Computer and Information Science; Vol. 17, No. 1; 2024
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Principal Components and Assortativity-based Assessment of the Similarity of Crime Metrics across Coterminous Wards in the City of Chicago

Natarajan Meghanathan[1]

[1] Department of Computer Science, Jackson State University, USA

Correspondence: Natarajan Meghanathan, Department of Computer Science, Mailbox 18839, Jackson State University, Jackson, MS 39217, USA. Tel: 1-601-979-3661. E-mail: natarajan.meghanathan@jsums.edu

Received: March 11, 2024 Accepted: April 17, 2024 Online Published: April 29, 2024

doi:10.5539/cis.v17n1p36 URL: https://doi.org/10.5539/cis.v17n1p36

Abstract

The City of Chicago (with 50 wards) has one of the highest crime rates in the US. We seek to quantitatively assess the similarity of crime metrics across coterminous wards using a combination of Principal Component Analysis (PCA) and Assortativity analysis. We first build a ward network (nodes are the wards and edges connect coterminous wards) of the city using the ward map. We parse through the 2022 crime dataset for the city and build a matrix whose entries correspond to the number of occurrences of a crime type in a ward. We conduct PCA of this ward-crime type matrix and determine a weighted average PC_crime_score (using the entries in the high-variance principal components and their variances as weights) for each ward. We observe the coterminous wards to exhibit moderate-strong assortativity with respect to three different crime metrics: hotspot classification, PC_crime_scores and the crime counts of the individual crime types.

Keywords: Chicago, Ward Network, Assortativity, Crime Hotspot, Crime Count, Principal Components

1. Introduction

Assortativity analysis (Noldus & Van Mieghem, 2015) is a well-known analysis procedure in Network Science to quantitatively assess the extent of similarity between the end vertices of the edges with respect to a node-level metric. If the node-level metric values are real numbers, the assortativity index (AssI) of a network is quantitatively measured as the Pearson's correlation coefficient (Strang, 2006) between the node-level metric values for the end vertices of the edges. The AssI value could range from -1 to +1: with -1 indicating the network is dissortative (i.e., for two edges (u_1, v_1) and (u_2, v_2), if the node-level metric value of u_1 is greater than that of u_2, the node-level metric value for v_1 is likely to be lower than that of v_2, and vice-versa) and +1 indicating the network is assortative (i.e., for two edges (u_1, v_1) and (u_2, v_2), if the node-level metric value of u_1 is greater than that of u_2, the node-level metric value for v_1 is likely to be greater than that of v_2 as well, and vice-versa). If the AssI value for a network is close to 0, it indicates the node-level metric values for the end vertices of the edges are independent of each other (i.e., for two edges (u_1, v_1) and (u_2, v_2), if the node-level metric value of u_1 is greater or lower than that of u_2, the node-level metric value for v_1 might be comparable to that of v_2, and vice-versa). Nevertheless, positive values of AssI is considered a measure of assortativity (Meghanathan, 2016) at the following levels weak ($0 \leq$ AssI < 0.40), moderate ($0.40 \leq$ AssI < 0.60), strong ($0.60 \leq$ AssI < 0.80) and very strong (AssI ≥ 0.80). Similar levels could be considered for dissortative networks (whose AssI values are less than 0). Throughout the paper, the terms 'node' and 'vertex', 'link' and 'edge', 'network' and 'graph' are used interchangeably. They mean the same.

The node-level metric typically used for analysis is degree centrality (Newman, 2010), the number of neighbors of a node. Nevertheless, recent studies (Meghanathan, 2016) have reported assortativity analysis of real-world networks with respect to other centrality metrics such as eigenvector centrality (Bonacich, 1987), betweenness centrality (Freeman, 1977) and closeness centrality (Freeman, 1979). Assortativity analysis of complex real-world networks has been rarely conducted and reported on node-level metrics other than centrality metrics, though it could be conducted with respect to any node-level metric that is characteristic of the network of study. Likewise, to the best of our knowledge, assortativity analysis has not been reported if the node-level metric values for the end vertices of the edges are binary. In this paper, we seek to conduct assortativity analysis on crime data for a city and intend to quantitatively assess whether coterminous wards (wards that share common border) of the city are similar or dissimilar to each other with respect both continuous and binary crime metrics.

We propose to use the Kendall's concordance (Strang, 2006)-based correlation coefficient (as the AssI value) to assess the assortativity of networks wherein the end vertices of the edges are represented in terms of binary values (0-1). The Kendall's concordance-based correlation coefficient is computed as the fraction (ratio) whose numerator is the difference in the number of concordant pairs and the number of discordant pairs and denominator is the sum of the number of concordant and discordant pairs. The range of values for the Kendall's concordance-based correlation coefficient is also -1 to 1. Hence, the above levels of assortativity and dissortativity (used to interpret the Pearson's correlation coefficient-based AssI values) could also be used for interpreting the Kendall's concordance-based correlation coefficient/AssI values.

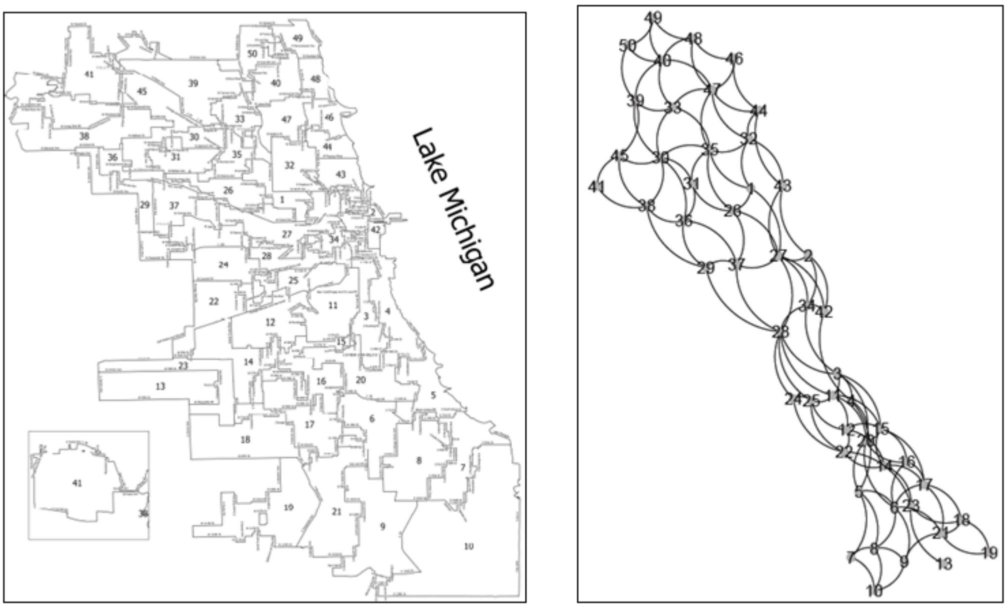
We chose the City of Chicago for our study, as the city has been one of the highest ranked cities with respect to crime rates (Chicago Crime Dataset). We build a ward network for the city (based on the ward map) wherein the nodes are the wards and there exists an edge between two nodes if the corresponding wards are coterminous (i.e., share a common boundary). We analyze the crime dataset for the City of Chicago for the year of 2022 and build a ward-crime type matrix wherein an entry (*i*, *j*) indicates the number of occurrences of crime type *j* in ward *i* during the year of 2022. We conduct Principal Component Analysis (PCA; Jolliffe, 2002) on this ward-crime type matrix and determine a weighted average score (referred to as the PC_crime_score) for each ward based on the entries for these wards in the high-variance principal components (variances ≥ 1.0; the variances are used as weights as well). We classify wards incurring positive PC_crime_scores to be potential crime hotspots and those with negative PC_crime_scores to be not potential crime hotspots. Per (Weisburd, 2015), a location in the city is a crime hotspot if the density of the crimes occurring in that location is significantly larger than the average in the city. We conduct assortativity analysis of the coterminous wards of the wards network at three levels: at a high-level, based on the potential crime hotspot classification (yes/no), at a mid-level, based on the PC_crime_scores of the individual wards and a low-level, based on the number of incidents of each crime type in the individual wards.

The rest of the paper is organized as follows: In Section 2, we present the Chicago Ward Network (CWN) constructed based on the ward map as well as present values for some of the salient network-level metrics. Section 3 presents the results of PCA of the crime dataset for the City of Chicago for the year 2022 and ranks the wards on the basis of the PC_crime_scores. Section 4 presents assortativity analysis of the CWN using three different set of node-level metric values for the end vertices of the edges. Section 5 presents the sensitivity analysis of our proposed PCA-based assortativity analysis approach across the different time periods of crime data collected for the City of Chicago. Section 6 reviews related work. Section 7 concludes the paper by summarizing its high-level contributions and also presents plans for future work.

2. Analysis of the Ward Network for the City of Chicago

The City of Chicago has a total of 50 wards and we use the ward map available at (Chicago Ward Map) as the basis for our study. Using this ward map, we construct a ward network wherein the nodes are the wards and there exists an edge between two nodes if the corresponding wards are coterminous (i.e., have a common border that is noticeable in the map). Figure 1 displays the ward map and the corresponding ward network (visualized per the Yifan Hu proportional layout algorithm: Gephi Tutorials) that we built based on this map. The topological layout is of the same orientation as that of the ward map. The Chicago Ward Network (CWN) has 50 nodes and a total of 118 edges connecting these nodes. We observe the city to be clearly a north-south oriented one with the ward # 28 (and to a lesser extent, ward #s 34 and 42) acting as a bridge between these two segments.

We now present values for some of the salient network-level metrics obtained for the CWN. The degree centrality-based assortativity index of the network is 0.1163. The randomness index (Meghanathan, 2017: measured as the Pearson's correlation coefficient of the degrees of the vertices and the average of the local clustering coefficient (Newman, 2012): LCC values for the nodes with a certain degree) is -0.9347 (indicating the edges of the network are not due to any random association). The spectral radius ratio for node degree (λ_k^{sp} : Meghanathan, 2014), a measure of the variation in node degree (irrespective of network size), is 1.1429, indicating the ward network is not scale-free (Barabasi & Albert, 1999). The link density (measured as the fraction of actual links and the maximum possible number of links between any two nodes in the CWN) is 0.0963. The algebraic connectivity of the network (Fiedler, 1973: a measure of the connectivity of the network with respect to the removal of two or more nodes) is 0.13. The bipartivity index (BPI: Ernada, E., & Rodriguez-Velazquez, 2005) of a network is a quantitative measure (ranging from 0.5 to 1.0) of the extent to which the nodes in the network can be partitioned to two disjoint sets such that majority of the edges are between nodes across the two partitions. A lower BPI value (0.6429: i.e., most of the edges are between vertices in the same partition: north and south) for the CWN confirms our earlier assertion (based on the visualization in Figure 1) that the city is oriented in the north-south direction, closely aligned with the city ward map (shown in the left image of Figure 1).



Chicago Ward Map

Chicago Ward Network (CWN)

Figure 1. Ward Map for the City of Chicago and the Corresponding Chicago Ward Network (CWN)

3. Principal Component Analysis of the Crime Dataset

We use the year 2022 crime dataset (Chicago Data Portal) for the City of Chicago as the primary source of data for our analysis and used the crime datasets for the years of 2020 and 2021 to conduct sensitivity analysis (reported in Section 4) of our proposed approach. The dataset reports a total of 238,801 crimes that fall under 25 crime types. Table 1 presents the names of the 25 crime types and the count for the number of crime incidents reported under each crime type. We build a ward-crime type matrix wherein the rows are the 50 wards and the columns are the 25 crime types. An entry (*i*, *j*) in this matrix corresponds to the number of incidents of crime type *j* that were reported to have occurred in ward *i*.

Table 1. Crime Types and their Count in the City of Chicago, Year 2022

#	Crime Type	Count	#	Crime Type	Count
C1	Arson	422	C14	Liquor Law Violation	203
C2	Assault	20803	C15	Motor Vehicle Theft	21448
C3	Battery	40924	C16	Narcotics	4731
C4	Burglary	7594	C17	Offense involving Children	1863
C5	Concealed Carry License Violation	177	C18	Other Offense	14574
C6	Criminal Damage	27242	C19	Prostitution	283
C7	Criminal Sexual Assault	1563	C20	Public Peace Violation	711
C8	Criminal Trespass	4224	C21	Robbery	8964
C9	Deceptive Practice	16524	C22	Sex Offense	1214
C10	Homicide	726	C23	Stalking	447
C11	Interference with Public Officer	393	C24	Theft	54844
C12	Intimidation	184	C25	Weapons Violation	8775
C13	Kidnapping	117			

We treat the wards as records and the crime types as features and conduct PCA on the 50 records x 25 features matrix to determine the high-variance principal components (of variance ≥ 1.0) that could be sufficient to capture the magnitude and variance of the # occurrences of the 25 crime types in each of the 50 wards. We observe just 4 high-variance principal components (of variances 13.63, 3.76, 1.94 and 1.33) are sufficient to capture (13.63+3.76+1.94+1.33 = 20.66)/25 = 82.64% of the information in the 50x25 ward-crime type matrix. For each ward (record/row), we then compute a weighted average of the entries for the record in the four high-variance principal components (with the corresponding variances used as weights). We refer to the weighted average as the PC_crime_score for the ward and could be considered a comprehensive measure of the number of occurrences reported for the ward with respect to all the crime types as well as relative to the number of occurrences reported in the other wards as well.

Figure 2 presents the PC_crime_scores for the 50 wards in the decreasing order of the values. We observe 19 of the 50 wards to incur positive PC_crime_scores and the rest 31 wards incur negative PC_crime_scores. Since the sum of the entries in any principal component should be zero, the sum of the PC_crime_scores of all the 50 wards should be zero as well. The wards with positive PC_crime_scores are candidates for classification as crime hotspots (shown in red color in Figure 2) in the city and the wards with negative PC_crime_scores can be regarded as relatively safe (i.e., NOT potential crime hotspots; shown in green color in Figure 2). We observe the wards in the northern areas of the city to be not potential crime hotspots, whereas several wards in the central and southern areas of the city are potential crime hotspots.

Ward #	PC_Crime_ Score	Potential Crime Hotspot
28	5.6191	Yes
6	4.6265	Yes
27	4.3917	Yes
24	4.1484	Yes
8	3.6457	Yes
20	3.5842	Yes
7	3.4983	Yes
42	3.3438	Yes
21	3.0983	Yes
17	3.0858	Yes
3	2.7330	Yes
9	2.5385	Yes
4	2.1659	Yes
16	2.0991	Yes
34	1.9940	Yes
37	1.5274	Yes
5	1.3116	Yes
29	1.0136	Yes
25	0.1937	Yes

Ward #	PC_Crime_S core	Potential Crime Hotspot
10	-0.1752	No
26	-0.3223	No
15	-0.9069	No
49	-1.0188	No
1	-1.0770	No
22	-1.2176	No
12	-1.2925	No
14	-1.3812	No
36	-1.4160	No
46	-1.6169	No
2	-1.6465	No
31	-1.6694	No
18	-1.7531	No
11	-1.7704	No
30	-1.8700	No
44	-1.8890	No
40	-1.9154	No
50	-1.9205	No
45	-1.9378	No
35	-2.0575	No
23	-2.1062	No
48	-2.1518	No
32	-2.2423	No
33	-2.2445	No
41	-2.3010	No
39	-2.3092	No
19	-2.3567	No
38	-2.3942	No
47	-2.5116	No
43	-2.5703	No
13	-2.5768	No

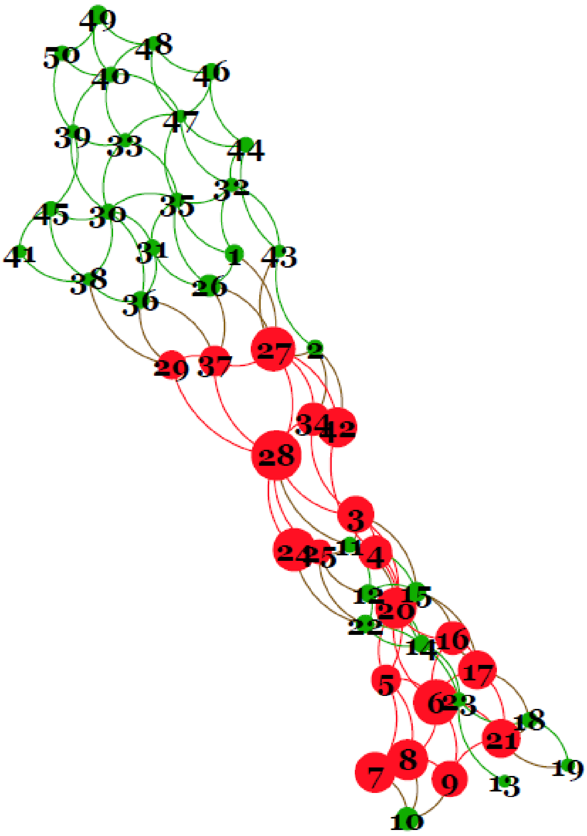


Figure 2. PC_crime_scores for the Chicago Wards and their Classification as Potential Crime Hotspots (Color: Red - Potential Crime Hotspots, Green - Not Potential Crime Hotspots
Node Size: proportional to the PC_crime_scores of the wards)

4. Assortativity Analysis of the Chicago City Ward Network

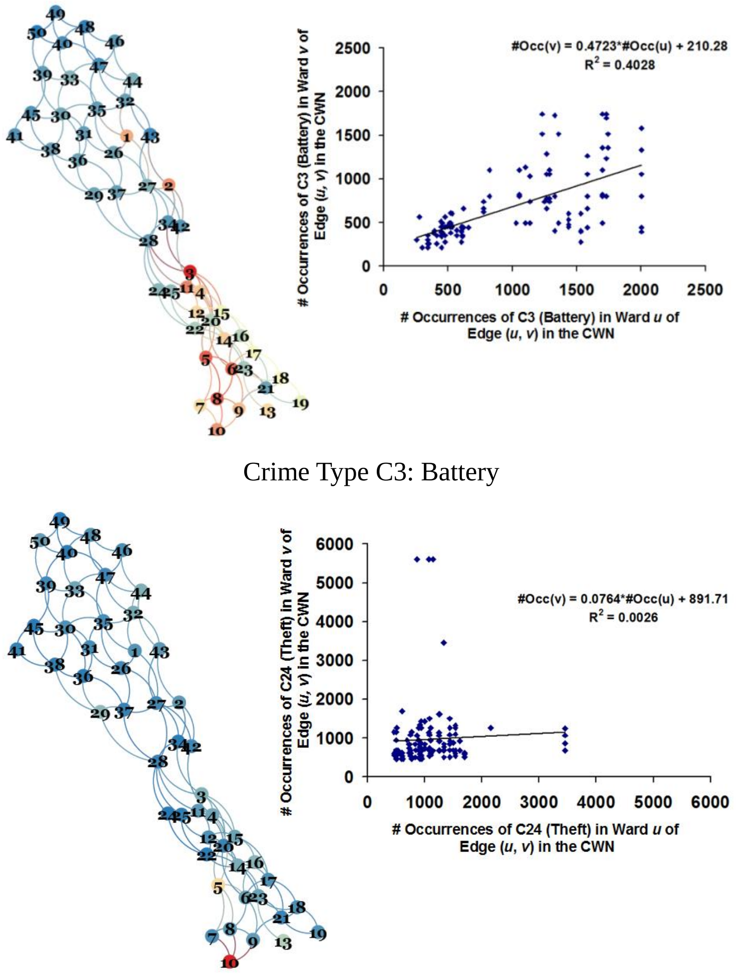
We conduct assortativity analysis of the Chicago city wards (modeled as CWN) at three levels. At a high-level, we seek to quantify the extent to which two wards both of which are not potential crime hotspots or two wards both of which are potential crime hotspots are coterminous. At a mid-level, we seek to quantify the similarity of coterminous wards with respect to their PC_crime_scores. At a low-level, we seek to quantify the similarity of the coterminous wards on the basis of the number of occurrences of crime incidents with respect to each of the 25 crime types. For the high-level hotspot classification-based assortativity analysis, we compute the Kendall's concordance-based correlation coefficient on the basis of the classification of the end vertices (wards) of the 118 edges in the ward network as potential crime hotspots or not potential crime hotspots. For the mid-level assortativity analysis, we compute the Pearson's correlation coefficient on the basis of the PC_crime_score values incurred by the end vertices of the edges in the ward network. For the low-level assortativity analysis, we compute the Pearson's correlation coefficient on the basis of the number of incidents of occurrence of each of the 25 crime types in the coterminous wards forming the edges of the ward network.

With respect to the computation of the high-level assortativity index (Kendall's concordance coefficient) on the basis of the classification of the end vertices (wards) of the edges as potential or not potential crime hotspots, we observe the following: of the 118 edges, 87 edges/pairs of end vertices are concordant (i.e., both the end vertices are either potential crime hotspots or both the end vertices are not potential crime hotspots) and 31 edges/pairs of end vertices are discordant (i.e., one of the end vertices is classified as a potential crime hotspot and the other end vertex is not a potential crime hotspot). The Kendall's hotspot concordance-based correlation coefficient is thus (87 - 31) / (87 + 31) = 0.4746, indicating a moderate-level of assortativity (in a scale of -1 to 1) for two adjacent wards to be both potential crime hotspots or both to be not potential crime hotspots. The right image of Figure 2 corroborates our findings. We observe the northern region of the city to be even strongly assortative (as any two adjacent wards are green in color) and the central region of the city also exhibits strong assortativeness (as any two adjacent wards are more likely red in color). It is the southern region of the city that exhibits lower assortativity with respect to the hotspot classification (for two wards that are coterminous, it is possible that one of these two wards could be a potential crime hotspot and the other ward is not a potential crime hotspot).

With respect to the AssI value computed on the basis of the PC_crime_scores of the wards, we observe the Pearson's correlation coefficient to be 0.6510, indicating a strong correlation between the coterminous wards with respect to the extent of # occurrences of all the crime types. In other words, for two coterminous wards (u_1, v_1) and (u_2, v_2), if the PC_crime_score(u_1) is appreciably greater (or lower) than the PC_crime_score(u_2), then the PC_crime_score(v_1) is also appreciably greater (or lower) than the PC_crime_score(v_2) as well. This implies that if a ward u_1 incurs a larger (or lower) # occurrences of the crime types than a ward u_2, then the wards adjacent to u_1 are also more likely to incur a larger (or lower) # occurrences of the crime types compared to that of wards adjacent to u_2. In Figure 2, such a trend is very much noticeable in the northern and southern regions of the city. That is, the neighboring wards of a ward in the northern region of the city incur a lower PC_crime_score compared to the neighboring wards of a ward in the southern region of the city. Hence, a larger AssI value of 0.6510 is also an indication that certain regions of the city are likely to be prominent crime hotspots compared to other portions of the city. The AssI value of 0.6510 obtained on the basis of the continuous PC_crime_scores is even more assortative compared to the AssI value of 0.4746 obtained on the basis of the binary classification (hotspot or not hotspot). The sizes of the nodes (wards) displayed in Figure 2 are proportional to their PC_crime_scores. The red-colored potential crime hotspot nodes are larger in size, whereas the green-colored not potential crime hotspot nodes are relatively smaller in size.

Figure 3 presents a color-coding of the CWN with respect to the number of occurrences of the crime types C3 (Battery) and C24 (Theft) in the wards. We observe the AssI values of the CWN with respect to C3 and C24 to be 0.6346 and 0.0509 respectively. In the ward network map color-coded on the basis of the number of foccurrences of C3, we observe that if a ward (say, ward # 33) in the northern region of the city reports fewer number of occurrences of C3 than a ward (say, ward # 10) in the southern region of the city, we observe the neighboring wards of ward # 33 also report fewer occurrences of C3 compared to the neighboring wards of ward # 10. Such a trend confirms the relatively larger positive Pearson's correlation coefficient (0.6346) for the CWN with respect to C3. This implies that certain wards (those in the southern region of the city) are more likely to be hotspots with respect to C3 compared to other wards (those in the northern region of the city). This also implies that the crime type (in this case, C3) being analyzed has the tendency to spread from a hotspot ward to its neighboring wards.

On the other hand, with respect to crime type C24, we observe the colors for the majority of the wards to be comparable to each other. In other words, if the number of occurrences of C24 in a ward (say, ward # 33) is lower than that of an another ward (say, ward # 10), the neighboring wards of ward # 33 need not also report relatively lower number of occurrences of C24 compared to that of the neighboring wards of ward # 10. This is very much observable in the color-coding map of the CWN with respect to C24 in Figure 3; a Pearson's correlation coefficient of 0.0509 for the CWN with respect to C24 confirms this observation that we cannot say anything about the neighboring wards of two wards *u* and *v* even if ward *u* reports a larger (or lower) number of occurrences of the crime type compared to that of ward *v*. Under such scenarios, there is less chance for any ward to be a hotspot; in the color-coded ward map with respect to C24, we also observe that only ward (ward # 10) could be really considered a hotspot in the city. On the other hand, there are several hotspots with respect to C34).



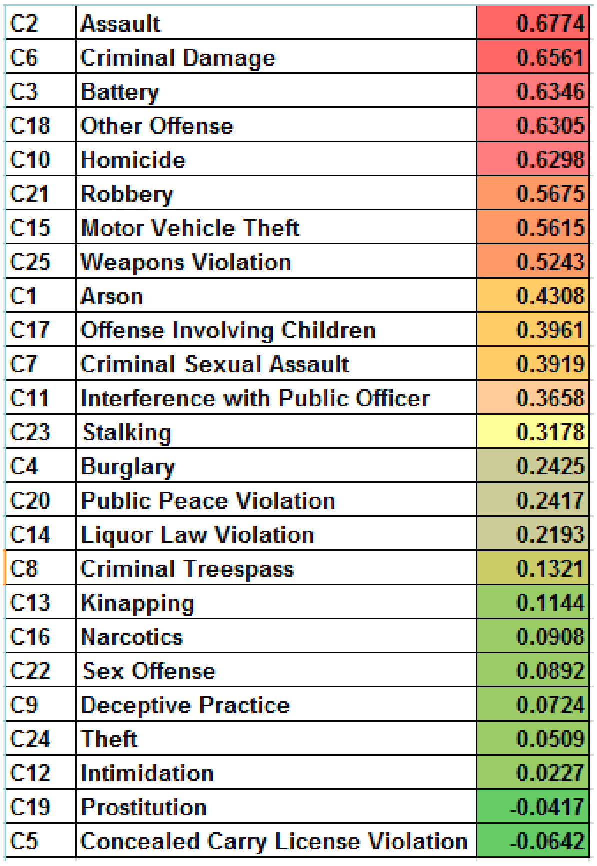
Crime Type C3: Battery

Crime Type C24: Theft

Figure 3. Assortativity Analysis of Crime Types C3 (Battery) and C24 (Theft) with respect to their Number of Occurrences in the Wards of the CWN

Figure 3 also displays the plot of the number of occurrences of wards *u* and *v* for all coterminous wards: edges (*u*, *v*) in the CWN with respect to C3 and C24. We represent the edges in the CWN in an order that for any edge (*u*, *v*): *u* < *v*. The Pearson's correlation coefficient of 0.6346 for the CWN with respect to C3 indicates a strong linear correlation between the number of occurrences of C3 in the coterminous wards *u* and *v*: i.e., we can even model a linear equation that could be used to predict the number of occurrences of C3 in a ward *v* of edge (*u*, *v*) using the number of occurrences of C3 in ward *u* of the edge. On the other hand, a close to zero Pearson's correlation coefficient of 0.0509 for the CWN with respect to C24 implies a lack of correlation between the number of occurrences of C24 in the coterminous wards and hence we cannot do any prediction of the number of occurrences of C24 in a ward *v* of an edge (*u*, *v*) using the number of occurrences of C24 in ward *u*.

Figure 4 presents the Pearson's correlation coefficient AssI values of the CWN based on the number of occurrences of the individual 25 crime types in the wards. Figure 4 also presents a heat map visualization of the AssI values for the CWN with respect to the 25 crime types. The CWN incurs larger AssI values for crime types (like C2: Assault, C6: Criminal Damage, C3: Battery and C10: Homicide) for which certain wards standout to be prominent crime hotspots compared to other wards. On the other hand, the CWN incurs close to zero AssI values for crime types such as C24: Theft, C12: Intimidation, C19: Prostitution, C5: Concealed Carry License Violation, for which no ward is likely to be a prominent hotspot. Thus, the AssI value (ranging from being closer to 0 to 1, as noticed in Figure 4) for the CWN with respect to the number of occurrences of a particular crime type in the wards could be considered as a quantitative measure of the extent to which any ward could be a hotspot ward for the particular crime type as well as the extent to which the "hotspot nature" for the crime type could spread from a ward to its adjacent wards.



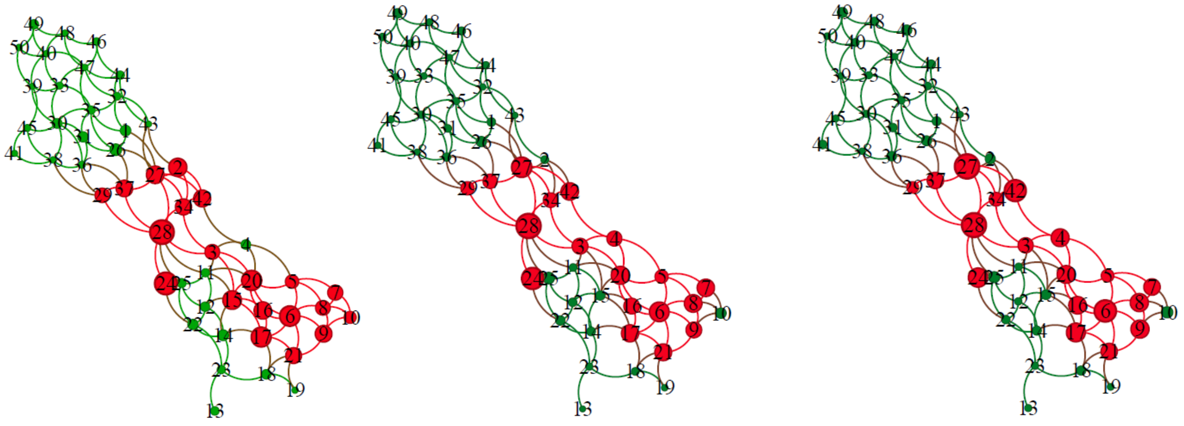
C2	Assault	0.6774
C6	Criminal Damage	0.6561
C3	Battery	0.6346
C18	Other Offense	0.6305
C10	Homicide	0.6298
C21	Robbery	0.5675
C15	Motor Vehicle Theft	0.5615
C25	Weapons Violation	0.5243
C1	Arson	0.4308
C17	Offense Involving Children	0.3961
C7	Criminal Sexual Assault	0.3919
C11	Interference with Public Officer	0.3658
C23	Stalking	0.3178
C4	Burglary	0.2425
C20	Public Peace Violation	0.2417
C14	Liquor Law Violation	0.2193
C8	Criminal Treespass	0.1321
C13	Kinapping	0.1144
C16	Narcotics	0.0908
C22	Sex Offense	0.0892
C9	Deceptive Practice	0.0724
C24	Theft	0.0509
C12	Intimidation	0.0227
C19	Prostitution	-0.0417
C5	Concealed Carry License Violation	-0.0642

Figure 4. Assortativity Analysis of Crime Types C3 (Battery) and C24 (Theft) with respect to their Number of Occurrences

5. Sensitivity Analysis

In this section, we demonstrate the stability of our proposed PCA-based approach to quantitatively assess the assortativity and spreadability of crimes across coterminous wards across different time periods (for the years of 2012, 2022 and 2023). In Sections 3 and 4, we exclusively use the Chicago crime dataset for the year of 2022 to present and assess the results. In this section, we compare the results obtained for the year 2022 with those for crime datasets retrieved for a year ahead (2023) and 10 years back (2012). Figure 5 presents the visualization of the CWN based on the hotspot classification of the wards (green: not hotspots and red: hotspots) as well as the PC_crime_scores of the wards (the larger the node, the more severe is the overall occurrence of crimes in the ward). Figure 5 also presents the assortativity index values calculated for the three years with respect to the binary hotspot classification and the continuous PC_crime_scores.

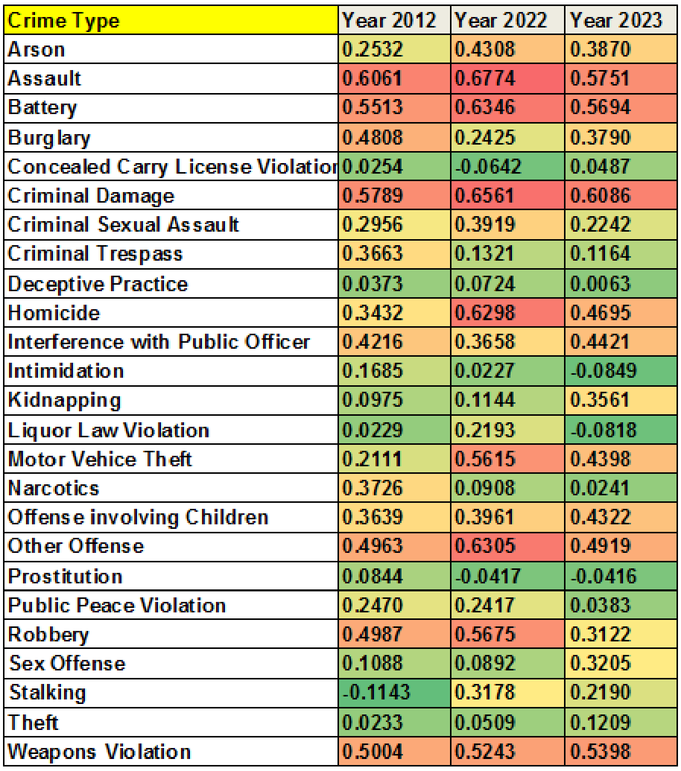
Overall, we observe the assortativity index results with respect to both the metrics (hotspot classification and PC_crime_scores) to be stable across the years. The general trend is that we see an overall decrease in the assortativity in the years 2022 and 2023 compared to the year 2012. This could be actually attributed to the change in the hotspot classification for wards 2, 4 and 10 in the central, south central and southern most regions of the city respectively from being a hotspot (in year 2012) to not a hotspot in years 2022 and 2023. With their neighboring wards undergoing no change (especially for wards 2 and 10), this resulted in the decrease of assortativity index. The hotspot classification for the rest of the 47 wards remained the same for all the three years (2012, 2022 and 2023): i.e., the northern region of the city exhibited no crime hotspots, whereas the central and southern region exhibited crime hotspots. We also observe the node sizes (especially for those wards that remained as hotspots in all the three years) to be relatively the same across the three years: i.e., wards exhibited the same severity of crime across the three years.



Year 2012 Year 2022 Year 2023

Hotspot classification-based assortativity index	0.5932	0.4746	0.4746
PC_crime_scores-based assortativity index	0.5425	0.5172	0.4851

Figure 5. Visual Comparison of the Hotspot Classification and PC_crime_scores of the CWN across the Different Years



Crime Type	Year 2012	Year 2022	Year 2023
Arson	0.2532	0.4308	0.3870
Assault	0.6061	0.6774	0.5751
Battery	0.5513	0.6346	0.5694
Burglary	0.4808	0.2425	0.3790
Concealed Carry License Violatior	0.0254	-0.0642	0.0487
Criminal Damage	0.5789	0.6561	0.6086
Criminal Sexual Assault	0.2956	0.3919	0.2242
Criminal Trespass	0.3663	0.1321	0.1164
Deceptive Practice	0.0373	0.0724	0.0063
Homicide	0.3432	0.6298	0.4695
Interference with Public Officer	0.4216	0.3658	0.4421
Intimidation	0.1685	0.0227	-0.0849
Kidnapping	0.0975	0.1144	0.3561
Liquor Law Violation	0.0229	0.2193	-0.0818
Motor Vehice Theft	0.2111	0.5615	0.4398
Narcotics	0.3726	0.0908	0.0241
Offense involving Children	0.3639	0.3961	0.4322
Other Offense	0.4963	0.6305	0.4919
Prostitution	0.0844	-0.0417	-0.0416
Public Peace Violation	0.2470	0.2417	0.0383
Robbery	0.4987	0.5675	0.3122
Sex Offense	0.1088	0.0892	0.3205
Stalking	-0.1143	0.3178	0.2190
Theft	0.0233	0.0509	0.1209
Weapons Violation	0.5004	0.5243	0.5398

Figure 6. Assortativity Index of the CWN with respect to the Number of Occurrences of the Individual Crime Types across the Different Years

Figure 6 presents the assortativity index for the CWN with respect to the individual crime types for the three years; the median of the assortativity index values are 0.2956 (year 2012), 0.3178 (year 2022) and 0.3205 (year 2023), remaining stable across the different time periods (though the assortativity index values for certain individual crime types exhibited a larger change). While crimes that incurred very high assortativity index (such as Assault, Battery, Criminal Damage, Weapons Violation) exhibited comparable levels of spreadability across the three years. On the other hand, crime types like Arson, Motor Vehicle Theft and Stalking appear to have become more spreadable (i.e., exhibited a relatively larger increase in assortativity index) with time. We observe that crime types like Burglary, Narcotics, Public Peace Violation and Criminal Trespass exhibit a tendency to have become less spreadable (i.e., incurred a relatively larger decrease in assortatvity index) with time. Also, crime types like Prostitution, Theft, Intimidation, Liquor Law Violation continue to remain least spreadable across the three years.

6. Related Work

The approach presented in this paper to assess crime hotspots on the basis of assortativity analysis could be very useful for the Police Department to effectively mobilize the resources for targeted intervention and take informed decisions as part of proactive policing. Very limited work is available in the literature on the use of assortativity analysis to identify crime hotspots or mitigate the spreading of the crime from a hotspot to its neighboring locations. Sivaranjani & Sivakumari (2015) presented a triangulation interpolation-based approach along with shortest paths-based centrality metrics (closeness and betweenness) to detect hotspots that are prone to crimes identical to each other (the similarity is assessed using assortativity analysis). Conrow et al (2015) conducted spatio-temporal analysis to study the influence of on-premise alcohol outlets on crime events in the City of Buffalo, NY. The study focused on assessing whether the occurrence of crime events increased since the opening of alcohol outlet(s) near by, as well as whether the crime events spread over a broader area. While the global bivariate space-time K-function analyses indicated a dispersion between the alcohol outlets and crimes over space and time, the local analyses indicated clustering of the crimes around an alcohol outlet at discrete space-time intervals.

Assortativity analysis of criminal networks has been reported in several studies. Duxbury, S., & Haynie (2020) found that intentional attacks on criminal networks decrease the degree-degree assortativity and promote preferential attachment. On the other hand, signal attacks (for e.g., attack on online selling networks using false fake reviews: opinion spamming) increase degree-degree assortativity, especially at low levels of intervention. Borgatti (2006) proposed the removal of critical nodes in a criminal network without which the communication among the remaining nodes would be severely disrupted. To bring down the resilience and redundancy of criminal networks, Catanese et al (2016) advocate attacking the weak nodes and links on redundant paths to make them unavailable when the critical nodes and links that are actually used for communication fail. Scale-free networks have been observed to be robust to random node and link removals, but very vulnerable to targeted node and link removals (Holme et al., 2002).

Unlike other social networks, it is hard to know the exact topology of criminal networks. Budur et al (2015) advocate the use of link prediction techniques to detect and incorporate hidden links in a criminal network after which the network could be subject to further analyses for resilience, robustness, etc. The success of the link prediction techniques lies in how effective are they in predicting links that happen to be the hidden links (that are not available for public information) in the network. In a related work Clauset et al (2008) observed that link prediction techniques based on the notions of common neighbors, shortest path length, product of degrees, etc are more effective for assortative networks and can be misleading if the network is not assortative. Xu & Chen (2008) used co-occurrence analysis of the crime incident records to predict the links between the 3917 criminals involved in gang-related crimes in Tucson from 1985 to 2002 to construct and analyze their criminal network. However, as argued in the same paper, co-incidental or co-occurrence links may not be a reflection of the real relationship between two criminals, especially if they did not commit a crime together. If a criminal network constructed mainly from anecdotal evidence and empirical evidence is not connected, network analysis studies were further conducted on the giant component (the largest connected sub graph of the network) alone. One such study (Xu & Chen, 2008) on the giant components of four criminal networks observed a terrorist network and gang network to be degree-assortative, but a drug network and Dark web network were observed to be degree-dissortative. This was attributed to the nature of the commands, businesses involved in these networks. Sometimes the presence of multiple components in a criminal network might also suggest the possibility of differing opinion among the members of the network with regards to their ideologies, kinships, etc.

Cavallaro et al (2020) constructed two different criminal networks based on information through the phone calls and physical meetings of members of a Mafia organization that operated in Sicily in the early 2000s. The authors evaluated the effectiveness of different intervention procedures (such as sequential node removal, node block removal and etc) by evaluating the centrality of the nodes before and after the interventions. It has been observed that by removing the top 5% of the nodes ranked based on betweenness centrality, the network connectivity could be made to drop by as large as 70%. The authors also did not observe any noticeable difference in the results of network analysis conducted on weighted vs. unweighted graphs. In a similar study, Mastrobuoni &

Patacchini (2012) observed the US Mafia networks (that were active in the 1950s and 1960s) to be extremely hierarchical; though enforcement agents could break the chain of command by arresting mobsters who occupy critical positions in the hierarchy, the hierarchy was observed to be kinship and trust-driven and hence, difficult to break. Criminals were observed to associate with other criminals with similar kind of business, both legal and illegal (i.e., a positive assortative matching); criminal networks that involved illegal businesses typically exhibited higher assortativity compared to criminal networks that involved legal businesses. For example, criminal networks based on drug offenses were observed to exhibit assortativity index as large as 0.939; whereas, criminal networks based on drug stores were observed to be almost neutral (with an assortativity index of 0.184). The study also observed that a sort of strategic endogamy prevails in the Mafia networks: i.e., marriages tend to happen between people within a trusted community rather than among people in two different communities. The centrality of the members in the Mafia networks is also observed to increase with age.

Gulumbe et al (2012) used PCA to assess the predominance of specific crimes in specific areas of Katsina state in Nigeria. The loading factors of the crimes in the principal components were used as the basis to come up with the assessment. In another crime analysis PCA study (Usman et al., 2012) based on Nigeria, the authors observed moderate correlation between crimes on properties and crimes on persons; the study also used exploratory factor analysis to come up with the assessment. While our work takes a predictive analytics approach (to assess the spreadability of crimes across a geographical region from one ward to the neighboring wards), the above two studies take a data analytics (exploratory factor analysis) approach of just assessing how crimes were reported in the regions across the state.

Wu & Meghanathan (2023) proposed a PCA-based spatio-temporal hotspot score for a particular crime in each sub region of a major city (Boston): the underlying matrix used to conduct PCA comprised of the sub regions as rows and the number of occurrences of a particular crime (Larceny) across each of the 24 hours of a day. The spatio-temporal hotspot score for a sub region is computed as the weighted average of the entries for the sub region in all the principal components, with the variance of the principal components used as the weights. While the spatio-temporal hotspot score could be used to rank the sub regions on the basis of the severity of the crime, it cannot be used to assess the spreadability of the crime across the sub regions. In a time-series analysis study (utilizing dynamic linear models; Garton & Niemi, 2019) of the crime records of the City of Chicago from 2007 to 2016 observed substantial differences in the trends of violent crimes vs. non-violent crimes as well as reporting pattern of the crimes were observed to differ depending on the nature of the crimes.

Mart ńez-Lanz et al (2021) used PCA and Chi-Square analysis (Strang, 2006) to assess the relationship between three factors (domestic violence, substance abuse, socio economic conditions) that were considered to contribute to male criminal behavior among the inmates of the inmates of the Federal Center for Social Rehabilitation number 7 in Mexico. While the PCA results showed the grouping of the domestic violence, alcohol consumption and substance abuse with a positive loading in the first principal component (that captured 56% of the variance in the dataset), Chi-square analysis indicated a higher degree of correlation between low socioeconomic conditions and criminal activity as well as statistically significant differences in the correlations between the sociodemographic factors, addictions, domestic violence and depressive symptoms. Crespo et al (2023) used multivariate statistical tools (such as clustering and HJ-Biplot) to assess the possible correlations between crime vs. {poverty and unemployment} in the country of Eucador from January 2021 to May 2022. While no strong correlation was observed on the overall dataset collected across all regions of Eucador, the Amazon provinces were observed to exhibit a strong correlation between rape (a crime) and poverty (a socioeconomic factor). Chung & Kim (2019) conduct multivariate spatial analysis (multivariate conditional autoregressive model) of the crime data for the city of San Francisco (the # occurrences of vehicle thefts, larceny and burglary in the 83 census tracts) for the year 2010 and generate *joint* crime risk maps that could potentially capture the spatial dependence between the different sites and the dependence between the crime types with respect to the number of occurrences of the crimes across the city.

7. Conclusions and Future Work

The high-level contributions of this paper are the following: (1) We propose the idea of constructing a ward network by identifying coterminous wards in a corresponding city map and using the constructed ward network to assess the assortativity of the wards on the basis of three levels of node-level crime metrics computed for these wards from crime datasets. (2) We propose to use the Kendall's concordance-based correlation coefficient for assortativity analysis involving binary data for the node-level metric (whether a ward is a potential crime hotspot or not) and we observe this approach to be very effective as the result resonates the assortativity analysis conducted by computing the Pearson's correlation coefficient on the continuous PC_crime_scores for the wards.

(3) We observe that if a ward network incurs larger values of assortativity index (Pearson's correlation coefficient) with respect to the number of occurrences of a crime type in the wards, some of the wards could be prominent hotspots for the crime type (compared to other wards that are typically away from it) and the neighboring wards are also more likely to be hotspots (indicative of a crime spread). We thus propose that the assortativity index value incurred for a crime type (on the basis of the # occurrences of the crime type in the individual wards) could be considered the "spreading index" (capturing the extent of spread possible) as well as the "hotspot index" (capturing the extent to which there could be hotspot locations) for the crime type. (4) We also show that Principal Component Analysis (PCA) could be effective in reducing the dimensionality of the crime datasets featuring multiple crime types across the regions (wards) of a city. We conducted PCA on a crime dataset featuring the number of occurrences of a suite of 25 crime types (as features/columns) in the 50 wards (as records/rows) of the City of Chicago. We observed that just four of the 25 principal components (PCs) to be of high-variance (variance ≥ 1.0) and showed that a weighted average PC_crime_score computed for each ward (based on the entries for these wards in the high-variance PCs and the variances of these PCs as weights) could be used as the basis for classifying the ward as either potential crime hotspot or not.

The assortativity index scores of the wards on the basis of the PC_crime_scores as well as the individual crime types could be used as the basis for targeted intervention and predictive policing. The CWN visualization for all the three years in Figure 5 indicate that the hotspot wards occur in clusters and are not isolated (i.e., surrounded by wards that are not hotspots). Our conjecture is that a cluster of hotspot wards needs to be disconnected to multiple disjoint components to reduce the spreading of crimes across coterminous wards. In this context, the law enforcement agencies could be to first target wards that incurred smaller but positive PC_crime_scores (and as a result are currently classified as hotspots) and take measures to reduce the crime rate at these wards so that these wards do not get classified as hotspots. Once this is accomplished, the targeted intervention could gradually shift towards combating the crime rates at wards that incur larger positive PC_crime_scores.

As part of future work, we plan to apply the proposed PCA-based technique to assess the crime assortativity of wards in other metropolitan cities as well as explore in detail the use of the Pearson's correlation coefficient-based assortativity index for individual crime types as the basis for quantifying the extent of crime spread possible within a neighborhood of wards. We also plan to conduct a partial correlation coefficient analysis (Meghanathan, 2019)-based study of the Chicago Ward Network with respect to the number of occurrences of the individual crime types to nullify the influence of the wards that are two hops away. We also plan to assess the crime similarity of wards (that need not be contiguous) by determining logical clusters of wards with respect to their PC_crime_scores.

Acknowledgments

Not applicable

Authors contributions

Not applicable

Funding

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-21-1-0264. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Competing interests

Not applicable

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education.

The journal’s policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

Copyrights

References

Barabasi, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509-512. https://doi.org/10.1126/science.286.5439.509

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, *92*(5), 1170-1182. https://doi.org/10.1086/228631

Borgatti, S. P. (2006). Identifying Sets of Key Players in a Social Network. *Computing & Mathematical Organization Theory*, *12*, 21-34, 2006. https://doi.org/10.1007/s10588-006-7084-x

Budur, E., Lee, S., & Kong, V. S. (2015). Structural Analysis of Criminal Network and Predicting Hidden Links using Machine Learning. arXiv: 1507.05739v3 [cs.SI].

Catanese, S., De Meo, P. & Fiumara, G. (2016). Resilience in Criminal Networks. *AAPP Physical, Mathematical and Natural Sciences*, *94*(2), 1-19. https://doi.org/10.1478/AAPP.942A1

Cavallaro, L., Ficara, A., De Meo, P., Fiumara, G., Catanese, S., Bagdasar, O., ... Liotta, A. (2020). Disrupting Resilient Criminal Networks through Data Analysis: The Case of Sicilian Mafia. *PLoS ONE*, *15*(8), e0236476. https://doi.org/10.1371/journal.pone.0236476

Chicago Crime Dataset, https://en.wikipedia.org/wiki/Crime_in_Chicago

Chicago Data Portal, https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

Chicago Ward Map, https://app.chicagoelections.com/Documents/general/Citywide%20Ward%20Map%202022.pdf

Chung, J., & Kim, H. (2019). Crime Risk Maps: A Multivariate Spatial Analysis of Crime Data. *Geographical Analysis*: *An International Journal of Theoretical Geography*, *51*(4), 475-499. https://doi.org/10.1111/gean.12182

Clauset, A., Moore, C., & Newman, M. E. J. (2008). Hierarchical Structure and the Prediction of Missing Links in Networks. *Nature*, *453*, pp. 98-101. https://doi.org/10.1038/nature06830

Conrow, L., Aldstadt, J., & Mendoza, N. S. (2015). A Spatio-Temporal Analysis of On-Premises Alcohol Outlets and Violent Crime Events in Buffalo, NY. *Applied Geography*, *58*, 198-205. https://doi.org/10.1016/j.apgeog.2015.02.006

Crespo, A., Brito, J., Ajala, S., Amaro, I. R., & Castillo, Z. (2023). Multivariate Statistical Techniques to Analyze Crime and its Relationship with Unemployment and Poverty: A Case Study. *Paper presented at the Computer Science Online Conference*. https://doi.org/10.1007/978-3-031-35314-7_18

Duxbury, S., & Haynie, D. L. (2020). The Responsiveness of Criminal Networks to Intentional Attacks: Disrupting Darknet Drug Trade. *PLoS ONE*, *15*(9), e0238019. https://doi.org/10.1371/journal.pone.0238019

Ernada, E., & Rodriguez-Velazquez, J. A. (2005). Spectral Measures of Bipartivity in Complex Networks. *Physical Review E*, *72*(4), 046105. https://doi.org/10.1103/PhysRevE.72.046105

Fiedler, M. (1973). Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal*, *23*(98), 298-305. https://doi.org/10.21136/CMJ.1973.101168

Freeman, L. (1977). A Set of Measures of Centrality based on Betweenness. *Sociometry*, *40*(1), 35-41. https://doi.org/10.2307/3033543

Freeman, L. (1979). Centrality in Social Networks: Conceptual Classification. *Social Networks*, *1*(3), 215-239. https://doi.org/10.1016/0378-8733(78)90021-7

Garton, N., & Niemi, J. (2019). Multivariate Temporal Modeling of Crime with Dynamic Linear Models. *PLoS One*, *14*(7), e0218375. http://doi.org/10.1371/journal.pone.0218375

Gephi Tutorials, https://gephi.org/tutorials/gephi-tutorial-layouts.pdf

Gulumbe, S, U., Dikko, H. G., & Bello, Y. (2012). Analysis of Crime Data using Principal Component Analysis: A Case Study of Katsina State. *CBN Journal of Applied Statistics*, *3*(2), 39-49. https://dc.cbn.gov.ng/jas/vol3/iss2/3

Holme, P., Kim, B. J., Yoon, C. N., & Han, S. K. (2002). Attack Vulnerability of Complex Networks. *Physical Review E*, *65*, # 056109, 1-14. https://doi.org/10.1103/PhysRevE.65.056109

Jolliffe, I. T. (2002). *Principal Component Analysis* (1st ed.) Springer Series in Statistics, Springer-Verlag.

Mart ńez-Lanz, P., Cuevas-Covarrubias, C., & Hern ández-Valdez, P. (2021). Principal Component Analysis of Male Criminal Behavior. *Health*, *13*(10), 1112-1128. https://doi.org/10.4236/health.2021.1310083

Mastrobuoni, G., & Patacchini E. (2012). "Organized Crime Networks: An Application of Network Analysis Techniques to the American Mafia," *Review of Network Economics*, *11*(3), article # 10, 1-42. http://doi.org/10.1515/1446-9022.1324

Meghanathan, N. (2014). *Spectral Radius as a Measure of Variation in Node Degree for Complex Network Graphs*. Paper presented at the 3rd International Conference on Digital Contents and Applications, Hainan, China. https://doi.org/10.1109/UNESST.2014.8

Meghanathan, N. (2016). Assortativity Analysis of Real-World Network Graphs based on Centrality Metrics. *Computer and Information Science*, *9*(3), 7-25. https://doi.org/10.5539/cis.v9n3p7

Meghanathan, N. (2017). Randomness Index for Complex Network Analysis. *Social Network Analysis and Mining*, *7*(25), 1-15. ttps://doi.org/10.1007/s13278-017-0444-3

Meghanathan, N. (2019). Centrality and Partial Correlation Coefficient-based Assortativity Analysis of Real-World Networks. *The Computer Journal*, *62*(9), 1247-1264. https://doi.org/10.1093/comjnl/bxy098

Newman, M. E. J. (2010). *Networks*: *An Introduction* (1st ed.) Oxford University Press

Noldus, R., & Van Mieghem, P. (2015). Assortativity in Complex Networks. *Journal of Complex Networks*, *3*(4), 507-542. https://doi.org/10.1093/comnet/cnv005

Sivaranjani, S., & Sivakumari, S. (2015). Mitigating Serial Hotspots on Crime Data using Interpolation Method and Graph Measures. *International Journal of Computer Applications*, *126*(7), 17-25. http://doi.org/10.5120/ijca2015906088

Strang, G. (2006). *Linear Algebra and its Applications* (4th ed.) Brooks Cole

Usman, U., Yakubu, M., & Bello, A. Z. (2012). Investigation on the Rate of Crime in Sokoto State using Principal Component Analysis," *Nigerian Journal of Basic and Applied Sciences*, *20*(2), 152-160.

Weisburd, D. (2015). The Law of Crime Concentration and the Criminology of Place. *Criminology*, *53*, 133-157. https://doi.org/10.1111/1745-9125.12070

Wu, Y., & Meghanathan, N. (2023). A Principal Component Analysis-Based Scoring Mechanism to Quantify Crime Hot Spots in a City. *Paper presented at the ITNG 2023 20th International Conference on Information Technology - New Generations*, Las Vegas, NV, USA. https://doi.org/10.1007/978-3-031-28332-1_6

Xu, J., & Chen, H. (2008). The Topology of Dark Networks. *Communications of the ACM*, *51*(10), 58-65. https://doi.org/10.1145/1400181.1400198