Analysis of Orlando Theft Crimes

Network analysis of the crimes in and around the city of Orlando, Florida by means of clustering.

Brenden Morton

University of Central Florida

Chart, scatter chart

Description automatically generated

Network Optimization

Dr. Wei Zhang

December 8, 2020

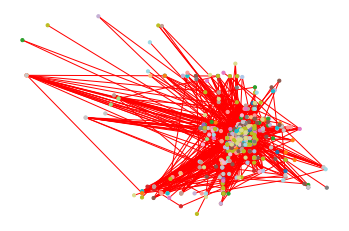
Throughout the history of human gathering, individuals or groups of individuals have committed theft crimes. In general, theft can be described as the reallocation of one’s goods in a non-consensual manner with the intent to deprive the rightful owner of personal property. Although the motivations for this genre of crime can develop from both economic and non-economic circumstances, there is always a victim. According to American sociologist Richard Quinney, there is a distinct relationship between a society and the crimes that occurs within it. Reflecting on this claim, I thought it would be interesting to see the relationships between a subset of a largely evident type of crime on a geographical scale centered in and around our community. In this paper, the crimes in the city of Orlando, Florida between the years of 2014 and 2020 are thoroughly examined through the use of network and data clustering algorithms to provide an insight into where exactly theft is most prominent and to determine whether or not there are relationships between theft crimes based on location. In the case of this report, all crimes that involve the loss of property that are not explicitly defined as theft, such as motor vehicle theft and burglary, will be generalized as theft. It is necessary to include these crimes since they involve similar, if not the same, motivations as regular theft.

To collect the geographical relationship between theft crimes around the city of Orlando, Florida, network clustering algorithms, as well as some data clustering algorithms, are employed. Markov Chain Clustering and Spectral clustering will be used for aggregating the dataset of crimes based on the locational positioning of the crime in terms of a street intersection. Additionally, the Mean-Shift data clustering algorithm is used to gain a perspective on the distribution of theft crimes across Orlando. Additional software such as Cytoscape will be used to visualize the graph constructed by the unique streets of the dataset as well as to illustrate the clusters formed by the clustering algorithms. Crude representations of the data are developed by the Networkx Python 3 package are also used to exemplify some dense regions in the network in a different perspective.

The city of Orlando provides a multitude of varying datasets on their website data.cityoforlando.net. Here, a few different datasets for criminal activity were available for public use, however, the most extensive dataset is the one used in this paper. In this dataset resides roughly 230,000 crimes that occurred in the city bounds of Orlando starting in the year 2014 and ending in September of 2020. According to the description of the dataset, all data conforms to FBI standards in terms of the definition of the crime such that each crime is rightfully defined. In the case of this report, the crimes of theft, motor vehicle theft, and burglary are verified to be accurate descriptions of the crime carried out. In general, the data from this dataset can be exported to a csv file which delimits 10 different attributes for each of the 230,000 crimes. The following is the list of attributes that are shown for each crime: Case Number, Case Date Time, Case Location, Case Offense Location Type, Case Offense Category, Case Offense Type, Case Offense Charge Type, Case Disposition, Status and Location. The attributes of interest in the analysis of the data included only the following: Case Location, Case Offense Type and Location. Case Location refers to the street intersection of the crime. The dataset consisted of two forms for Case Location where only one of which was adequate for analysis. The Case Location could be given in the form of [street 1 / street 2] which can easily be tokenized and analyzed with an adjacency matrix or of the form [Block of X on Street] which can be tokenized but not in an advantageous way. The latter form does not provide enough information to be used for network analysis. In the case of the former form of [street 1 / street 2], these streets can be tokenized and provides some intersection in Orlando that works well with an adjacency matrix. Whereas the form [Block of X on Street] represents a residential crime and this data is not able to be easily translated to a position in an adjacency matrix. To remedy the situation, I filtered out this form such that only the crimes with valid [street 1, street 2] form existed in the dataset. The next attribute of interest is the Case Offense Type. This attribute can be used to filter out any crimes that do not belong to the subset of theft which includes theft, motor vehicle theft and burglary. The last attribute of interest is the Location of the crime. This differs from the Case Location in that it represents the longitude and latitude coordinates of the crime in the form [long. / lati.].

<Brief introduction of the implementation and details of your results>

To properly analyze the crime data, I began by filtering out all the crimes that did not belong to the subset of theft crimes. In the previous paragraph, the topic of improper formatted data is discussed in detail. This data was also filtered out. Next, an adjacency matrix was created where the rows and columns of the matrix represented all the unique streets of the crimes. In this way, for a crime that occurred on [street A / street B], the position that corresponds to this cell in the adjacency matrix was incremented by 1 symmetrically([street A][street B] and [street B][street A]). After filling the entire matrix with all the crimes, the matrix was used as an input to a Markov Chain Clustering algorithm to develop clusters of the data. These clusters can be found in ***clusters.csv***. Using these clusters, I was able to output the clusters using the Networkx Python package. Here is the image below:

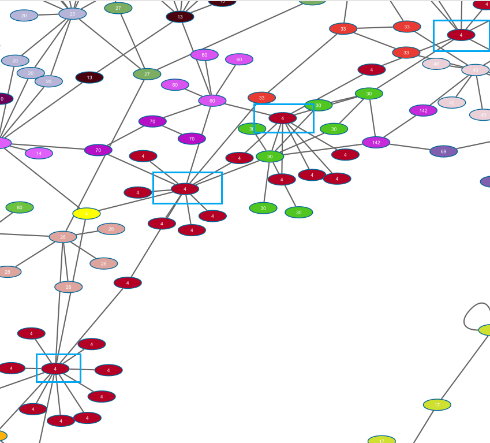
The resultant representation of the clustering is quite complicated to observe since the data is packed tightly at the center. To gain a better image of the clustering and to see the relationships between the crimes clearly, it was determined to use Cytoscape. To do this, the adjacency matrix had to be exported as a .gml file. After importing the matrix into Cytoscape, the network was created. First, a Spectral Clustering algorithm was applied to the network and a subset of the network can be seen below:

Chart, scatter chart

Description automatically generatedA clearer version of this image can be seen in the FullSpectral.png image.

Looking at the graph, this is a single connected component where each node represents a street intersection of at least one crime. Some of the key features of this network include some of the larger clusters such as cluster 1 which has the most nodes in the cluster and node 4 which can be considered an intermediary network (bridge network) that links other subnetworks together. In the case of cluster 1, this shows a high crime rate among this street and its neighboring streets. A closer look on cluster one gives us this:

Diagram

Description automatically generatedThe nodes encased in red rectangles not only have a high number of neighbors but also act as parent nodes each of which have several neighbors that are leaf nodes. This relationship in the crime network can imply that these are major streets that are used in areas that frequent crime. One can infer by looking at this relationship that these nodes represent unsafe streets to be on. Cluster 4 presents another interesting relationship in the network.Some of the nodes of cluster 4 are inscribed in blue rectangles solely for emphasis. Looking at some of the neighboring subgraphs, cluster 4 acts as sort of a bridge between other clusters. This could imply that these streets can be used for movement to other high crime streets.

Next, the Markov Clustering algorithm was applied to the original network and this yielded some well-defined clusters. The clusters were segmented such that each cluster was its own connected component. In this implementation, an inflation value of 2.5 was applied to the clustering and an expansion value of 2 was also used. The clusters seen below are the top 4 clusters in terms of nodes in a cluster meaning that these are the top 4 worst streets in terms of theft crime. From left to right are clusters 1, 3, 2 and 4. Again, there all these clusters have the many parent to many leaf node relationships that shows that the parent nodes are central to some of the crime in each cluster. In terms of a physical community, the parent nodes are streets that transfer crime to other streets. The parent nodes are used as steppingstones to other streets that have theft crimes.

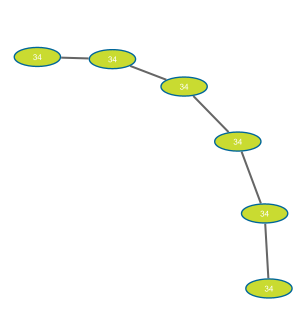
A picture containing diagram

Description automatically generated

I had a series of unsuccessful attempts to import the street data for each of the nodes. To see the streets of the nodes, there is a csv file called node\_labels.txt. At the end of the report, I will include the names of the most prominent intersection of the streets. From the clusters above, cluster 1 is the most prominent in terms of number of leaf nodes. Like cluster 1, Spectral clustering shows a high rate of theft occurring on these streets. Another observation is that the average distance between leaf nodes is only 5 edges. In terms of physical streets, there is not much of a long trip to get to these other streets. One can postulate that crime sprees can happen in this network where thieves can easily travel through the main streets (parent nodes) to some of the leaf nodes for theft.

The clusters below represent some interesting relationships between theft crimes that both compare with some of the other clusters already covered.

Diagram

Description automatically generated

The cluster on the left is a very well-connected network with multiple cycles in it whereas the cluster on the right has a non-trivial acyclic tree structure. One observation of the cluster on the left is that there are substantially less leaf nodes. The theft crimes in this street network are closely related. However, the cluster on the right, has linear with only one edge connected adjacent nodes. This cluster is of the form of a linked list. This cluster has much less connectivity because it is full of cut-vertices. A crime network like this could be easier to breakdown since there are not currently any profound existing connections between streets due to the linearity of the cluster.

To conclude, using these clustering algorithms has highlighted some interesting and unique relationships between the theft crimes in the dataset. Overall, the highest level of crime occurred in cluster 1. Three parent nodes in the center of this cluster provided access to other streets where theft crime is evident. A relationship like this insists that there is high cohesion with crime in this community centering at the parent nodes. Another interesting relationship is of cluster 4 in the spectral analysis. This cluster acts as a bridge network between multiple other clusters. This is an intermediary between crimes or targets. Clusters such as cluster 4 that connect a range of different clusters poses a high threat in the mobilization and transport of theft crime around these communities. I would consider networks such as cluster 1 with multiple parent nodes with high order of leaf nodes and cluster 4 which acts as an intermediary between other clusters to be some of the more dangerous streets in terms of theft. The data for the names of the streets can be found in node\_labels.txt.

As a side analysis, I implemented a Mean-shift data clustering algorithm on the Longitude and Latitude coordinate data for all crimes generalized as theft.

Chart, bubble chart

Description automatically generated

The axes on this chart are Longitude and Latitudes of all the crimes. The mean shift algorithm clustered the data into 11 total cluster with 5 main clusters of theft data. The circles represent the center of the data clusters. Observing the results of the data, the 5 main clusters, represent the highest concentration of theft crime in the city of Orlando. Based on the geographic coordinates, the main clusters are position on the center of Orlando which is about where the CBD or central business district that downtown Orlando is. The concentration of crime in the main clusters emphasizes networks of theft crimes as seen in the previous analysis of Markov and Spectral clustering.

All in all, this data clustering shows some of the more prominent locations of theft crime.

To Brenden,

The next step that you can do is begin to create the readme and provide documentation for the source code.

Citations

1. MetroWest Public Safety. (created 2016, last updated 2020) City of Orlando Crimes [Crimes in Orlando, Florida]. Retrieved 23:44, December 1, 2020, from <https://data.cityoforlando.net/Orlando-Police/City-Of-Orlando-Crimes/hm2t-fd4m>
2. Wikipedia contributors. (2020, November 28). Crime. In *Wikipedia, The Free Encyclopedia*. Retrieved 04:43, December 1, 2020, from <https://en.wikipedia.org/w/index.php?title=Crime&oldid=991063902>