Chicago Crime Analysis

Jake Reyna, Qinyi Chen, Elizabeth Law, Jamie Atlas, and Xie He

University of California, Los Angeles

We create two models to analyze crime patterns of the city of Chicago in 2014. We create two network models to explore both geographic and qualitative community structure inherent in the data. The optimal communities turn out to match with the actual districts and socioeconomic divides of Chicago. We show that our methods are robust, and could be extended to analyze crime patterns in other locations.

1. Introduction

The city of Chicago is notorious for its high crime rates. Using data from the Chicago Police Department, we looked for (and found) patterns and community structure within crime events. We set out to answer the following questions:

- Do the Police Department's districts make sense?
- Where are the "dangerous" areas? Are they all dangerous for the same reasons?
- How does the layout of the city affect criminal behavior?
- How does weather affect crime?
- How does Chicago crime compare with that of other cities?

Our data set consists of information about crime in Chicago over 15 years. In order to make our task manageable, we focused on January and July, 2014. This gave us a much more workable matrix of size 34,000 x 22. An in-depth description of the data is in the Supplementary Materials. This gave us a lot of data, but it wasn't in network form. We looked through Andrea Bertozzi's research on crime in Los Angeles for inspiration [1] [2]. We realized that we couldn't answer all our guiding questions with a single model, so we decided to create two.

2. Model 1—Geographic Clustering

A. Motivation. We want to determine whether the city-drawn districts are an appropriate division of the city based on crime patterns. We therefore need a network that has a distance component to it, so that our communities will be connected subsets of the map. Inspired by the ideas in [3] of repeat victimization, we focus on crimes that have a high spatial and temporal locality. If an edge exists between two nodes, that gives us a high probability that a person will commit a crime at one location, and then the other.

B. The Model. We choose the node to be the blocks of the city of Chicago. For January we have 8,606 blocks with crimes, while for July we have 10,509 blocks with crimes. We define there to be an edge between two blocks if the following two assumptions are satisfied:

- There is a day on which both blocks were targeted.
- The two blocks are close to each other. We say nodes i and j are close if we have $max\{|x_i - x_j|, |y_i - y_j|\} \leq d$ *

This definition gives us a weighted, undirected network with no self-edges. Edge weight takes the place of multiple edges; that is, the weight of an edge corresponds to the number of days that two close blocks both had a crime occurrence. Notice that we do not incorporate crime type, instead focusing only on geographical information. Though simple, this model does offer insights into the crime patterns of Chicago.

C. Modularity Maximization. To partition our data based on our model, we performed modularity maximization on the network. That is, we seek to find a partition P of communities that maximizes

$$Q = \frac{1}{m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{m}]\delta(c_i, c_j)$$

where m is the sum of all the edge weights in the network, k_i is the sum of the weights of the edges of node i, and $\delta(c_i, c_j)$ is set to 1 if nodes i and j are in the same community, and 0 otherwise. In order to find the optimal partition, we used Louvain's method. [5] From this, we hope to find clusters in which criminals move from location to location, committing several crimes within a short period of time. Then, when an incident occurs at one spot, the police department can send units to the areas to which the perpetrators most likely moved.

* d is a threshold parameter with units of feet, which we vary throughout this section.

Significance Statement

With better drawn districts, the police department can send patrols to locations most likely to have crime events. Furthermore, our focus on same-day crime and transitivity means that even if a criminal moves on from a location after the police is called, we have a good idea of where they may go next. For well-studied cities like Chicago or New York our results are not as surprising (they fit in to our preexisting knowledge about their layouts), the fact that this model works so well for both indicates that we can use it to learn more about currently unknown locations.

.

Author Contributions:

Jake Reyna worked on the conclusion and further discussion sections. He also looked at published reports and presentations to give us an idea of how to format our own. **Qinyi Chen** worked on the Geographic Clustering section. She cleaned the data, chose the community detection algorithms, coded the model, and analyzed the results. She spent time working with different parameter values. **Elizabeth Law** worked on the social research section, finding socioeconomic information that matched our findings. She also created and set up the presentation slides. **Jamie Atlas** worked on the Manhattan and comparison between models section. He looked at papers on crime analysis to get ideas for the models. He also did the LaTeX for the report. **Xie He** worked on the Similarity Clustering section. She cleaned the data, chose the community detection algorithms, coded the model, and analyzed the results. She also found information on the corresponding narcotics areas.

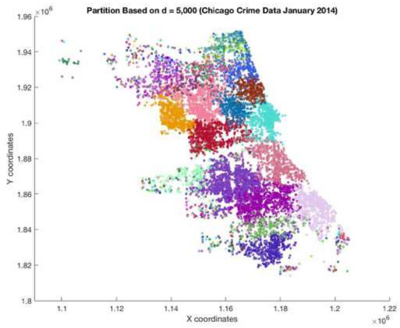

Fig. 1. January 5,000 partition

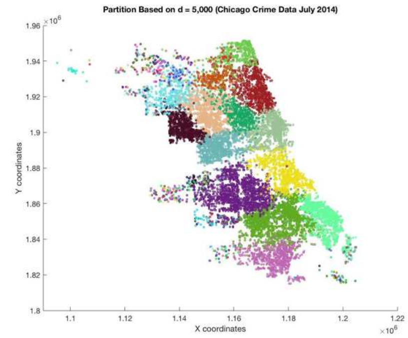

Fig. 2. July 5,000 partition

D. Influence of the Threshold Parameter. The choice of d has a noticeable effect on the structure of the maximal partition. On average, the dimensions of a Chicago district are roughly 20,000 x 20,000 square feet. We perform community detection based on four different values of d: 20000, 10000, 5000, 3000 (square feet). These are illustrated in Figures 16, 17, 1, and 18. We see from these that as d gets smaller, the number of communities tends to grow, and the mean community size goes down. The following table shows us the statistics of each choice for d.

d	# Communities	# 100+ Nodes Communities	Modularity
20,000	9	7	0.4608
10,000	40	5	0.6490
5,000	259	18	0.7695
3,000	1,031	17	0.8507

Looking at the 10,000 partition shown in Figure 17 qualitatively, a relationship can be seen between our communities and the actual structure of Chicago. Comparing it with Figure 19 (a map of the community areas of Chicago [4]), we see that the 10,000 partition roughly separates Chicago into the North (consisting of the North Side, Northwest Side, and Far North Side areas), Central Chicago, the West Side, the South (South Side, South West Side) and the Far South (Southwest Side, Southeast Side). When the threshold gets very small, as is shown in Figure 18, the geographical niceness breaks down, and the communities begin to overlap. Setting $d = 5{,}000$ yielded the best results. The boundaries between communities are still defined (for the most part), and it corresponds nicely with the predefined areas in Figure 19 .

E. January and July. Surprisingly, despite the vast weather differences of the two months, our communities turned out to be very similar. The following table shows the results of applying the Louvain method to data from July 2014.

d	# Communities	# 100+ Nodes Communities	Modularity
20,000	7	5	0.4649
10,000	24	7	0.6254
5,000	204	14	0.7852
3,000	1,382	22	0.8714

While the distribution of communities are slightly different from the January partitions, our general observations regarding how the threshold d affect our partition remains. The d = 5,000 partition, as shown in Figure 2 still gives a reasonable number of communities with fairly large sizes, and will

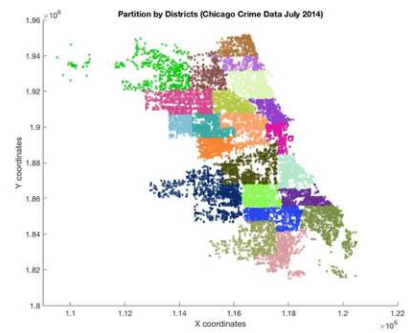

Fig. 3. July actual districts

thus serve as the most appropriate choice in the comparison to Chicago districts.

F. Comparison Against Chicago Districts. Looking at the July 5,000 partition side by side with the Chicago Districts map (Figure 3, we see a lot of similarities, particularly in the northern half of the map. The south matches up a little less well. Notice that the boundaries of the actual districts are much straighter than those of our modules—this is unsurprising, as a police department may wish to sacrifice some accuracy in order to make it easier to remember where one district ends and another begins.

G. Clustering Coefficient. While community detection is the focus of Model 1, we calculate the clustering coefficients for this model to gain an intuition about areas of interest. Here, we again use the $d = 5{,}000$ partitions. For the January model, the global clustering coefficient is 0.3089, while the mean local clustering coefficient is 0.5031. For the July model, the global clustering coefficient is 0.3025, while the mean local clustering coefficient is 0.4904.

Plotting the nodes with local clustering coefficient over 0.75, we obtain the graphs shown in Figures 4 and 20.

Note that in the plots, we omit the nodes with local clustering coefficient 1. Since in our definition, we require that two blocks can only gain edges between them if their l1 distance is less than 5,000, it is relatively easy for nodes on the corners of Chicago to form triangles with nodes around them. (This is also the reason why our mean local clustering coefficients were significantly higher than our global clustering coefficients: these perimeter nodes have low degree but very high clustering coefficients, and get counted the same as every

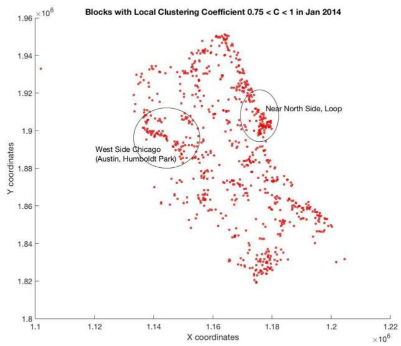

Fig. 4. January blocks with high clustering coefficients

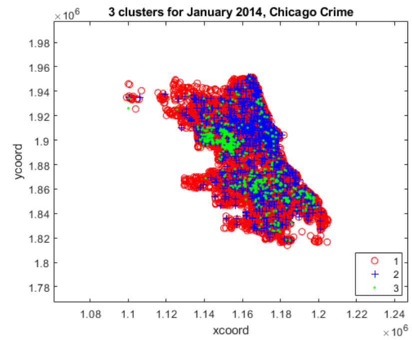

Fig. 5. January 3 clusters

other node, thus bringing up the mean.) Therefore we only look at nodes with local clustering coefficient in the open interval (0.75, 1) to avoid this bias. In the plot, we can clearly see areas with "clusters" of blocks with high local clustering coefficients. Two notable areas are the West Side of Chicago (consisting of Austin and Humboldt Part) and the Central Chicago (consisting of the Near North Side and The Loop), which are circled in Figures 4 and 20. They tend to include a large number of nodes with high local clustering coefficient, which indicates that these areas tend to be densely connected and many triangles are formed around those blocks.

For there to be a triangle, we need either crimes occurring on the same day for all three blocks, or each pair of blocks to have same-day crimes. These types of criminal activities might refer to thefts performed by one person (usually on the same day in nearby blocks) or to gang activities. The high occurrence of triangles around particular blocks can thus imply that these blocks form a neighborhood that shares many same-day criminal activities.

These two areas with high local clustering areas agree with the two areas that have the highest larceny rates in Chicago. As mentioned in the Introduction, theft is the main crime type occuring in Chicago in both in January and July 2014. By data provided by the Chicago Police Department (see Figure 21), the West Side (Austin) and the Central Chicago (Loop, Near North Side, Near West Side, Lake View) are the areas with the most counts of thefts in the past year. We will discuss later that the West Side is known for high criminal rates especially in larceny and narcotics, while Central Chicago is the home to Chicago business districts and the main area of tourist attractions, thus having a high larceny rate.

Model 2—Similarity Clustering

H. Motivation. While the Geographic Clustering model yielded strong results, we wonder whether physical distance is the only factor at play. In particular, we have many qualitative attributes in the data set that the first model didn't utilize. Therefore, we introduce a second model, which focuses on similarity between crime events [6] [11].

I. The Model. We use the same definition for nodes as we did in the first model—that is, each node is a city block on which a crime occurred. This time, however, we don't base our edges on physical distance, but instead create feature vectors for each node. Our feature vectors are all $1x22$ vectors, each element representing the number of occurrences of a different crime type.

J. K-Means Clustering. To perform community detection for this model, we use the method of **k-means clustering** Given a parameter k, and an initial set of means $\bar{m}_1^{(1)}, \bar{m}_2^{(1)}, \ldots, \bar{m}_k^{(1)}$ we perform an iterative process, to assign each node to a cluster $S_i^{(t)}$ at each step, so as to minimize

$$\sum_{i=1}^{k} \sum_{x \in S_i} ||\bar{x} - \bar{m}_i^{(t)}||$$

We find the optimal assignment using the algorithm outlined in [9].

K. Choice of k. The success of this algorithm is largely dependent on an appropriate choice for k, the number of clusters. Initially, we tried $k = 22$, as there are in total 22 types of crimes happened in January 2014 in Chicago. The result were not promising; this choice gave 14 empty groups, leaving us with 8. However, among these 8 groups, two are of specific interest because their main crime type is not theft, like the others, but narcotics, and these two are focused in certain areas of Chicago.

Lowering our k, we continue to see this behavior. In the 7 cluster plot, we have two groups, 6 and 7, gathering in the western and southern parts of the city (Figure 22). When we set k to 3, we can see a significant difference between the groups (Figure 5). As we see, all the green dots—group 3—are mostly located in western Chicago and the blue dots (group 2) are focused on the left part of the map, the center of Downtown Chicago. The red nodes don't have any visible focus. To understand the different patterns, we look individually into each group.

L. Comparison Between Groups. Figure 6 shows the overall frequency of crime types from Chicago in January 2014. Figures **??** and **??** show the frequency graph and heat map for Group 1. Group 1 looks very similar to the overall distribution, and we can see in Figure 22 that it is spread out throughout the entire city. Thus, we draw the conclusion that the Group 1 nodes serve as background in our model, from which we can draw conclusions by looking at the differences between it and Groups 2 and 3.

Looking at Group 2, we can see key differences. Figures 23 and 7 shows the frequencies and heat map. Notice that while the main type of Group 2 is theft, like before, the geographical

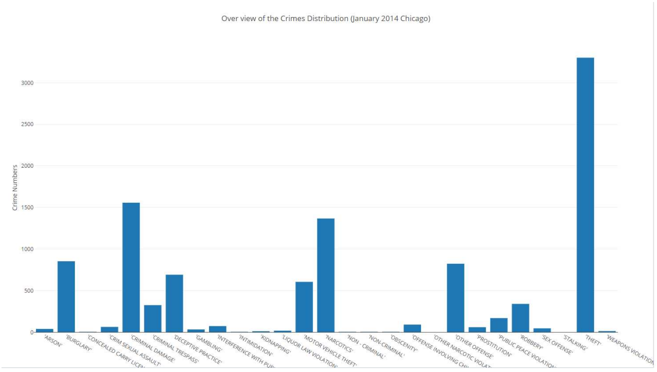

Fig. 6. Crime frequencies of all three groups

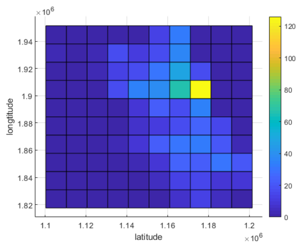

Fig. 7. Heat map of crimes in Group 2

distribution is heavily centered on Central Chicago. Also, the ratio of the number of theft to other crimes is much higher than it was in Group 1.

The aforementioned "interesting" group that have a high amount of narcotics events manifested as Group 3. Figures 8 9 shows its distribution and heat map. This time, the distribution looks very different, as narcotics is the most significant crime type. Furthermore, Group 3 crimes primarily take place in Western Chicago, though we do see some emergence in the south.

Similar to our first model, when we ran the model on the July data, we found the same patterns of crime. There is a lower total number of events in January (likely less people are out and mobile), but the relative amounts were the same throughout the year.

M. Causes of our Findings. In Figure 24 we see a 'tourist attraction map' for Chicago which coincides with our cluster of

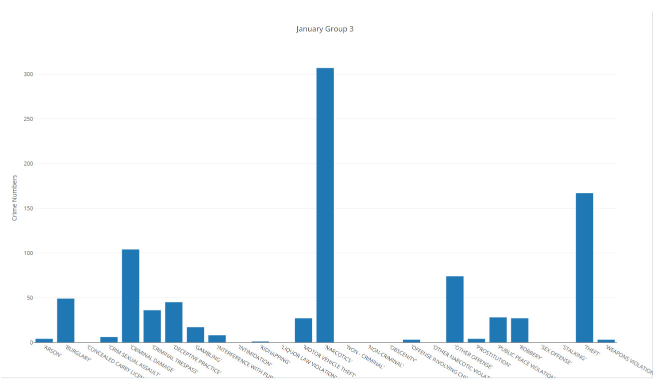

Fig. 8. Crime frequencies of Group 3

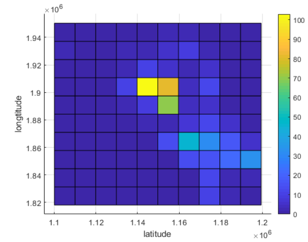

Fig. 9. Heat map of crimes in Group 3

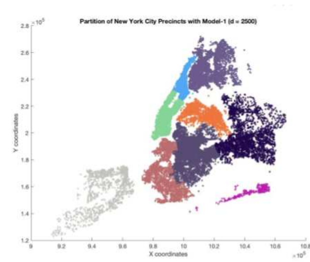

Fig. 10. Geographic clustering for Manhattan

Group 2. In Figure 25 we see a 'drug usage heat map' for Chicago from 2001 to 2008 which coincides with our cluster of Group 3.

These result makes sense; theft tends to occur where there are known to be tourists, and the western Chicago area has historically had problems with drugs. Our group had no knowledge of these danger areas in advance, so we were pleased to find that our methods work well to detect crime communities even when unsupervised. Thus, we are confident in their robustness. To test this further, we tried the same system of analysis on a new data set.

3. Manhattan

Using information from the NYPD that had a similar structure to the Chicago dataset, we created analagous models for New York City.

A. Model 1. We can see in Figure 10 a strong delineation of the regions of the city. Though Manhattan, Brooklyn and Queens are each broken into two communities, we see the boundaries of all five boroughs are clearly defined. Focusing on Manhattan, we see the two modules are split across the Upper East Side, Grand Central Park, and Morningside Heights. It makes sense that the split would occur by the Park, as our data has no crimes taking place in it, and so the nodes around it would have a lower degree. Furthermore, the rightmost split is between the Upper East Side and East Harlem, which are known to be relatively safe and dangerous parts of the city, respectively.

B. Model 2. Once again, the best results come from setting k to 3. Overlaying our data with a map of the city, we are able to see which areas suffer from which types of crime (Figure 11). And like Chicago, our findings are in accordance with what we know about the city. We see theft, in red, in the Upper East and West sides, which are extremely affluent areas. Thus, it would be a likely target for burglary. We also see in yellow are very serious crimes, such as sexual assault or child

Fig. 11. Similarity clustering for Manhattan

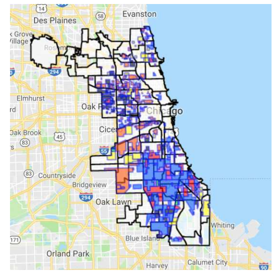

Fig. 12. Gang territories [14]

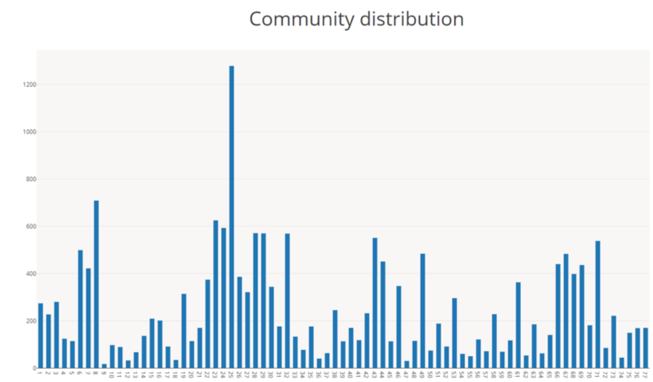

Fig. 13. Crime breakdown by community

abandonment. These are mostly concentrated in the north of Manhattan, in Harlem, which is known to house a lot of poorer communities.

C. Comparison with Chicago. To the credit of our models, we found very similar results for both cities. Though our Chicago dataset was more extensive than the New York one, the Geographic Clustering Model allowed us to recreate much of the drawn districts of both cities, and the Similarity Clustering Model identified areas by social issues.

4. Differences Between Models

Though both models were designed to help identify crime patterns in Chicago, their structure—and therefore the results of each—are very different. The Geographic Model consists of connected, mostly convex components, with clear boundaries. The communities of the Similarity Model, by contrast, each have nodes spread throughout the entire city. Fortunately, our results were clear enough so that the concentrated areas were visible by inspection. This may not be true for every data set, however, so it is less useful than the first model for drawing districts. It has a strength, however, in that it can be used to identify alike areas that are separated by a large distance. For example, we found in [Fig] that drug usage is concentrated in the Southern as well as the Western regions of Chicago. Model 1 can find us communities in both of those areas, but not that they are their own communities for the same reason.

5. Social Research

Gang violence has been ingrained in Chicago's history, from prominent figures such as Al Capone or the Gangster Disciples. We can see current territories of different gangs in Chicago in Figure 12. It is interesting to look at where gangs are located because it is important to see the effect of gang territories on the landscape of Chicago, and specifically, our research of crime in Chicago. We see that these territories are predominantly in the west and southern regions of Chicago, with most of these territories almost perfectly overlapping with the different communities within Chicago.

Most of the violence in Chicago is spurred by gang violence, from arms race between gangs, or turf wars over regions to sell drugs and firearms. As we can see from Figure 6, the top crimes are: theft, criminal damage, and narcotics (organized selling of drugs, primarily via criminal organizations) We are interested in whether the patterns we found from our heat maps overlap with gang territories from the previous graph. Even for crimes not directly linked to specific gangs, we can still assume that some crime may have been influenced by gangs, ie. higher crime rates in a certain area may be influenced by lack of police protection in the area because of a gang's stronghold in the neighborhood.

We can also track some of the violence that occurs in Chicago from hate crimes, and racial tensions. Racially segregated neighborhoods have strong social tensions between socioeconomic groups, stemming from history. From laws in place for segregated neighborhoods, to real estate industries prohibiting colored and other minority racial groups from living in "white" neighborhoods, even in terms of Modern Segregation [15], Chicago is a city plagued with racial tensions and the highest crime rate in the nation. The 20k partitioned graph from Model 1 is strongly reflects the general racial segregation of Chicago between the north side and the south side that's carried over from the 1900's. The 10k partitioned graph from Model 1 reflects the general communities of Chicago and is a representation of how crime occurs among different social groups.

We first look into the different communities within Chicago in order to compare the differences between them (Figure 13, Figure 14). For example, we see that community 25, located on the west side of Chicago, has the highest occurrence of crime within the city.

We first look into two communities: Austin (Figure 26), and Lake View (Figure 27). We note that Austin, community 25, is the community that had the highest occurrence of crime

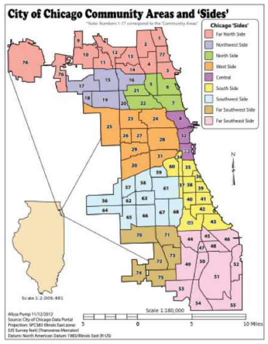

Fig. 14. Placement of communities

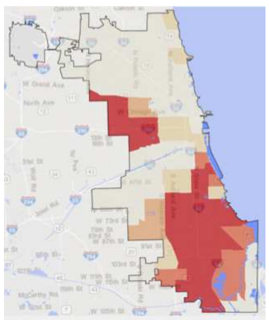

Fig. 15. Black population [16]

within Chicago, and Lake View is a community with a very low occurrence of crime. Lake View is located in the north of Chicago, with a very large White population. Austin is located in the west, with a large Black population (Figure 14). Lake View has a higher household income than Austin as well.

However, we want to expand our view to all of the communities within Chicago before we can conclude on a pattern.

In Figure 29, Figure 30, and Figure 15, we see a heatmap of the locations of where different racial groups reside in within the different communities of Chicago. We see that the white population is primarily located in the north, the latino population in the west, and black population in the west and south. Comparing these graphs, we notice that regions where minorities (black and latino) reside in are also the primary regions of gang territories (Figure 12).

We also notice from our heat maps generated from model 2, that the location of a high percentage of crime (primarily narcotics) are located in the western and southern regions of chicago, which overlaps with the areas of a high black population.

We see in Figure 28 that the communities located in the northern region of Chicago have higher median income as compared to western and southern regions of Chicago. The crime rate in these regions are significantly lower as well.

We notice that the areas with higher income are also cities that are part of gated communities: ie. Lakeview, Lincoln Park, Ravenswood, Uptown Chicago, and DePaul, Rogers Park, Edgewater, Logan Square. We can link the higher income in these areas to money that is able to afford protection/ etc. Because in these gated neighborhoods, they can afford security systems, extra security personnel, etc.

It is also interesting to look into census data of food stamps and post-secondary degree, as food stamps are linked with lower-income households, and post-secondary education is linked with higher-income households. We notice that food-stamp recipients are located in the West and the South of Chicago (Figure 31), while those with post-secondary degrees are located in the North and the East of Chicago (Figure 32).

6. Further Discussion

While we did gain deeper insight into the geography and crime distribution of Chicago through modeling and researching the social background of the city, there are still smaller aspects of the models that could give more insight into crime in Chicago if studied further. While crime was found throughout the city, there were gaps in certain areas where crime was sparse. While we do believe that much of that can be attributed to the bodies of water located throughout the city, there may also be other causes for these gaps. They could possibly be due to other factors such as gated communities or other types of areas where crime is prevented significantly. Our data also did not find any significant differences between the months of January and July other than there being more crime in July than January, so a deeper analysis into the nature of the crimes between these months would help in understanding how crime changes throughout the year. While our model did prove to align with the communities of Chicago, it would be helpful to use this model to compare Chicago with other years to see how crime has transformed year to year.

7. Conclusion

Comparing our models with our social research, the models did successfully represent the communities and social groupings of Chicago. While there was no clear evidence that weather had any impact on crime other than a higher rate of crime in July than in January, it is clear that the social environment is a key factor for the concentration and type of crime that occurred. From the partitioning of crime districts by the Chicago police department, the natural racial segregation carried on from the early 1900's into modern day, and from the disparity of wealth among the citizens of the city, our models reflected the impact Chicago has had in splitting its residents into social groups. Our model was also able to successfully give an overview of crime in Manhattan. Like Manhattan, Chicago demonstrates how social aspects contribute to different types of crime within the city.

ACKNOWLEDGMENTS. Thank you for the dataset, Professor Porter! And thank you to both you and Yacoub for the help and direction you've given us for this project.

8. Bibliography

1. Short et al. A Statistical Model of Criminal Behavior. *Mathematical Models and Methods in Applied Science* (2007) 18, 1249-1267. http://paleo.sscnet.ucla.edu/ShortEtAl2008-M3AS.pdf
2. Kuang et al. Crime topic modeling *Crime Sci* (2017) 6:12 https://doi.org/10.1186/s40163-017-0074-0
3. M. B. Short, A. L. Bertozzi, and P. J. Brantingham. Nonlinear Patterns in Urban Crime: Hotspots, Bifurcations, and Suppression. SIAM Journal on Applied Dynamical Systems, 2010, Vol. 9, No. 2 : pp. 462-483 https://doi.org/10.1137/090759069
4. Peter Fitzgerald. Chicago community areas map. *Wikipedia.* April 2008, https://commons.wikimedia.org/wiki/File:Chicagocommunity-areas-map.svg.
5. Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (9 October 2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment.* 2008 (10): P10008. arXiv:0803.0476. doi:10.1088/1742-5468/2008/10/P10008.

6. A.L. Bertozzi and Arjuna Flenner. Diffuse Interface Models on Graphs for Classification of High Dimensional Data. *Multiscale Model. Simul.*, 10(3), 1090–1118. (29 pages) https://doi.org/10.1137/11083109X
7. Ulrike von Luxburg. A tutorial on spectral clustering, 2006. Technical Report No. TR-149, Max Planck Institute for Biological Cybernetics, 2017. https://dblp.org/rec/bib/journals/corr/abs-0711-0189
8. Kriegel, HP., Schubert, E. Zimek, A. *Knowl Inf Syst* (2017) 52: 341. https://doi.org/10.1007/s10115-016-1004-2
9. Hartigan, J. A.; Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society.* Series C (Applied Statistics). 28 (1): 100–108. doi:10.2307/2346830. JSTOR 2346830.
10. Newman, M.E.J. 2010. Networks: An Introduction. Oxford, UK: Oxford University Press.
11. Yongli Li, Peng Luo, Chong Wu. A new network node similarity measure method and its applications. *arxiv*. 2014. arXiv:1403.4303
12. Chicago printable tourist map. *Sygic*. 2012. https://travel.sygic.com/lp/Chicago-Printable-Tourist-Map
13. MantasCode. Chicago narcotics solicitations in public-ways; incident maps. *Chicago Data*. 2018. http://mantascode.com/chicago-narcotics-solicitations-in-public-ways-incident-maps.
14. Elliot Ramos. In Chicago, gangs abound, but where are they? *WBEZ News*. September 2012. https://www.wbez.org/shows/wbez-news/in-chicago-gangs-abound-but-where-are-they/0d956c2e-171a-483a-8666-f5228d98812d
15. Whet Moser. Chicago Isn't Just Segregated, It Basically Invented Modern Segregation *Chicago Magazine*. March 2017. http://www.chicagomag.com/city-life/March-2017/Why-Is-Chicago-So-Segregated/
16. United States Bureau of the Census. Household Income in Austin, Chicago, Illinois (Neighborhood). *Statistical Atlas of the United States*. Washington :Govt. Print. Off., 1914. Print. https://statisticalatlas.com/neighborhood/Illinois/Chicago/Austin/Household-Income

9. Supplementary Materials

A. Data Set. The 22 fields for each crime were:

- ID — each crime is given a unique ID
- CaseNumber — similar to ID
- FBICode — similar to ID and CaseNumber
- Date — in the format of "09/24/2014 07:42:00 PM"
- Year — year of event
- Block — corresponds to Chicago city blocks
- IUCR — stands for "Illinois Uniform Crime Reporting", are four-digit codes used to classify crimes
- PrimaryType — used to classify crimes
- Description — short description of the event
- LocationDescription — e.g. "street", "home", etc
- Arrest — boolean value
- Domestic — boolean value
- District — Chicago Police Department's division of the city
- Beat — subdivison of districts, refers to the area in which an officer patrols
- Ward — geographical distance used for voting and polling
- CommunityArea — geographical division of the city
- XCoordinate — location where a crime takes place, unit of distance is feet
- YCoordinate — see "XCoordinate"
- Latitude — location of event
- Longitude — longitude of event
- Location — combination of latitude and longitude
- UpdatedOn — year that crime was added to the database

Some of these were redundant, such as "Year" and "Latitude/Longitude". Others were simply not useful, like "ID"—we can't do meaningful numerical analysis when every ID is distinct. We ended up considering only 7 of the fields: "Date (only using the day, not the time)", "Block", "PrimaryType", District, "Location Description", "XCoordiante", and "YCoordinate".

B. Extra Figures.

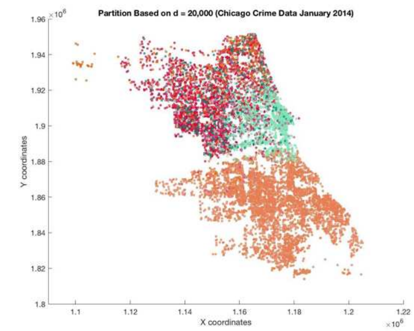

Fig. 16. January 20,000 partition

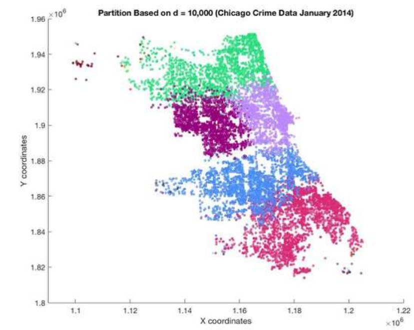

Fig. 17. January 10,000 partition

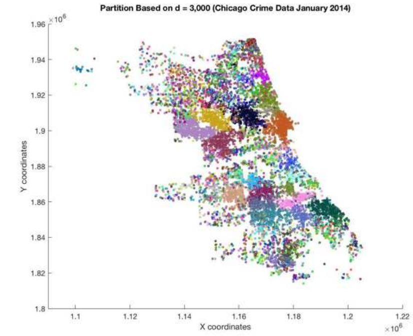

Fig. 18. January 3,000 partition

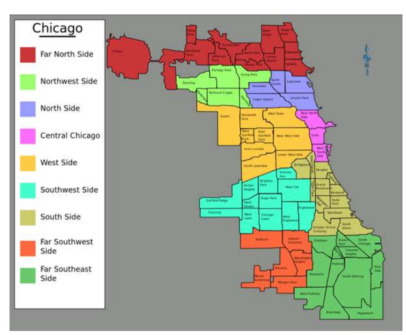

Fig. 19. Chicago communities

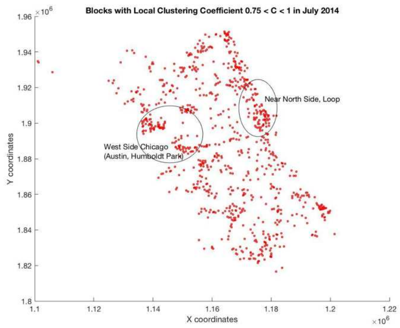

Fig. 20. July blocks with high clustering coefficients

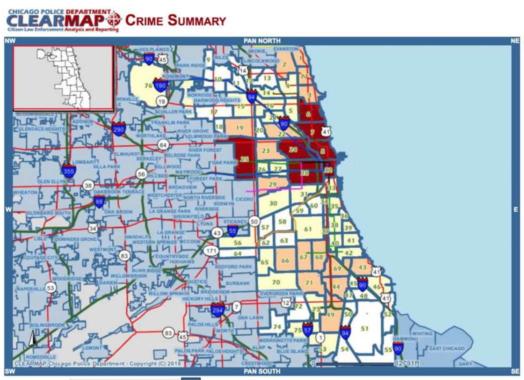

Fig. 21. Map of theft-related crimes in Chicago

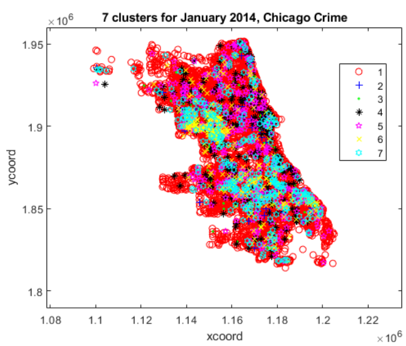

Fig. 22. January 7 clusters

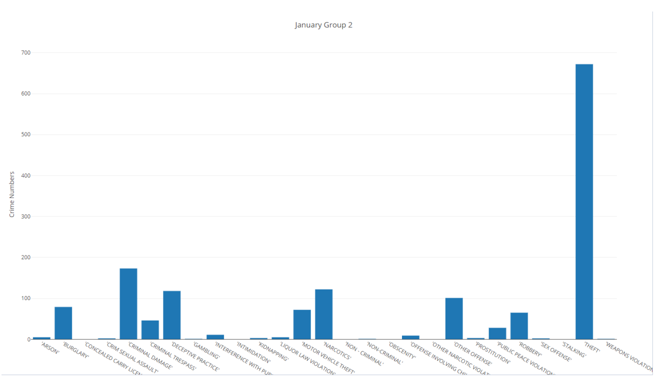

Fig. 23. Crime frequencies of Group 2

Fig. 24. Tourist attractions [12]

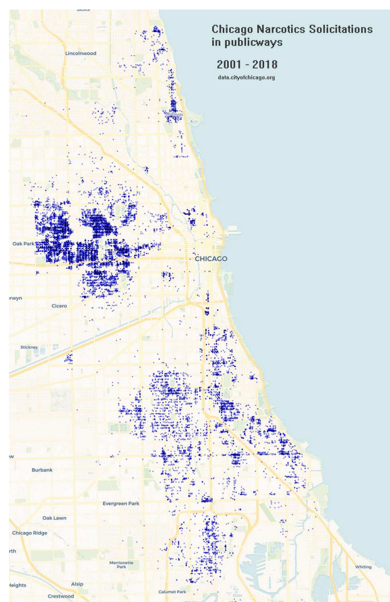

Fig. 25. Drug usage map [13]

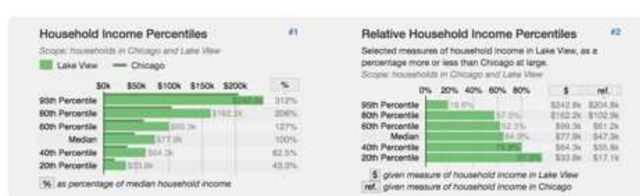

Fig. 26. Austin income [16]

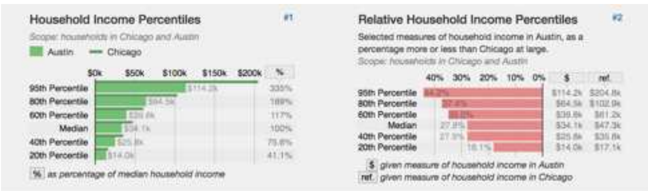

Fig. 27. Lake View income [16]

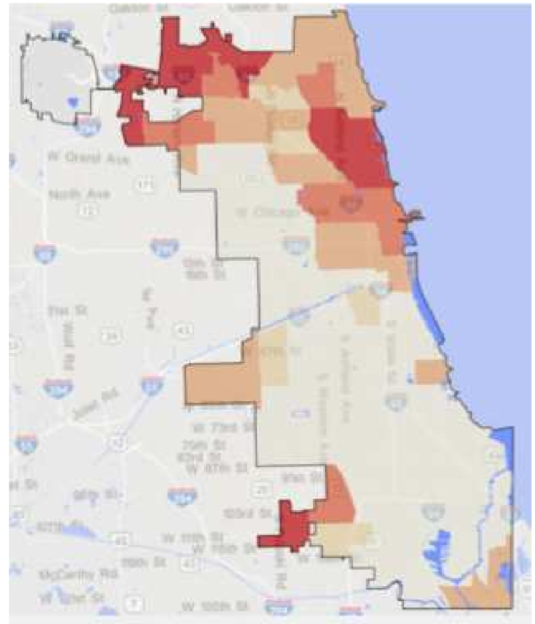

Fig. 29. White population [16]

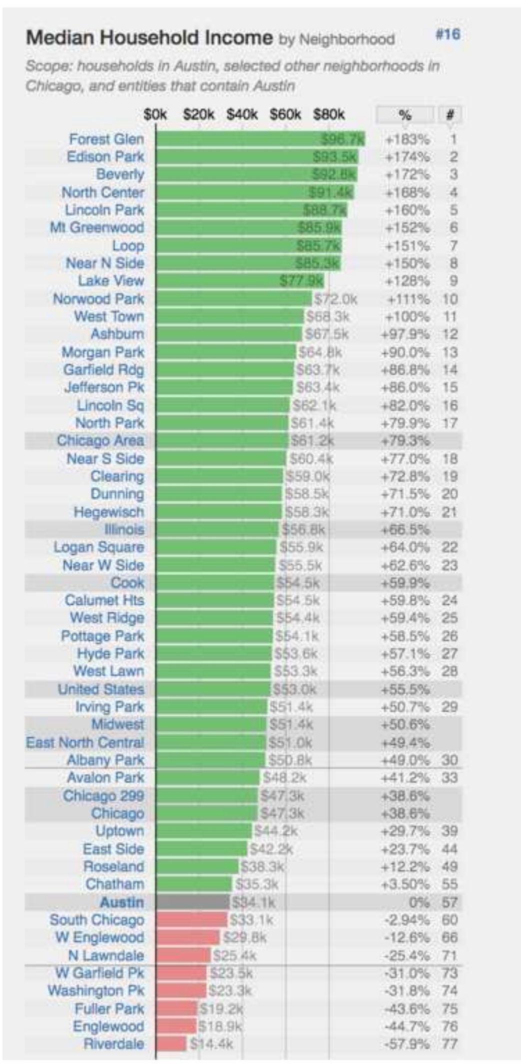

Fig. 28. Household incomes of all communities [16]

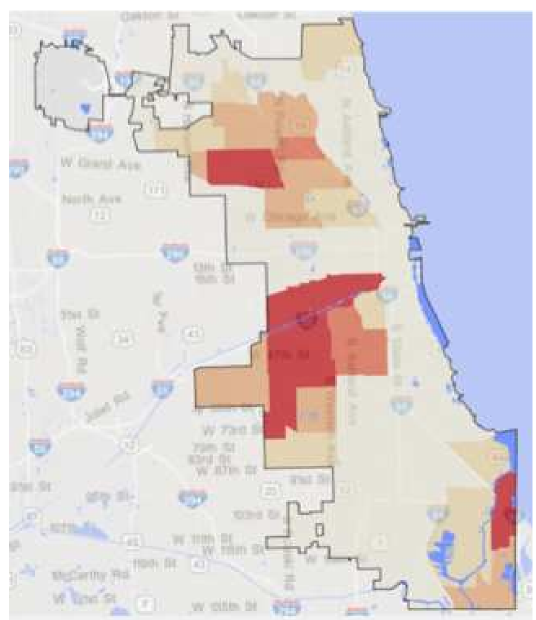

Fig. 30. Latino Population [16]

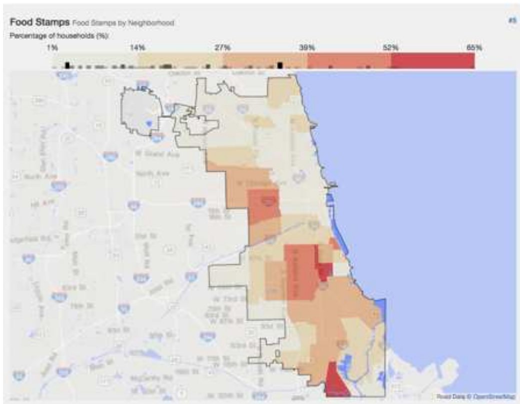

Fig. 31. Food stamp recipients [16]

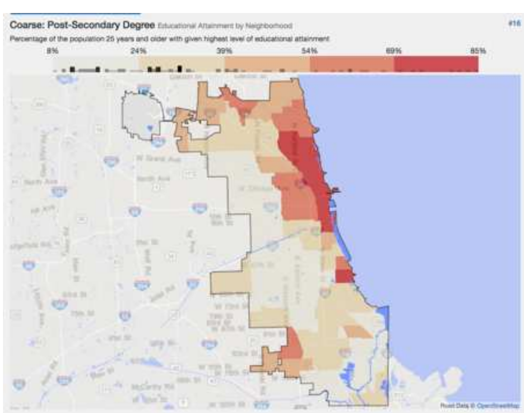

Fig. 32. Post-secondary degree holders [16]