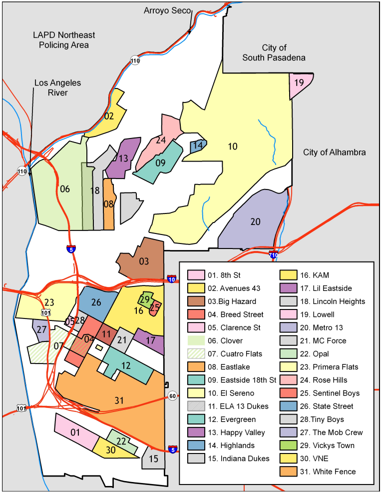
Cleaning Up the Neighborhood: Duplicate Detection and Community Analysis of Hollenbeck Gangs

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

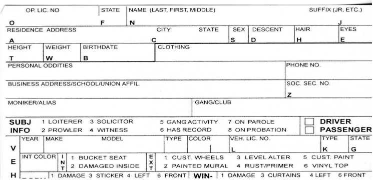
August 8, 2012

Hollenbeck



- 200,000 residents, 15.2 square miles
- 19 miles east of UCLA
- Home to 31 distinct gangs
- Bordered by Los Angeles River, Vernon, and several freeways
  - Creates social insulation making it desirable for sociological study

Data Collection


OP. LIC. NO | STATE | NAME (LAST, FIRST, MIDDLE) | SUFFIX (JR, ETC.)
O | F | N | J
RESIDENCE ADDRESS | CITY | STATE | SEX | DESCENT | HAIR | EYES
A | C | S | D | H | E
HEIGHT | WEIGHT | BIRTHDATE | CLOTHING
T | W | B
PERSONAL ODDITIES | PHONE NO.
BUSINESS ADDRESS/SCHOOL/UNION AFFIL. | SOC. SEC. NO.
Z
MONIKER/ALIAS | GANG/CLUB
SUBJ INFO 1 LOITERER 3 SOLICITOR 5 GANG ACTIVITY 7 ON PAROLE DRIVER
2 PROWLER 4 WITNESS 6 HAS RECORD 8 ON PROBATION PASSENGER
V E H | YEAR | MAKE | MODEL | TYPE | COLOR | VEH. LIC. NO. | TYPE | STATE
L | K | G
INT COLOR | INT 1 BUCKET SEAT 2 DAMAGED INSIDE | EXT 1 CUST. WHEELS 2 PAINTED MURAL 3 LEVEL ALTER 4 RUST/PRIMER 5 CUST. PAINT 6 VINYL TOP
1 DAMAGE 3 STICKER 4 LEFT 6 FRONT | WIN- | 1 DAMAGE 3 CURTAINS 4 LEFT 6 FRONT

- Every time the poilice stop to talk to someone, they fill out a "Field Interview (FI) Card".
- Includes Name, Address, SSN, Gang Affiliation, Moniker, Location of stop, etc.
- Gang members are typically honest about gang affiliation.
- This data was collected, stored, and given to us, by the LAPD

Task 1: Data Cleaning

- Miscommunications, mistakes, and inconsistencies in data
  - eg. "Aug 18 2007" vs "18-08-07"
- Need to eliminate any duplicates to create most accurate social data
- Very large initial data set - over 34,000 entries!

Task 2: Data Analysis

- Spectral Clustering
  - Our runs are modeled after Van Gennip and Hunter et al. and 2011 UCLA REU
- Modularity:
  - Implement another clustering algorithm and compare its results to spectral clustering
- Intergang Communities:
  - Analyze incidents involving different gangs

Data Cleaning

- Initially provided a large excel sheet
  - 34303 Entries, 71 fields
  - Each entry is a single entry on an FI Card
  - Want to identify duplicate entries of people

Last	First	M.I.	OLN	GangAff
Bruin	Joe			C.E. Young Crew
Bruin	Joseph	D.	E123456	Charles E. Young Crew
Trojan	Tommy	A.	N654321	SoCal Uni

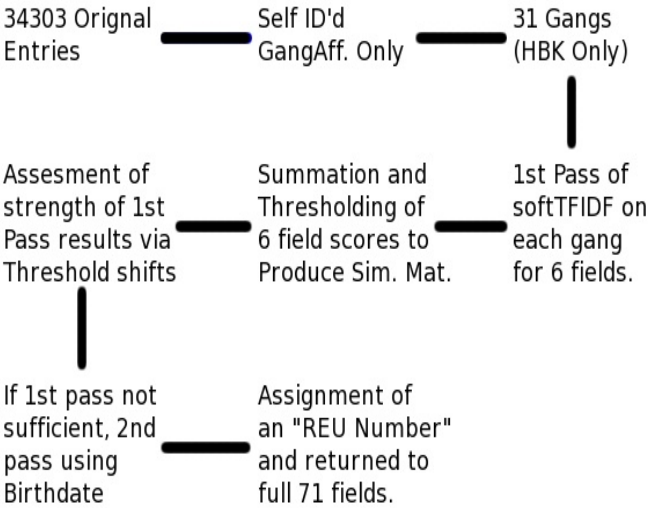
Matching People

- Want to match Joe, Joey, and Jeoy; but also Shadow, Ghost Shadow, and Shadow/Killer
- Jaro-Winkler distance

$$JaroDist_{1,2} = \frac{1}{3}\left(\frac{\lambda}{S_1} + \frac{\lambda}{S_2} + \frac{\lambda - t}{\lambda}\right)$$

- Tokenization via softTFIDF scheme and then application of Jaro-Winkler

Matching People - cont.



Results

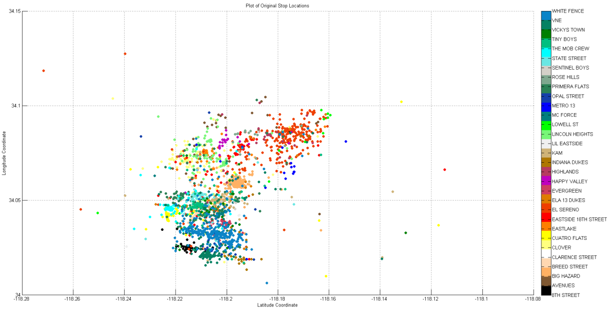
- 34303 entries —> 8834 self reported gang members—> 3163 unique gang members
- 22610 distinct FI card numbers —> 2987 events (with at least one gang member)
- **Sparsity of Data**
  - 1633 singletons (never seen with another gang member)
  - $\sim 0.5\%$ expected intragang connections observed
    - Last year: 2.66%
  - Average degree per person: 1.65 ± 3.17

Spectral Clustering

Why Spectral Clustering?

- It is simple to implement
- Can be solved efficiently
- Applications ranging from statistics, computer science, biology, and social sciences
- Determine the communities into which gang members in Hollenbeck organize themselves because it is an important step to determining their behavior
- Extend on last year's REU paper with hopes of less sparse data and therefore better results

How it works

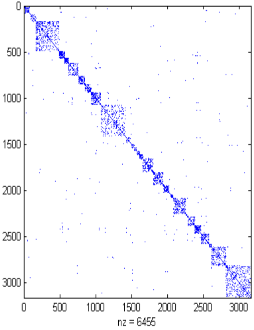


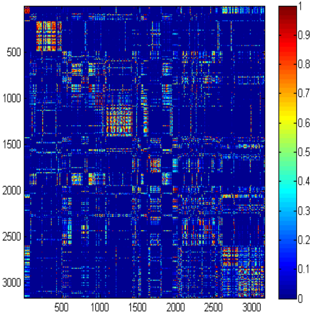
- Goal: divide data points into distinct clusters
- Create a normalized affinity matrix that includes both geographic and social data
- Compute the eigenvectors of the affinity matrix
- Use k-means to separate the data into distinct clusters
- inbed data points in space spanned by first k eigenvectors

Normalized Affinity Matrix

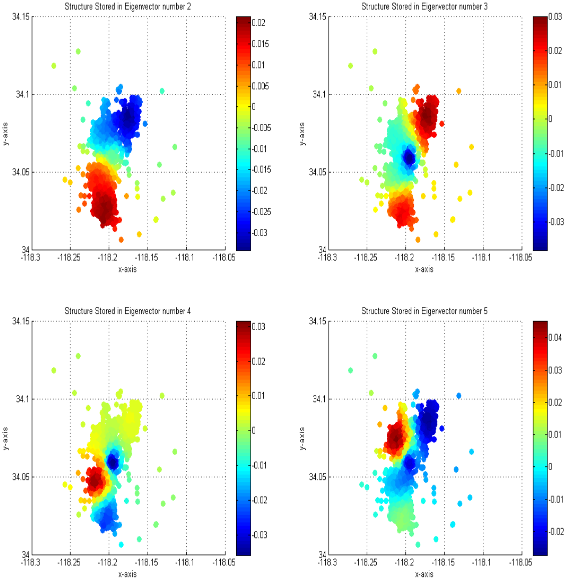
$$W_{i,j} = \alpha S_{i,j} + (1-\alpha) e^{-d_{i,j}^2/\sigma_i \sigma_j}$$

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

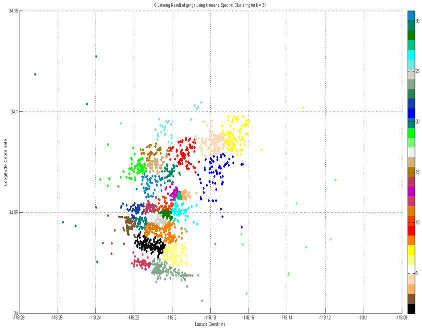




Clustering Structures Embedded in the Eigenvectors



Results of Spectral Clustering Algorithm

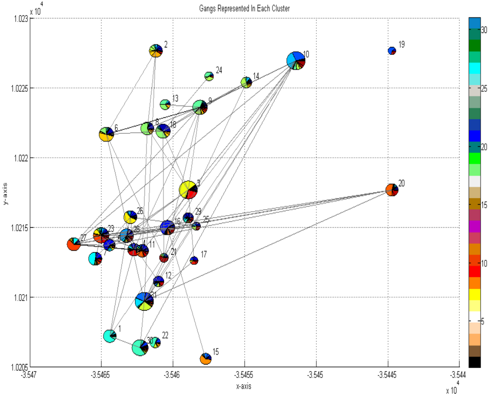


$$Purity = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

Z-Rand: the number of standard deviations which $\omega_{1,1}$ is removed from its mean value under a hypergeometric distribution of equally likely assignments

Reference Z-Rand: 1030

Clustering Gangs in Hollenbeck



Results for this particular plot

$$\alpha = 0.7, \text{ Purity} = 42.85\%, \text{ Z-Rand Score} = 495.1689$$

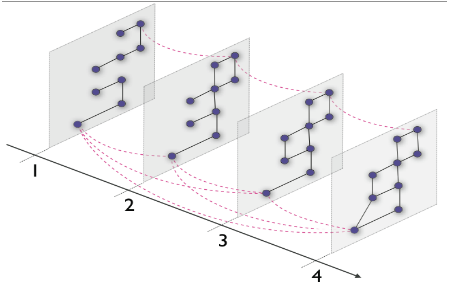
Modularity Method

- Why modularity?
- Modularity: The number of edges falling within groups minus the expected number of edges placed at random

$$Q = \frac{1}{4m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(i,j)$$

- For $A_{i,j}$, we use an adjacency matrix similar to the one we use in spectral clustering
- Newman 2006
- Used code from Mason Porter et al.

Multiplex Method

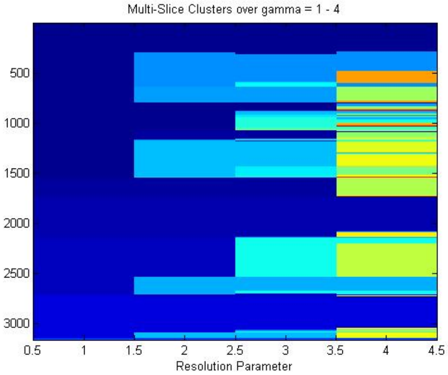


- Multi-Slice: Multi-slice method utilizes modularity for networks with different types of connections by coupling multiple adjacency matricies.

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

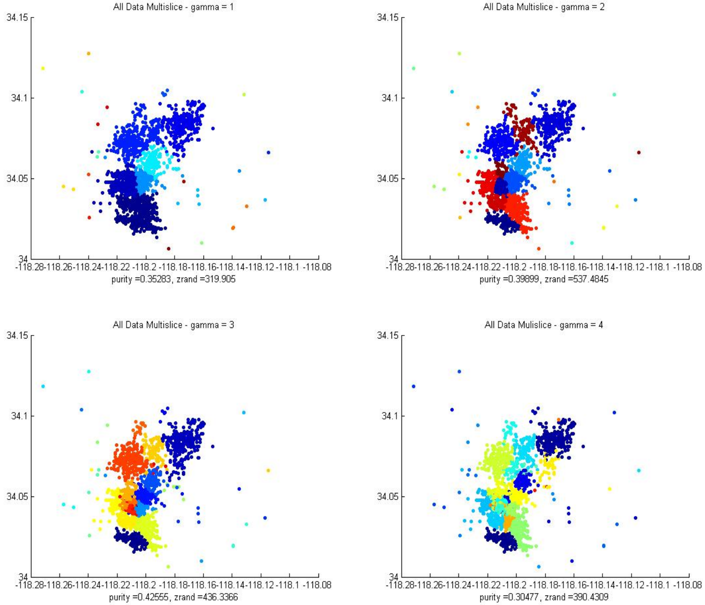
- Why Multi-slice?
- Traud et al. 2011

Multi-slice vs Modularity



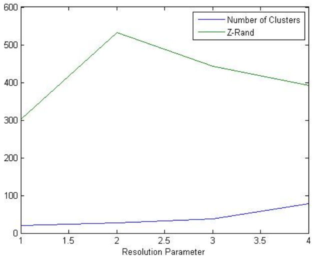
- Multi-slice method allows you to impose consistencies between slices.

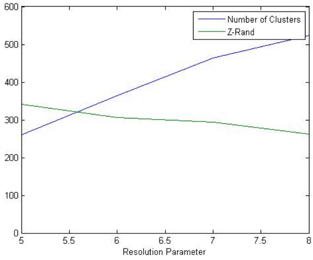
Multi-slice by resolution parameter



Performance of Multi-slice

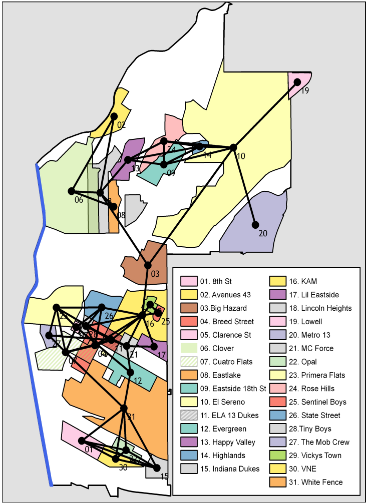
- We can see the clusters breaking up as resolution increases.





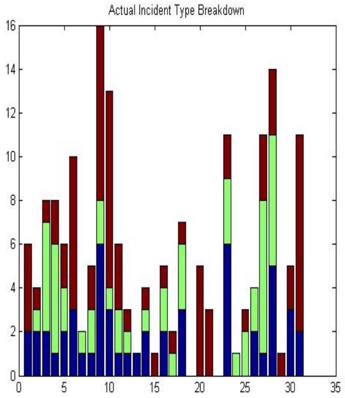
Intergang Relations

- $\binom{31}{2} = 465$ pairwise gang relations
- 61 are Rival Relations (*RR*)
  - By map
- 92 are Common Enemy Relations (*CER*)
  - "The enemy of my enemy is my friend"
- 312 are Non-Relations (*NR*)
  - The rest



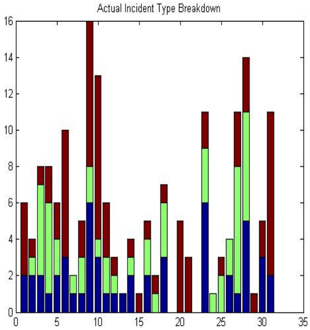
Intergang Incidents

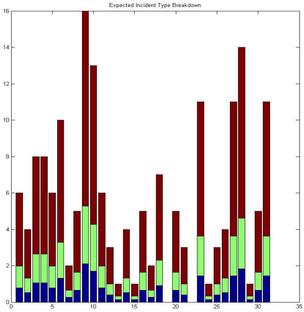
- 176 incidents involving multiple gangs
  - 52 are Rival Relations
  - 50 are Common Enemy Relations
  - 74 are Nonrelations



Intergang Analysis

Relation	Actual %	Expected %	Act.-Exp.%
RR	29.55%	13.12%	**+16.43%**
CER	28.41%	19.78%	**+8.63%**
NR	42.05%	67.10%	**-25.05%**





Remaining Questions

- Effect of Distance
  - Dist(RR) $\approx$ 2*Dist(CER) $\approx$ 2*Dist(NR)
  - Product of geography?
- Territory Trends
  - Does relation affect meeting place?
  - % of incidents in one gang's territory:
    - *RR* 76.92%
    - *CER* 60%
    - *NR* 45.95%
- Trend By Size
  - Do the relations of a gang depend on size of gang?
  - Hypothesis: Smaller gangs will have more *CER*s because they require more collaboration to compete with larger gangs

Acknowledgements

Yves van Gennip and Blake Hunter
Huiyi Hu, Matthew Valasik, Christina Garcia
George Tita, Kristina Lerman, Rumi Ghosh
Andrea Bertozzi
UCI Data Processing Group
Los Angeles Police Department
Gangs of Hollenbeck

Bonus Slide: Runtime of Data Cleaning