more about what it means to be a gang member, police can better understand how intergang tolerance can be encouraged, and mathematicians are provided knowledge about the complicated relationships that may exist within social network data.

3.4.1 Intergang Relations

In order to study the relations between gangs in Hollenbeck, the different possible relation types must be defined. The intuitive place to start is with the most obvious type of relation that gang exist between two gangs, a rivalry. Using 21 where rival gangs are connected by edges, we created a 31x31 rivalry adjacency matrix RR documenting the Rival Relations (RR) in the Hollenbeck area, such that

$$RR_{ij} = RR_{ji} = 1$$

if gangs i and j are rivals. With this method, we determined that there are 61 RRs in Hollenbeck.

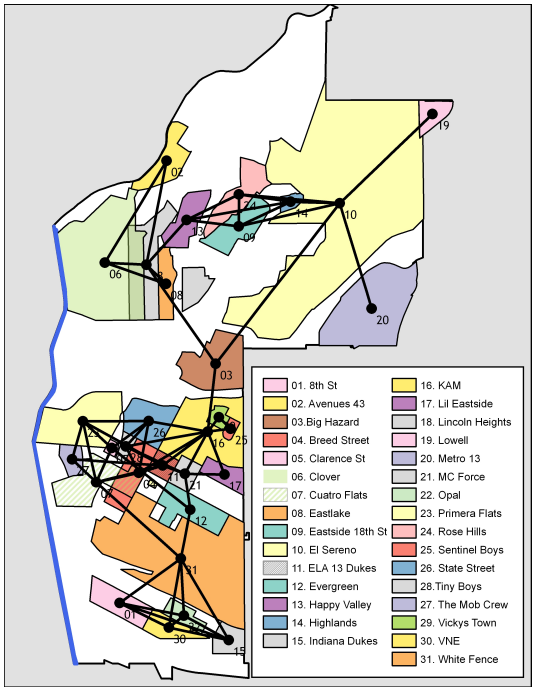

Figure 21: Gang rivalries in Hollenbeck (LAPD).[14]

The next type of relation was inspired by the saying, "The enemy of my enemy is my friend." Interested in examining the possibility of intergang collaboration, we created the Common Enemy Relation (CER). The 31x31 Common Enemy adjacency matrix CER was created such that

$$CER_{ij} = CER_{ji} = 1$$

if gangs i and j are not rivals, and $\exists$ gang k such that $RR_{ik} = 1$ and $RR_{jk} = 1$. In other words, if i and j are both rivals with some k, but are not rivals with one another, we label their relation CER. Again, using 21, we determined there to be 92 CERs in Hollenbeck.

To determine if the incidents observed by the police that featured more than one distinct gang deviate at all from what would be observed at random, we need to determine the probability of two gangs of each type of relation being seen together. As $\binom{31}{2} = 465$, there are 465 different ways that any two gangs can be seen together. Furthermore, as stated above, 61 of these 465 would involve RR gangs, and 92 of the 465 would involve CER gangs. This leaves 312 pairs of gangs that share neither an RR nor a CER. We call these remaining 312 relations Non-Relations (NR). Thus, if the intergang incidents that were actually witnessed in the field were occurring at random, we could say the following: given an incident where two different gangs are present, there is a $\frac{61}{465} * 100\% = 13.12\%$ chance that the relation between the gangs is RR, a $\frac{92}{465} * 100\% = 19.78\%$ chance that the relation is CER, and a $\frac{312}{465} * 100\% = 67.10\%$ chance that the relation is NR.

The next step was to analyze the raw data from the LAPD and determine if the relations in the intergang incidents actually reflected these percentages.

3.4.2 Intergang Activity

As the identification numbers given to each gang member found in the raw data-set began with two numbers identifying the number of said member's gang (i.e. a person with ID number XY-ABCD would be a member of gang XY), filtering the list of events was a fairly straightforward process. Given an event from the full list of events in the large data set, each attendee's REU-ID number was inspected. If an attendee was a member of a gang (i.e. his or her number was of the format XY-ABCD), his or her gang (i.e. XY), was placed in a vector. The gang number for each attendee of the given event was added to the same vector. Next, the

number of unique elements in the "Gangs Present" vector was counted. If this number was greater than 1, then there were at least two gangs represented at the event and the event was transferred to an "Intergang Events" database. After running every event on file through this process, 106 events were detected. However, after doing a manual inspection of each of these events, a number of duplications were found (as a result of their appearing more than once in the initial list of events), bringing the final number of intergang events down to 90. Only one of these 90 events involved more than two distinct gangs, and instead involved 3 distinct gangs. Because we were interested in looking at the relations present at each event, for the 89 incidents involving exactly 2 distinct gangs, there was only 1 relation per event (giving us 89 relations). For the event involving 3 gangs, 3 relations were present (e.g. if gangs A, B and C were present, then there is a relation between A and B, A and C, and B and C). Thus, there were a total of 92 intergang relations observed in the field. The next step was to categorize each of these relations as either RR, CER or NR, and compare the frequencies to what they would be at random.

Iterated through every intergang event, each pairing of gangs present at a particular event (e.g. gangs i and j) were examined in the adjacency matrix, RR. If $RR_{ij} = 1$, the event was categorized as RI, or Rival Incident. If not, it was examined in the adjacency matrix, CER. If $CER_{ij} = 1$, the event was categorized as CEI, or Common Enemy Incident. If the event was neither RI nor CEI, it was categorized as NRI (non-relational incident).[1] At the end of this sorting, there were a total of 27 RIs, 26 CEIs and 39 NRIs. Of the total 92 incident relations, this means 29.55% were RI, 28.41% were CEI and 42.05% were NRI. The difference between these "actual" frequencies of the relation-types and the "expected" frequencies, determined in the above section, are expressed in the following table[2]

Relation Type	**Actual**	**Expected**	**Difference**
RRI	29.55%	13.12%	+16.43%
CEI	28.41%	19.78%	+8.63%
NRI	42.05%	67.10%	−25.05%

Figures **??** and **??**provide a visual of the disparity between what was expected and what was observed in the field. The left figure shows the actual results, while the right depicts what was expected. In both graphs, each bar represents one of the 31 Hollenbeck gangs. The y-axis represents the number of intergang incidents in which said gang was involved. Within a given bar, there are, at most, 3 colors. Blue represents the number of incidents of type RI; green represents type CEI; red represents type NRI.

[1]The event featuring 3 distinct gangs featured two relations of type NR, and one relation of type RR, and, thus, counted 3 times in the categorizing, twice as an NRI and once as an RI.

[2]Statistically significant: 0.1% chance of occurring at random by χ^2 test.

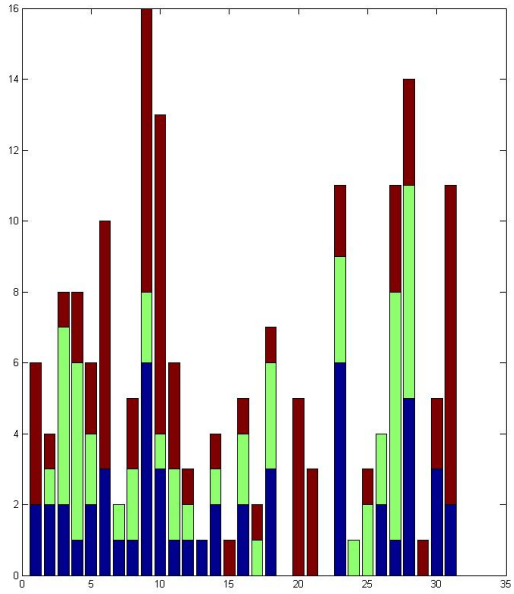

Figure 22: Actual Incident Type Breakdown

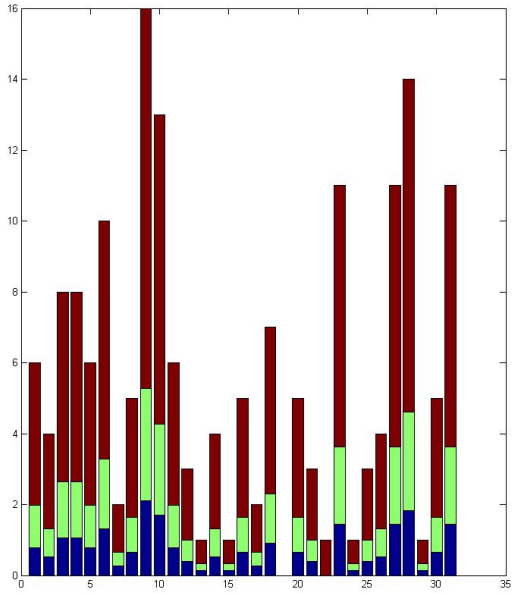

Figure 23: Expected Incident Type Breakdown

Several interesting observations can be made from these results. Most obviously, both rival gangs and gangs with common enemies are seen together far more frequently than one would expect if found together at random. While this is not necessarily surprising in the case of rivals, it is certainly interesting in the case of common enemy gangs. While the rival relationships between gangs were determined by information from the LAPD, the common enemy relationship was one purely fabricated on a hypothesis that gangs are more likely to interact if they share an enemy. However, the fact that these types of gangs are seen together more frequently than expected implies that this theoretical relationship may actually be grounded in truth. If this tendency can be shown to be true, it would undermine a common belief among criminologists: that gangs either hate or merely tolerate one another. Usually, and particularly in Hollenbeck, the idea of alliance or collaboration is considered an anomaly. As our data implies that certain gangs are given preference in interaction, it seems to contradict this general line of thinking.

On another note, while it may not be surprising that rivals are seen spending time together, just *how* they're spending their time may be. After being categorized, the events specifically involving rival gangs were inspected. By examining the incident details provided by police, each event was coded as either hostile or non-hostile.[3] Of the 27 rival incidents, only 2 were determined to be hostile. The remaining were incidents such as drinking in public, loitering, etc. With this in mind, the percentages of each incident type become much more intriguing. If when rival gangs were seen together, it was in instances of dispute, then seeing them together so frequently would seem to make sense. Instead, about 93% of the time that two rival gangs were seen together, they were apparently interacting peacefully. What does this say about the common conception that one is forbidden to interact with a member of a rival gang? It seems to say, quite simply, that this is not the case. However, this is admittedly a very distant analysis, and to assume that there are no other factors at play would be fallacious.

3.4.3 Distance-Based Analysis

What other factors might be at play? In examining a map of Hollenbeck, one might easily notice a defining characteristic of the metropolitan area: two freeways, Route 60 and Interstate 10, intersect Hollenbeck

[3]This process assumes that the police officers make a proper assessment of the incident upon arrival, fully express their assessment on the FI card and that their assessment is correctly interpreted.

horizontally, dividing it fairly evenly into 3 sections. As a result, no rivalries traverse these geographic barriers. Thus, we should expect that RRs tend to cover less geographic distance than CERs and NRs. Furthermore, because rivalries are contained within these three "pockets" of Hollenbeck, it is more likely for two gangs in the same pocket to share a rival than two gangs in different pockets. Again, this would imply that CE gangs, while farther apart than rivals, are closer to one another than two gangs with no relation. In fact, a rough mathematical verification of this can be performed. If we measure the average Euclidean distance between two gangs involved in each type of relation[4], we see the following:

Relation Type	**Average Distance**
RR	0.0105
CER	0.0185
NR	0.0407

Clearly, as explained above, there is an increase, on average, of the distance between two gangs, depending on what their relation is, a factor that has, so far, gone unaccounted for in this report. Perhaps the only reason that there are so many more events involving rivals is because rivals happen to be closest together. Thus, the likelihood is higher than two rivals know each other from a common job, or because they live on the same block, etc. This might also explain why we see CEIs occurring more frequently than expected, and NRIs less frequently. A deeper investigation into the impact of geographic distance and intergang trends is planned for the near future.

3.4.4 Future Analysis

With the vast amount of information contained in these intergang events, we are now only beginning to scratch the surface. Many more mathematical, criminological and sociological questions remain to be answered, and many remain to even be asked. One question that we have begun a preliminary investigation into concerns the relationship between a gang's size and the relations that it has with other gangs. In theory, one might think that larger gangs would tend to have more rivals, as they often take up larger areas and thus share more borders with other gangs. Consequently, large gangs might be expected to have fewer CERs, as the more rivals a gang has, the fewer remaining gangs are eligible for a CER. In addition, we might expect that smaller gangs have a larger number of CERs, due to the fact that they are in greater need of collaboration and resources to defend themselves and compete against larger gangs. A small step has been taken towards beginning to investigate what exactly the relationship between a gang's size and its associations: assuming that a gang's size is accurately represented by the number of incidents involving someone from said gang, we plot the difference between the actual and expected CEIs, divided by the number of times that a gang is involved in an incident. This division is performed to account for the extremely variable gang sizes in Hollenbeck. A gang with 20 members having one more CEI than expected is certainly more significant than have one more CEI than expected in a gang of close to 50 times as large. The result is as follows:

[4]The distances measured were those between the coordinates of the centroids of each gang involved. Centroids are located at the geometric center of a gang's territory, and are shown as nodes on any of the included maps of Hollenbeck.

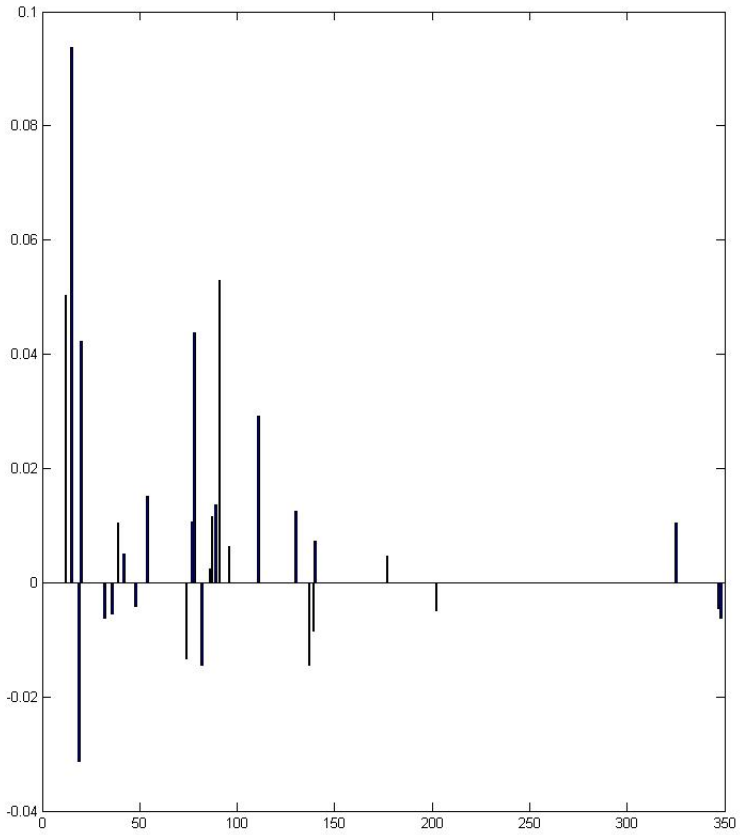

Figure 24: (Actual CEIs - Expected CEIs)/Number of Incidents Involving Gang.

Though no clear trends are visible from this bar graph, new insight is provided when we, instead of looking at individual gangs in Hollenbeck, bin the gangs and *average* their ΔCEIs. Because we are looking at small gangs, medium gangs and large gangs as categories, this may make more sense. The following bar graph shows the average ΔCEI, where the bars are increments of 100 stops. In other words, the first bar represents the average ΔCEI of all gangs that were stopped between 0 and 100 times, the second bar represents the average ΔCEI of all gangs that were stopped between 100 and 200 times, etc.

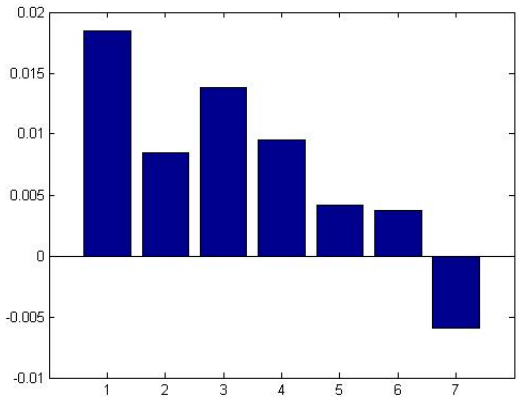

Figure 25: (Actual CEIs - Expected CEIs)/Number of Incidents Involving Gang.

In this depiction, there is a clearer implication that, as gangs get larger, they have fewer CEIs than they would at random. While this is a visually convincing graph, the actual data has not been rigorously statistically verified, which is something we intend to look into in the near future.

4 Future Work

4.1 Inclusion of Unused Data

Though the data used in the presented work is sufficient to perform operations upon, the large number of non-self-identifying persons which were not used may still hold relevant and useful information, especially when building social connections between gangs and when analyzing police bias and/or FI record patterns, an issue not addressed here.

Furthermore, the use of actual stop locations in the calculation of geographic proximity may aid in clustering. While not difficult to compute theoretically, the memory demands are quite high. These demands would be reduced as the number of persons is reduced by finding the unique persons in the non-self-identified entries, but prototype code took in excess of 24 hours to complete, and so it is expected that algorithmic changes need to be made. The authors believe that holding a block of memory for the distances between each event and a block for the comparisons for each person and the listing of people and events causes the processors to 'thrash', i.e. reload the memory caches for each iteration of some loop. It should be noted, however, that this is unproven and difficult to observe.

The construction of a binary social adjacency matrix may be able to be improved upon with the inclusion of the non-self-identifying persons. While preliminary tests upon just the cleaned data showed no improvement or worse performance from various social adjacency schemes taking into account more than directly observed contact between persons, with the inclusion of more connections and more people this might improve. This may have interesting sociological implications a well, such as the average degree of separation between gangs socially, a statistic which may become relevant with a less sparse graph.

4.2 Temporal Analysis

While the data is sparse, there may be some advantage to computing a temporal distance between events as well as a geographic distance. Because the data spans such a long period, connections may be inappropriately draw between two groups meeting at the same place months or years apart. Furthermore, preliminary analysis showed a large increase in data during summer months. There may be sociologically significant information about different age groups and their appearance in the data that accounts for this.

Gangs have cliques within their gangs that are usually composed of gang members around the same age [10]. A possible use of the multislice method is finding the cliques within groups because there is no constraint in the number of clusters the method must produce.

5 Discussion and Conclusions

This years Social Networks REU Program received more than five years of FI Card data in hopes that the larger amount of data would reduce sparsity. Our first task was to clean up the data, as there were inconsistencies and repeated persons in the data set. To do this, we implemented the softTFIDF method as well as the Jaro Winkler metric. After processing the original entries, we had a relatively clean data set to work with for purposes of clustering our network.

In an effort to cluster our data into the 31 gangs of Hollenbeck, we performed a graph partitioning method, spectral clustering, and a community detection method, modularity on our data. These methods did not produce the clusters we were hoping for. When we investigated the sparsity of our data, we found that our data is much more sparse compared to last years REU data. We also investigated types of gang interactions to draw conclusions about the sociological relationships between gang rivalries or lack there of. In doing so, we concluded that there exists some meaning behind the collaboration of members of two different ganga.

We also applied the method of spectral clustering to the LAPD FI card data set. Using the social and geographic information in the data set we were able to cluster all the individuals into various social groups. We showed that the geographic information as a stand alone yields a purity of about 41% compared to the ground truth gang affiliations provided by the Los Angeles Police Department. We also have seen that adding social data to the geographic information can improve our results. The data that was used in the REU this year is even more sparse than the data that was used from 2009 [18]. The purity from the trials of spectral clustering performed on this data set had a range from 30-40%, compared to last years 50-60%. This is not abnormal given the sparsity of the newer data set. There have been ideas proposed to use FI card information including members of these data who did not self-identify with a gang. The goal of looking into this information is to be able to create social connections that were not there before which could reduce the sparsity of the social connections. One of the questions that we had set out to answer was: is it possible to identify social structures in human behavior from limited observations of individuals in Hollenbeck? Or more specifically do we benefit from adding social data to geographic data? Considering the sparsity of the data set it is still hard to answer this question. Although it was shown earlier that adding geographic data to social data increased cluster accuracy, this increase was not large.

6 Acknowledgements

We would like to extend our sincerest gratitude to Yves van Gennip and Blake Hunter[5], our mentors who oversaw the operation of this project. Also, to our criminology expert, Matt Valasik[6], and other mathematicians Huiyi Hu and Cristina Garcia[7].

Furthermore, we would like to thank Andrea Bertozzi [8] for organizing this summers REU Program and overseeing all of our work. Thank you, also, to last years REU group who provided an excellent jumping off point for this years project.

We would like to thank George Tita [9] for keeping us motivated and being an excellent informational resource.

We also want to acknowledge the Los Angeles Police Department (LAPD) for allowing us access to their information for the production of this project, as well as the undergraduate students at UCI who helped to initially clean the raw data from the LAPD.

[5]Applied Mathematics Professors at UCLA.
[6]Ph.D. student in Criminology at UCI.
[7]Ph.D. students in Applied Mathematics at UCLA
[8]Head of Applied Mathematics Dept. at UCLA.
[9]Professor of Criminology at UCI

References

[1] Raymond Ahn, Peter Elliott, and Kyle Luh. Social network clustering: An analysis of gang networks. *UCLA Technical Report.*, 5 August 2011.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unforming of communities in large networks. July 2008.

[3] William W. Cohen, Stephen E. Fienberg, and Pradeep Ravikumar. A comparison of string distance metrics for name-matching tasks. *American Assoication for Artifical Intelligence*, 2003.

[4] Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[5] Melanie Herschel and Felix Naumann. An introduction to duplicate detection. *Synthesis Lectures on Data Management*, 2:1–87, 2010.

[6] Inderjit S. Jutla and Peter J. Mucha. A generalized louvain method for community detection implemented in matlab, 2011.

[7] Kristina Lerman, Rumi Ghosh, and Shang-Hua Teng. The impact of dynamic interactions in multi-scale analysis of network structure. 2012.

[8] Kristina Lerman, Rumi Ghosh, Konstantin Voevodski, and Shang-Hua Teng. Non-conservative diffusion and its application to social network analysis. 2011.

[9] Jitendra Malik and Jianbo Shi. Normalized cuts and image segmentation. *Transactions on PAttern Analysis and Machine Intelligence*, 22, August 2000.

[10] Joan Moore, Diego Vigil, and Robert Garcia. Residence and territoriality in chicano gangs. December 1983.

[11] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks.

[12] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103, June 2006.

[13] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[14] Steven M. Radila, Colin Flinta, and George E. Tita. Spatializing Social Networks: Using Social Network Analysis to Investigate Geographies of Gang Rivalry, Territoriality, and Violence in Los Angeles. March 2010.

[15] George E. Tita, K. Jack Riley, Greg Ridgeway, and Peter W. Greenwood. Reducing Gun Violence: Operation Ceasefire in Los Angeles. February 2005.

[16] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Comparing community structure to characteristics in online collegiate social networks. 53, 2011. 526-543.

[17] Amanda L. Traud, Eric D. Kelsic, Peter J. Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. August 2011.

[18] Yves van Gennip, Blake Hunter, and et al. Community detection using spectral clustering on geosocial data. Submitted, June 2012.

[19] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.