Journal of Forensic and Legal Medicine 55 (2018) 76–86

Contents lists available at ScienceDirect

Journal of Forensic and Legal Medicine

journal homepage: www.elsevier.com/locate/yjflm

Research Paper

Adopting data interpretation on mining fine-grained near-repeat patterns in crimes

Ke Wang[b], Zhiping Cai[b], Peidong Zhu[a,*], Pengshuai Cui[b], Haoyang Zhu[b], Yangyang Li[c]

[a] Department of Electronic Information and Electrical Engineering, Changsha University, Changsha, China
[b] College of Computer, National University of Defense Technology, Changsha, China
[c] Innovation Center, China Academy of Electronics and Information Technology, Beijing, China

ARTICLE INFO

Keywords:
Crime analysis
Near-repeat effect
Data interpretation
Crime patterns mining
Knotted-clues method

ABSTRACT

The near-repeat effect is a well-known phenomenon in crime analysis. The classic research methods focus on two aspects. One is the geographical factor, which indicates the influence of a certain crime risk on other similar crime incidents in nearby places. The other is the social network, which demonstrates the contacts of the offenders and explain "near" as degrees instead of geographic distances. In our work, these coarse-grained patterns discovering methods are summarized as bundled-clues techniques. In this paper, we propose a knotted-clues method. Adopting a data science perspective, we make use of a data interpretative technology and discover that the near-repeat effect is not always so near in geographic or network structure. With this approach, we analyze the near-repeat patterns in all districts of the dataset, as well as in different crime types. Using open source data from Crimes in Chicago provided by Chicago Police Department, we find interesting relationships and patterns with our mining method, which have a positive effect on police deployment and decision making.

1. Introduction and related works

Criminological studies have demonstrated that repeat crimes are essential fundamental phenomenons.[1] And the near-repeat effect is widely known because it reveals the elevated tendency between crime incidents taking place nearby in both space and time.[2] The major near-repeat researches concentrate on two aspects. One aspect pays attention to the crimes in particular type.[3] The near-repeat phenomenon is first discovered in burglary,[4] which is still a hot topic even today.[5,6] There are also researches on other single crime types, including robbery,[7] shooting[8,9] and assault[10].[11] assesses the near-repeat phenomenon by Mean Frequencies in three crime types of shooting, robbery and auto theft. But the depth of this research inclines to the unique spatio-temporal pattern of each type, without penetrating deep into the relationships of the types. The other aspect takes notice of the individuals or the relationships among criminal suspects with the help of social network thoughts and research methods.[12] utilises epidemiological methods to investigate the phenomenon in the offence of burglary.[13] talks about the same offenders involved in near-repeat burglaries.[14] analyzes the initiator and near-repeat events in burglary and motor vehicle theft.

When it comes to near-repeat phenomenons, we have to mention the importance of locating crime scene in criminal cases, which is the core of the criminal investigations.[15] Through the analysis of forensic experts, the police can get the autopsy report,[16] the botanical analysis[17] and analysis of evidences, such as hanging marks[18] and mobile devices.[19] In order to make best uses of the information that has been mastered in similar cases, the studies of near-repeat effect and crime patterns are of great significance.

From the perspective of crime patterns mining, patterns are based on time or space.[20] finds the spatial behavioural patterns of the individual burglar.[21] makes the offenses cluster in time, and finds that crimes often occur at particular time. However, given the problems of non temporal or spatial factors, there is no proper method to deal with it. The techniques are usually on the basis of some mathematical distributions, such as, *Poisson distribution*[22] and non-hierarchical clustering method, for example, the k-means clustering.[23] However, crime data in the real world rarely matches specific data distribution. And the k-means clustering completely depends on the coordinates of each point, which is limited by Euclidean distance and the mean value. In our paper, we choose hierarchical clustering as a basic step, which is an efficient technique for data mining.[24]

In the researches aforementioned, the measurement of "near" is usually related to geographical positions or degrees in social network.

* Corresponding author. Hongshan Road 98#, Changsha, Hunan, China.
E-mail addresses: wangke@nudt.edu.cn (K. Wang), zpcai@nudt.edu.cn (Z. Cai), pdzhu@nudt.edu.cn (P. Zhu), cuipengshuai@nudt.edu.cn (P. Cui), zhuhaoyang@nudt.edu.cn (H. Zhu), yli@csds.lab.net (Y. Li).

https://doi.org/10.1016/j.jflm.2018.02.015
Received 26 September 2017; Received in revised form 17 January 2018; Accepted 11 February 2018
Available online 17 February 2018
1752-928X/

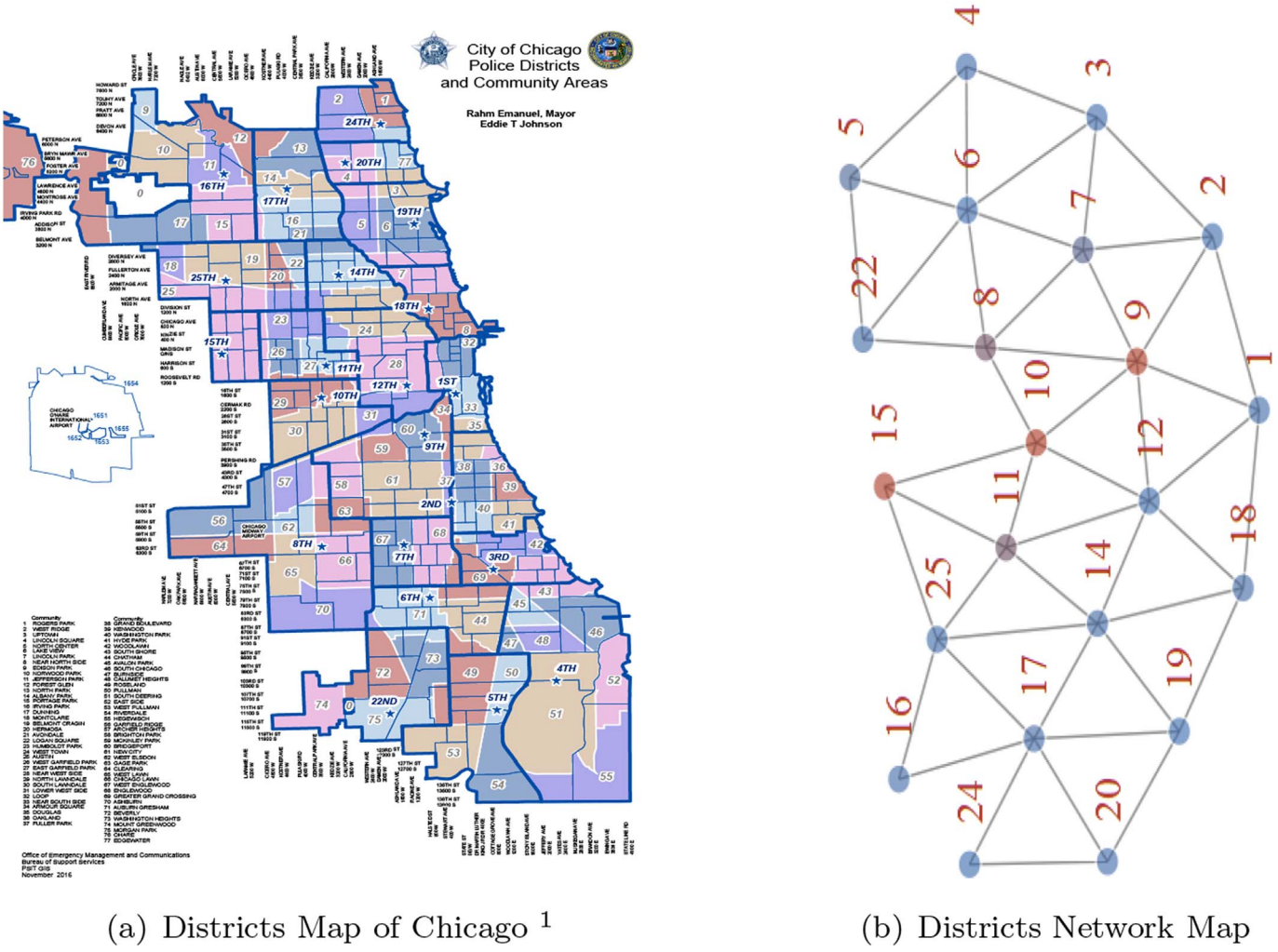


Fig. 1. Districts in Chicago. (a) Districts map of Chicago .[1] (b) Districts network map.

In these works, every clue leads to a result, and the final conclusion is a combined effect. This approach puts all the clues together into a bundle, so we call it the bundle-clues method. However, these kinds of definitions lead to a problem, which is whether the same measurement of "near" means the same influence of repeat crimes. In other words, this kind of "near" definition is coarse-grained. When predicting the crime incidents, we may care more about which one is the next place in the vicinity rather than whether within the nearby locations. For instance, in Fig. 1(b), around node 9, it is good that if we can predict a certain crime incident will happen in the near nodes. But if we can make an accurate prediction that it will happen in node 8 or 10, that is better. In order to achieve this goal, we have to process the clues in a fine-grained way, just like to knot them into a long one, which is named the knotted-clues method. More significantly, the sorting of the steps in this new method needs to have a reasonable explanation of data processing order, which is called data interpretation in our work.[1]

With the development of information technology, the scope of people's activities and capabilities are further expanded, including the criminal activities. Accordingly, the crimes may become more and more complicated, which means different types of crimes interrelate with each other or the criminal companions have never met face to face. At this point, the traditional methods may be incompetent. Taking Fig. 1 for example, (b) is the network structure diagram of (a) which is the districts map provided by the Chicago Police Department. In this network, District 10 is next to District 8, 9, 11, 12 and 15, where the degree is one. All the five districts are in the near vicinity of District 10, but the crime risk of District 10 is impossible to affect these five areas equally. And also the repeatability of crimes in district cluster 8-10-11 is in fact different from cluster 9-10-15, according to either time or crime types. Then, from the perspective of data science, we take advantages of data science methods[25] to interpret the facts and solve the problems.

Through the open source data provided by Chicago Police Department, we find that the near-repeat effect does not always choose the near geographical distance or small degree. And we mine the near-repeat patterns from 25 different crime types. The paper is organized as follows. Section 2 specifies our knotted-clues method. In Section 3, we provide experimental results and analysis. Discussions are made in Section 4. Finally, the paper ends by the conclusions in Section 5.

2. Methods

The bundled-clues can be interpreted as Formula (1), where c_n stands for clue n, f_n represents some kind of function or correspondence, r_n means one of the direct results, R is the final result and g indicates the integrated function. The researchers extract clues from the data, and then utilize one or several methods or functions to get some results, and integrate them together in the end. It is a very effective methodology, from which our work has received a great deal of inspiration. However, for the near-repeat research, the accuracy of this method needs to be improved. On the basis of this method, we propose the knotted-clues

[1] https://home.chicagopolice.org/community/community-map/.

approach, as shown in Formula (2), which takes the results of this step as input to the next step. Our method takes the bundled-clues as a first step. In another word, if there is inadequate clues or functions, a simple one step knotted-clues can be seen as the bundled-clues.

$$\left.\begin{array}{l} f_1(c_1, c_2, ..., c_n) = r_1 \\ f_2(c_1, c_2, ..., c_n) = r_2 \\ \cdots \\ f_n(c_1, c_2, ..., c_n) = r_n \end{array}\right\} \Rightarrow R = g(r_1, r_2, ...r_n) \tag{1}$$

$$\overbrace{\left.\begin{array}{l} f_{11}(C_n) = r_{11} \\ f_{12}(C_n) = r_{12} \\ \cdots \\ f_{1n}(C_n) = r_{1n} \end{array}\right\}}^{\text{step 1}} \Rightarrow \overbrace{\left.\begin{array}{l} f_{21}(R_{1n}) = r_{21} \\ f_{22}(R_{1n}) = r_{22} \\ \cdots \\ f_{2n}(R_{1n}) = r_{2n} \end{array}\right\}}^{\text{step 2}} \Rightarrow ... \Rightarrow \overbrace{\left.\begin{array}{l} f_{m1}(R_{m-1\,n}) = r_{m1} \\ f_{m2}(R_{m-1\,n}) = r_{m2} \\ \cdots \\ f_{mn}(R_{m-1\,n}) = r_{mn} \end{array}\right\}}^{\text{step m}}$$
$$C_n = c_1, c_2, ..., c_n \qquad R_{1n} = r_{11}, r_{12}, ..., r_{1n} \qquad R_{m-1\,n} = r_{m-1\,1}, r_{m-1\,2}, ..., r_{m-1\,n}$$
$$\Rightarrow \underbrace{R_{final} = g(R_{mn})}_{R_{mn} = r_{m1}, r_{m2}, ..., r_{mn}} \tag{2}$$

2.1. Knotted-clues method

In this paper, we adopt the calculation of correlation coefficients as the first step, the hierarchical clustering technique as the function of the second step, and the frequency items pattern mining method as the last step. Compared to using the three methods separately, our approach aims at the characteristics of the problem and refines the accuracy of the result.

2.1.1. Correlation coefficient

The correlation coefficients are the measurement of linear correlation between the variables, of which the most famous one is,*Pearsoncorrelationcoefficient*[26] as shown in Formula (3), where X, Y are variables, $cor(X, Y)$ stands for the *Pearsoncorrelationcoefficient* of X and Y, $cov(X, Y)$ represents the covariance of X and Y, $var(X)$ is the variance of X and $E(X)$ means the mathematical expectation of X. If $|cor(X, Y)|$ is close to 1, the correlation of X and Y is high. Conversely, if the value is close to 0, the linear correlation is low.

$$cor(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{3}$$

2.1.2. Hierarchical clustering

Hierarchical Clustering treats every variable as a single cluster, and merges the two closest clusters into one.[27] The specific steps are shown in Table 1. The key point of this technology is the distance calculation in step (2). There are seven common algorithms to implement the clustering, "complete", "ward", "single", "average", "mcquitty", "median" and "centroid". Each algorithm has different emphases on distance calculation. If the algorithm is selected individually, we might choose "ward", "average", "mcquitty", "median" or "centroid", which cares more about the whole or average level of the elements in the cluster. But here, due to the data interpretation, "complete" is selected, in which the calculation focuses on the farthest two elements in each cluster. And it will be discussed in Section 2.2.

2.1.3. Frequency items patterns mining

Frequency items patterns mining is an important technology of data mining, which reflects the interdependence and relevance between objects. Frequent items mining algorithms have been mature, of which the most well-known is the Aprior algorithm.[28] To illustrate the algorithm, some basic definitions should be explained here. The minimum support threshold, in simply words, is a manually set value. If the occurrence probabilities of the items are greater than this value, we call them the frequency items. If an item contains k elements and satisfies the minimum support threshold, it is called the frequent K itemsets, written as L_k. The support and confidence of a frequency item are shown as Formula (4) and (5). The basic steps of Aprior are shown in Table 2. For example, there are 4 items in initial data $\{\{a, b, c, d\}, \{a, b, c\}, \{a, b, e\}, \{a, f\}\}$, and the minimum support threshold is 0.5. Then we can get $L_1 = \{\{a\}(support = 1), \{b\}(support = 0.75), \{c\}(support = 0.5)\}$, and $L_3 = \{\{a, b, c\}(support = 0.5, confidence((a, b) \Rightarrow c) = 0.67)\}$.

Table 1
Basic steps of hierarchical clustering.

steps	explanation
(1)	(initialization) define each variable as a cluster
(2)	calculate the distance between each two clusters
(3)	find the closest two clusters and combine them into one
(4)	repeat step 2 and step 3 until all variables fall into one cluster

Table 2
Basic steps of Aprior algorithm.

steps	explanation
(1)	set the minimum support threshold according to the problem
(2)	generate L_1 from initial data
(3)	generate L_k from L_{k-1} until there is only one itemset in L_k

$$support(X) = \frac{the\ number\ of\ occurrences\ of\ frequency\ item\ X}{the\ number\ of\ all\ initial\ items} \tag{4}$$

$$confidennce(X \Rightarrow Y) = \frac{support(X, Y)}{support(X)} \tag{5}$$

2.2. Data interpretation

In the knotted-clues method, to simplify the problem, an internal relationship between the steps is formed by the specific algorithms, which reflects our ideas for solving the problem, and is called the data interpretation. In this work, our goal is to simplify the problem and refine the results. Assuming that each clue is a variable which is represented by an n-dimensional vector. There are m($m \ll n$) variables, and the size of the initial data is $n*m$. In the first step, we screen the relationships of the variables preliminary. Through the calculation of Pearson correlation coefficient, two vectors are converted to a single value. At the end of the first step, the data become $m*m$ size, which contains the correlations between every two variables. After removing duplicate numbers and the self correlations, the computationally significant size is $\frac{m(m-1)}{2}$.

In the second step, the $\frac{m(m-1)}{2}$ data is used as an input for hierarchical clustering, which is supposed to be m $(m-2)$-dimensional vectors compared with each other. There is a toy example shown in Table 3, in which (a) displays the correlation coefficients of 5 variables, (b) demonstrates the distance calculation of (a) and (c) gives the clustering dendrogram. The measurement of distance plays a vital important role. Take the distance of A and D for example. To give full consideration of closeness, we compare the other m-2 coefficients of two variables one by one, and calculate the differences. To B, C and E, the coefficients of A are (0.97,0.91,0.79), while those of D are (0.88,0.96,0.98), and the differences here are (0.09,0.05,0.19). In the crime analysis, the similarity of crime incidents cares more about overall stability rather than individual difference. For B and C, A seems to be similar with D. However, for E, A and D differ widely, just because of which we can come to conclusion that A and D cannot be fit in one cluster directly. As the input data here is the correlations, a single value represents the overall relevance of the two variables, which determines that the deviation of one value results in the difference of distance between the two variables. So we choose the maximum distance as the measurement of two variables, as shown in Formula (6), which is called

Table 3
An example of hierarchical clustering.

(a) input data

A				
0.97	B			
0.91	0.97	C		
0.82	0.88	0.96	D	
0.79	0.88	0.96	0.98	E

(b) distance calculation

	C	D	E
A,B	0.06	0.06	0.09
	B	D	E
A,C	0	0.14	0.17
...	...	...	...
	A	B	C
D,E	0.03	0	0

(c) clustering dendrogram

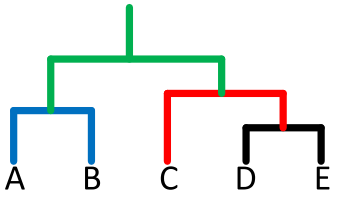


Fig. 2. District data summary.

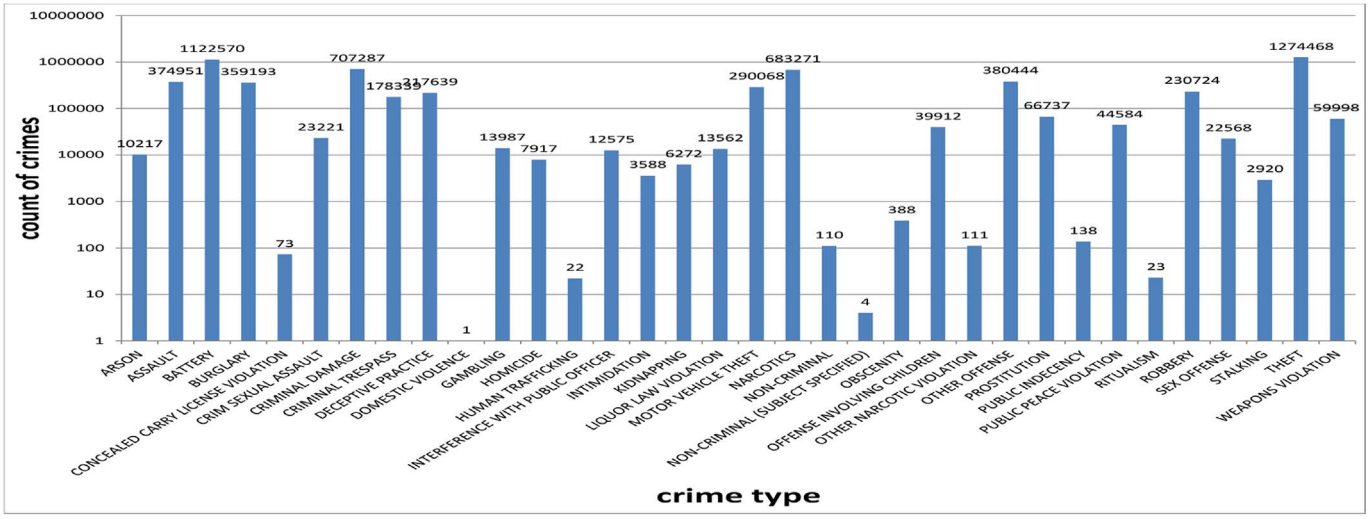
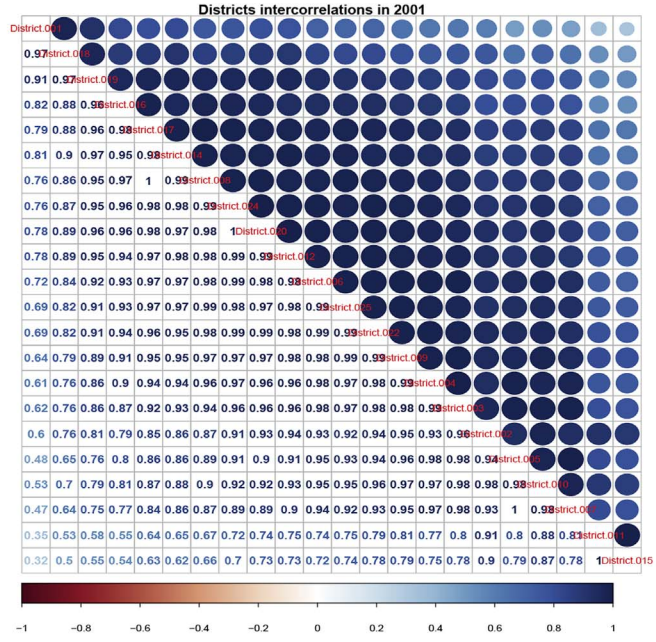


Fig. 3. Crime type data summary.

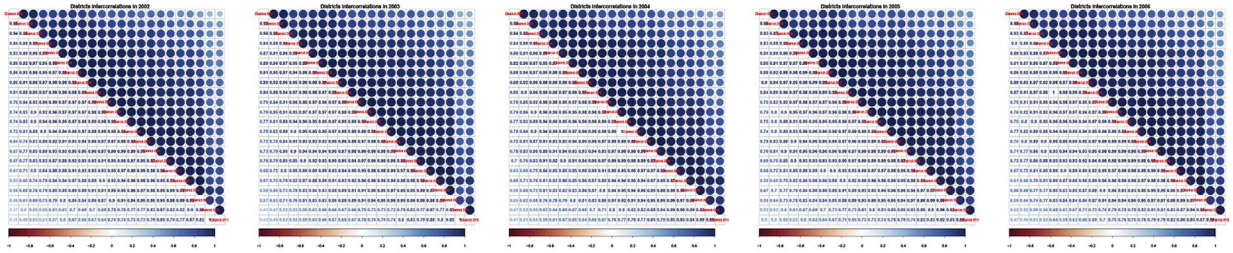
"complete" algorithm in Section 2.1. In the first round, we put D and E into a cluster. Then, A and B are selected into one group. Next, C is assigned into the (D,E) cluster. At last, both (A,B) and (C,D,E) integrate together, which is the end of clustering process. If we set the number of clusters as two, we get the result $\{\{A, B\}, \{C, D, E\}\}$. This result can be already called a pattern, though a little random. In order to make the results more stable, we use the next step.

$$Distance(X, Y) = \max_{x_i \in X, y_i \in Y} Distance(x_i, y_i) \quad (6)$$

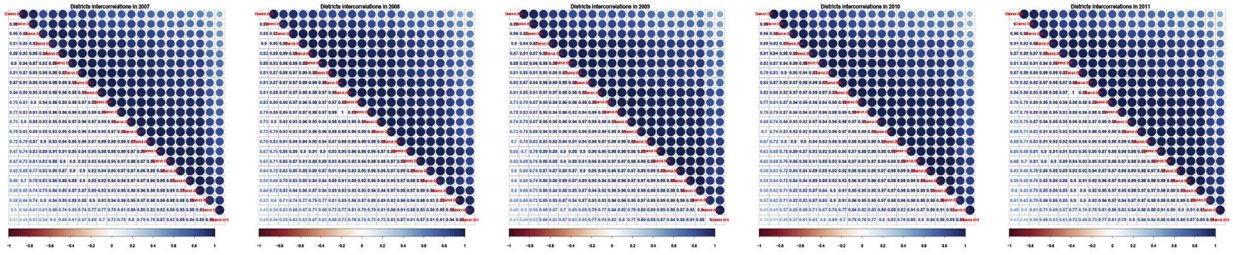
For sake of making the results more convincing, we divide the data into groups. In each group, we repeat the above two steps and get a pattern. Then we collect the patterns as the input data for this step and mine the frequency items. In all or most of the groups, the frequency patterns are the final result. Continuing with the example above, we divide the initial data into 3 groups by year and get 3 patterns $\{\{A, B\}, \{C, D, E\}\}$, $\{\{A, B\}, \{C, D\}, \{E\}\}$ and $\{\{A\}, \{B, C, D, E\}\}$. With minimum support threshold 0.67, we obtain the final pattern $\{\{A, B\}, \{C, D\}\}$ and make a conclusion that A and B are close to each other in data interpretation, which means the probability of occurrence



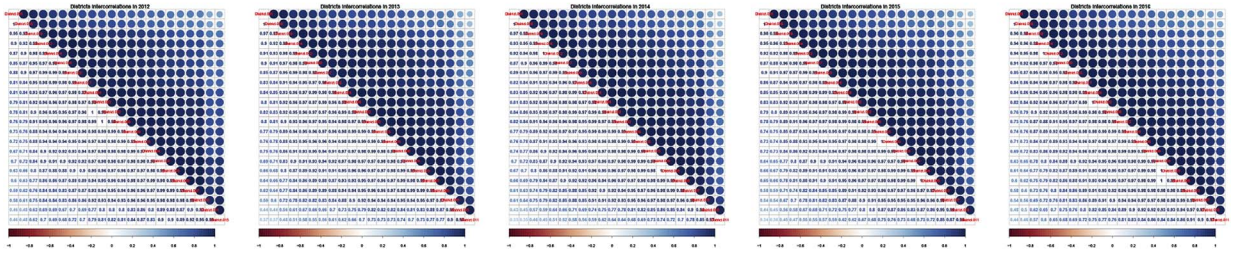
(a) 2001



(b) 2002 (c) 2003 (d) 2004 (e) 2005 (f) 2006

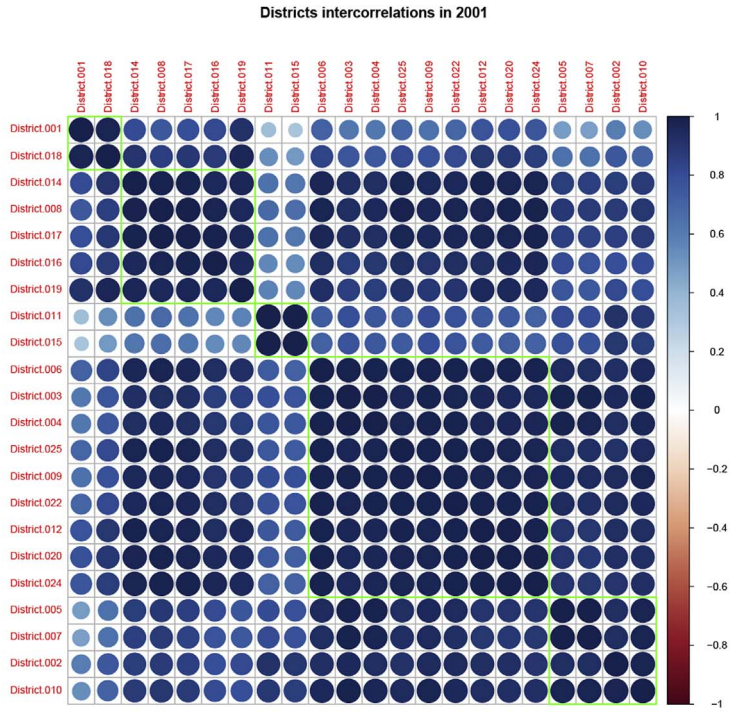


(g) 2007 (h) 2008 (i) 2009 (j) 2010 (k) 2011

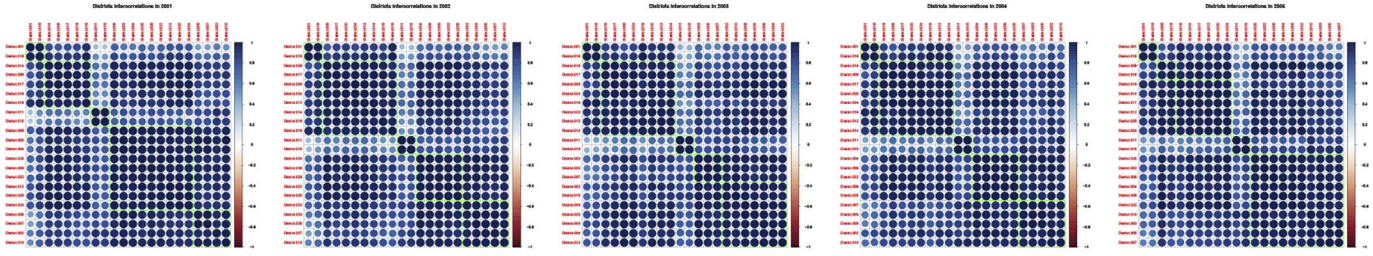


(l) 2012 (m) 2013 (n) 2014 (o) 2015 (p) 2016

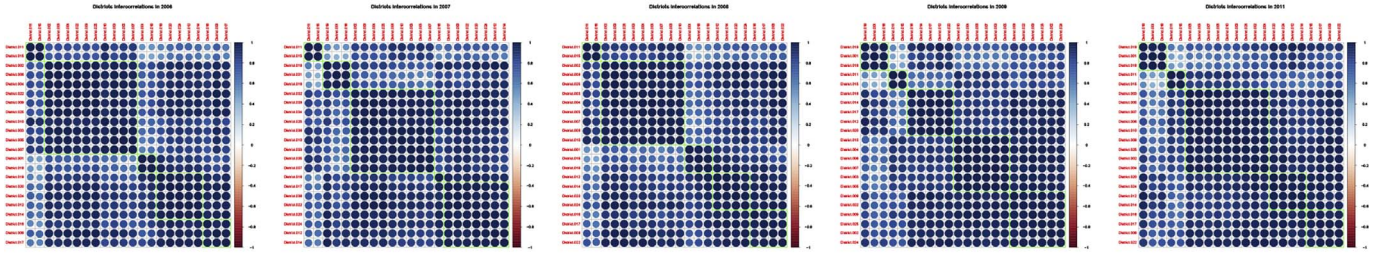
Fig. 4. Districts correlation coefficients of Chicago in 2001–2016.



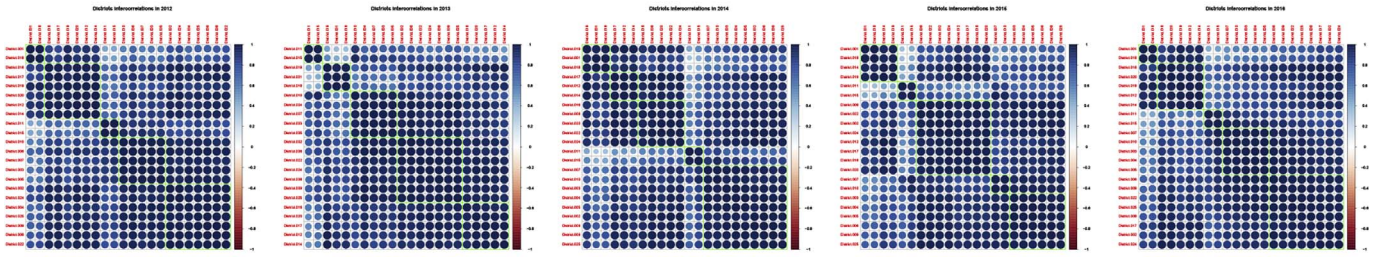
(a) 2001



(b) 2002 (c) 2003 (d) 2004 (e) 2005 (f) 2006



(g) 2007 (h) 2008 (i) 2009 (j) 2010 (k) 2011



(l) 2012 (m) 2013 (n) 2014 (o) 2015 (p) 2016

Fig. 5. Districts hierarchical clusters of Chicago in 2001–2016.

Table 4
Association rules for over 14 years.

{5,7,10}	{3,5,7}	{6,9,25}
{10,5} ⇒ {7}, *conf*: 1.0	{3,7} ⇒ {5}, *conf*: 1.0	{9,6} ⇒ {25}, *conf*: 1.0
{10,7} ⇒ {5}, *conf*: 0.933	{3,5} ⇒ {7}, *conf*: 0.933	{25,6} ⇒ {9}, *conf*: 1.0
{5,7} ⇒ {10}, *conf*: 0.933	{5,7} ⇒ {3}, *conf*: 0.933	{9,25} ⇒ {6}, *conf*: 0.875

of near-repeat phenomenon between them is higher than those between them and others. C and D can reach the same conclusion.

In the data interpretation, we specially arrange the order of the three steps and algorithms for implementation. Though focusing on the relevance during the whole process, each step has its own characteristics as well. In the first step, we make a preliminary calculation and simplified the data. In the second step, we rigorously compare the variables and carry out patterns mining. In the last step, the results are strengthened and refined. The algorithms for each step may not be the most efficient one, but this integration arrangement brings better final result, which is the core of the data interpretation.

3. Experiments and results

There is a famous open source dataset in the field of crime data analysis, which is provided by the Chicago Police Department[2] and records millions of reported incidents of crime that occurred in the City of Chicago from 2001 to present. Our experimental Data is extracted from January 1, 2001 to August 26, 2016, with 6,147,883 records and 22 features. Our experiment has two objectives: one is the nearby areas, which concentrates on refining adjacent locations; The other is the near-repeat effect in the crime types, which pays attention to the correlation patterns of various crime types.

3.1. Experimental data analysis

We summarize the initial data that contains 26 districts and 34 crime types, as illustrated in Fig. 2 and Fig. 3. Then we eliminate districts 13, 21, 23 and 31, in which the occurrences of crime incidents are less then 120 during the 16 years. In these districts, the records of criminal incidents are too low to be statistically significant and not suitable for the research methods of data science. For the same reason, we preserve 26 crime types without considering the crime categories fewer than 400 occurrences in 16 years. According to the data interpretation of our knotted-clues method, we divide the data into 16 groups by year, from 2001 to 2016. In order to avoid the data being too scattered and sparse, we recount the number of crimes by day and type.

3.2. Experimental results

With the preprocessed data mentioned above, we implement the data interpretative knotted-clues method described in Section 2. In the second step of hierarchical clustering, we set 5 as the cluster number. Each variable must appear in one of the five clusters without repeated. In the frequency patterns mining step, there are 80 input items(16 years × 5 clusters). We hypothesize the frequency requirement here is over 14 years, for which the minimum support threshold is $0.175(\frac{14\ years}{16\ years \times 5\ clusters})$. The confidence threshold in the association rule calculation is 0.9. From the perspective of the two experimental objectives, we analyze the experimental results separately as below.

[2] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Map/c4ep-ee5m.

3.2.1. Fine-grained districts near-repeat patterns

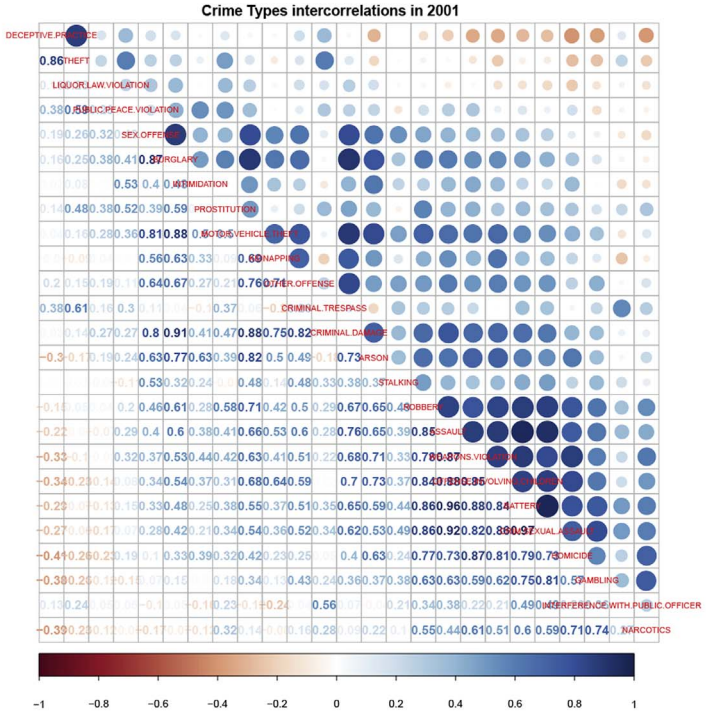
In this experiment, our goal is to find the more accurate near-repeat relationships between districts. The correlation coefficients of the 16 years are demonstrated in Fig. 4. The hierarchical clustering results of the 16 years are shown in Fig. 5.

Here we discuss the interesting patterns, consulting the districts map shown in Fig. 1. There are 3 clusters appear in all the 16 years, {1,18}, {9,25} and {11,15}, where the confidences between elements in each cluster are all 1.0. It is a very strong association rule, which means if there is a repeat crime following District 1, we can directly infer that it occurs in District 18 rather than 2, 9 or 12. More significantly, the geographic distance or network degree of the pattern {9,25} is not so near, which gives us great assurance to infer that there is a strong connection between the two districts in a truly hidden criminal network. We suggest that the police force of the two districts should strengthen their cooperation. The situation is similar in 15 years. There are {3,5}, {5,7} and {7,10}, with the same confidence 0.9375, without directly next to each other.

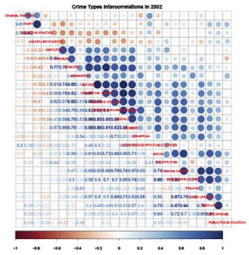
In 14 years, the patterns are more rich and interesting. There are 9 patterns, of which some are in the relationship of subset, {12,14}, {12,20}, {5,10}, {5,7,10}, {3,7}, {3,5,7}, {6,9}, {6,25} and {6,9,25}. Among them, {12,20}, {5,7,10}, {6,9,25} are not in the contiguous districts in the map. From the angle of association rules, the patterns satisfying the confidence requirement are concentrated in groups with more elements, as shown in Table 4. In {3,5,7}, {3,7} plays an absolute leadership role to {5}. With the help of association rule in 15 years ({5} ⇔ {7}, *conf*: 0.9375) and the new rule ({3,5} ⇒ {7}, *conf*: 0.933), we come to the conclusion that {3} has more influence on {5} than on {7}. We can draw similar conclusions that {10} has a greater impact on {7} than on {5}. From {3,5,7} and {5,7,10} , we associate {3,5,7,10} and test our conjecture by checking out the results in 13 years, in which we do find this pattern. In {3,5,7,10}, there are 5 rules of which confidence = 1.0, {3,10} ⇒ {5}, {3,10} ⇒ {7}, {3,10} ⇒ {5,7}, {3,5,10} ⇒ {7} and {3,7,10} ⇒ {5}. This result validates our inference again, in real crime network, {3,10} acts as a trigger for {5,7}. If we restrain the crimes in {3,10}, the near-repeat crime risk will reduce in {5,7}, which is to say {3,10} is the center of the near-repeat cluster {3,5,7,10}. At last, we talk about {6,9,25} with rule ({9} ⇔ {25}, *conf*: 1.0) mentioned above. Among the three, the confidence of one leading the other two is 0.875. In contrast, as illustrated in the last column of Table 4, the confidence of {9,25} ⇒ {6} is lower than the setting value, while those of the others are 1.0. On the three, {9} and {25} are equal in status, but {6} is not very cooperative with the actions of the other two. If the police force want to disrupt the criminal gang {6,9,25}, {6} is the breakthrough.

3.2.2. Crime types near-repeat patterns

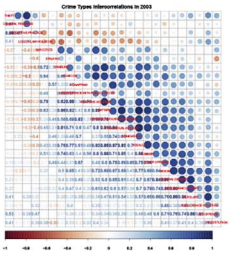
In this experiment, we aim to find near-repeat patterns between crime types. The correlation coefficients of the 16 years are demonstrated in Fig. 6. The hierarchical clustering results of the 16 years are shown in Fig. 7. In all the 16 years, there are some types always in the same cluster, {assault, battery, robbery, weapons violation}(for short P_{16-1}) and {deceptive practice, theft}(for short P_{16-2}). In the 4 elements cluster, the confidences of all rules are 1.0 in both (one ⇔ the rest three) and (one pair ⇔ the other pair). These four types often occur simultaneously in one crime incident and occur frequently, which is an expected result. In the two elements pattern, the rule (deceptive



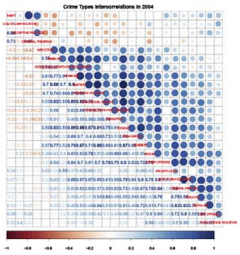
(a) 2001



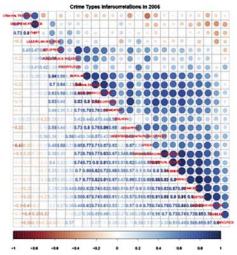
(b) 2002



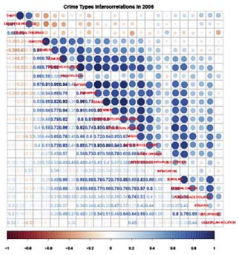
(c) 2003



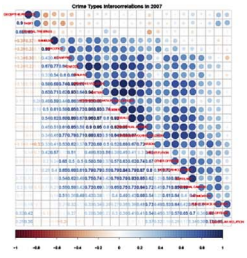
(d) 2004



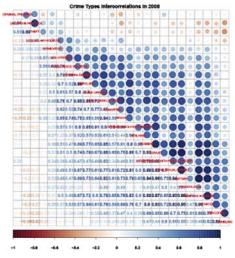
(e) 2005



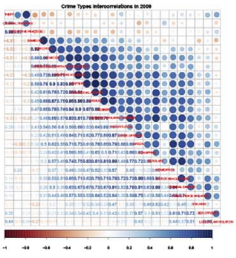
(f) 2006



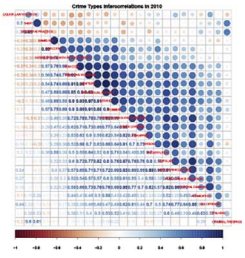
(g) 2007



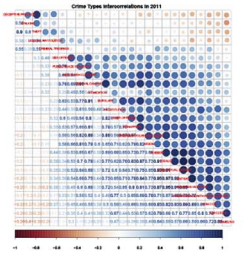
(h) 2008



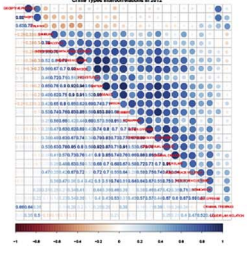
(i) 2009



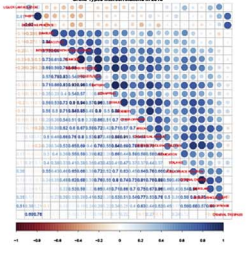
(j) 2010



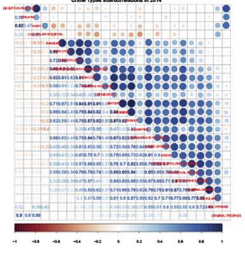
(k) 2011



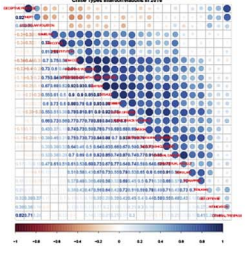
(l) 2012



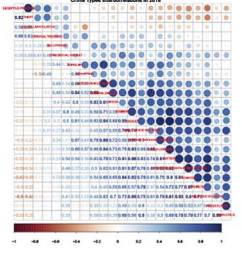
(m) 2013



(n) 2014

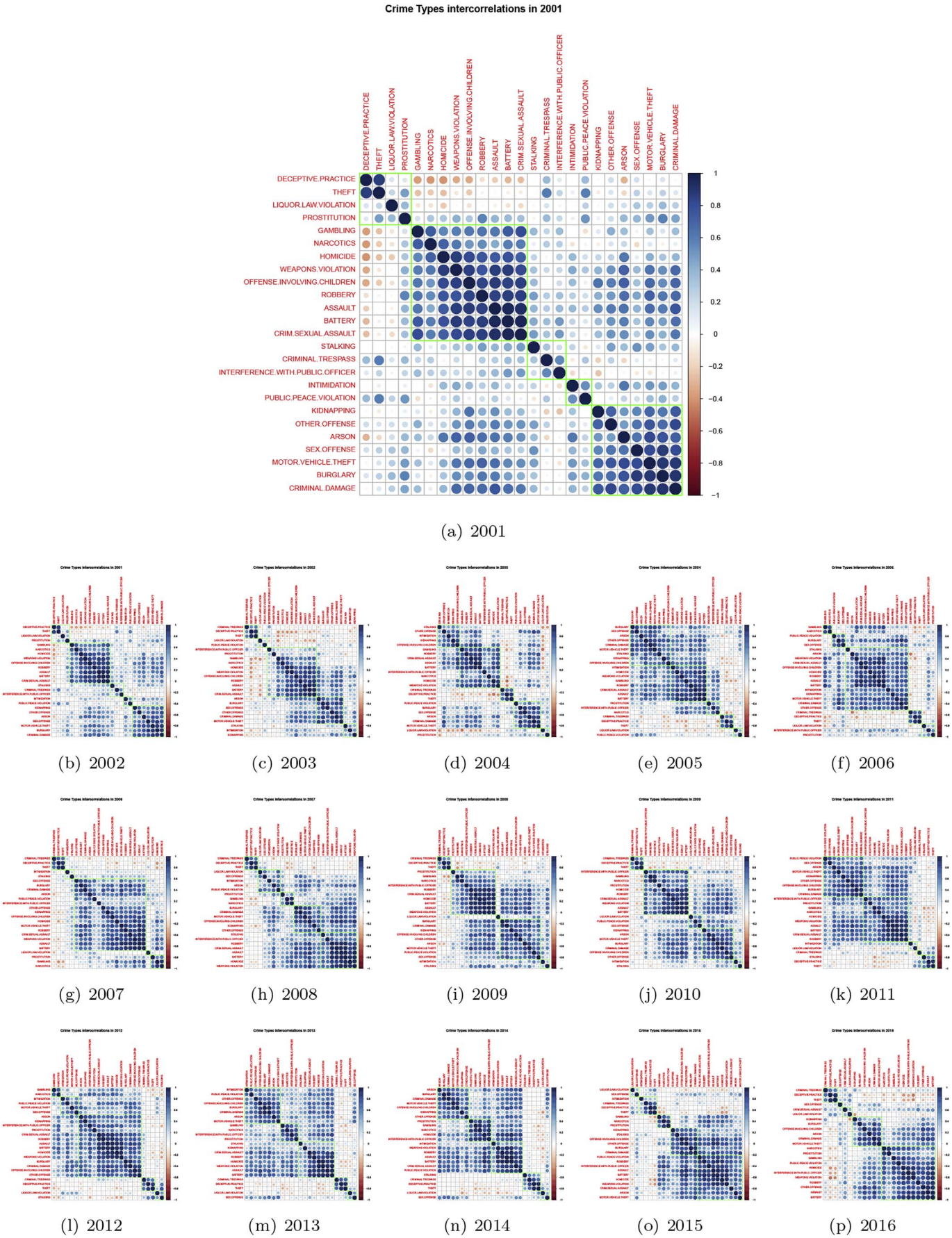


(o) 2015



(p) 2016

Fig. 6. Crime types correlation coefficients of Chicago in 2001–2016.



(a) 2001

(b) 2002 (c) 2003 (d) 2004 (e) 2005 (f) 2006

(g) 2007 (h) 2008 (i) 2009 (j) 2010 (k) 2011

(l) 2012 (m) 2013 (n) 2014 (o) 2015 (p) 2016

Fig. 7. Crime types hierarchical clusters of Chicago in 2001–2016.

practice ⇔ theft, conf:1.0) reveals the two types are near by each other in a data relation, which maybe bring us a new perspective on the relationship of both the types.

Then we look into the patterns in 15 years. Simple rules (burglary ⇔ criminal damage, conf:0.9375) and (gambling ⇔ narcotics, conf:0.9375) tells us that the two types in one pattern are closely associated with each other and always occur together or successively in nearby places. This situation should be taken into account when deploying police force. There are also patterns seem to be complex, such as {assault, battery, homicide, robbery, weapons violation} and {assault, battery, crim sexual assault, robbery, weapons violation}. After analyzing, we find that both patterns are actually a combination, which integrate a single type X (homicide or crim sexual assault) into the above pattern P_{16-1}. And the rules are in 3 categories, ($X \Leftrightarrow P_{16-1}$,conf:0.9375), ($X + Y \Rightarrow P_{16-1} - Y$,conf:1.0) and ($Y \Rightarrow X + P_{16-1} - Y$,conf:0.9375), where Y is a nonvoid proper subset of P_{16-1}. Unlike the cooperations of crimes in districts, the relationships of crime types are not easy to make clear. Because of environmental factors and the emotional state of criminals, this relationship is a hierarchical deepening process. For instance, in an armed robbery, assault and battery are common followed. If the offender is stimulated at this point, he will further act aggressively, resulting in homicide or crim sexual assault. A thorough study of crimes without serious consequences will help us minimize the damage and find a good way to protect the victims.

In the 14-years-patterns, there are 4 two-elements-clusters, {arson, criminal damage}, {arson, motor vehicle theft}, {criminal damage, motor vehicle theft} and {deceptive practice, narcotics}, from which the confidences of the rules are all 0.875. Although below the threshold we set, the proximities between these types are still of reference value. Next, we come to the patterns in conformity with the setting confidence. From {criminal trespass, deceptive practice, theft}, two rules are extracted, ({criminal trespass, deceptive practice} ⇒ theft, conf:1.0) and ({criminal trespass, theft} ⇒ deceptive practice, conf:1.0), while the rest confidences are 0.875. Taking into account the previous rules in P_{16-2}, we arrive at a conclusion that criminal trespass is an additional sufficient condition for rule ($S \Rightarrow P_{16-2} - S$), where S is a nonvoid proper subset of P_{16-2}. In simple word, it reinforces our inference about deceptive practice that a theft and a criminal trespass occur simultaneously. But it is a a probabilistic problem when we infer the criminal trespass from deceptive practice or theft. At last, here is the pattern {assault, battery, crim sexual assault, homicide, robbery, weapons violation}, which can be obtained from P_{16-1} and {homicide, crim sexual assault}(for short P_{deep}). According to P_{deep}, we divide the rules into three categories, ($P_{deep} + Y \Rightarrow P_{16-1} - Y$,conf:1.0), ($Y \Rightarrow P_{deep} + P_{16-1} - Y$,conf:0.875) and ($D + Y \Rightarrow P_{deep} - D + P_{16-1} - Y$,conf:1.0), where Y is a proper subset of P_{16-1} and D is a nonvoid proper subset of P_{deep}. Here we interpret P_{deep} a deeper crime collection of the crime types. If all the types in P_{deep} occur, we can obtain definite results, as rules in P_{16-1}. Just the opposite, if we want to infer all types in P_{deep}, the accuracy of inference is relatively low. In the middle, if we speculate about part of P_{deep} from the rest types, the probability improve.

4. Discussions

For the fine-grained districts near-repeat patterns, the selection of the number of years is worth being discussed. This number is negatively related to the number of patterns. The more years, the less patterns. These patterns are stable but not good for detecting new trends. The patterns in few years contain more new information as well as more noise. A future research direction is to find the subset relations between the patterns in most years and a few years, analyze the cascade relations and discriminate the noise and new trends.

When it comes to the types of crime, it is a man-made division by researchers. Crime is a complex human activity. Perpetrators may not understand any legal knowledge, let alone the boundaries of the crime types. In order to achieve his goal, he may use various criminal means. These lead to the one sidedness and complexity of crime type analysis. So we make the comprehensive analysis of the crime types, and discover the nearby relations among them. Through the discussions of frequency patterns and association rules, we analyze the correlation between the types. Each pattern can be seen as a big crime category, in which the specific crime types should not be treated differently. In this situation, the neat-repeat phenomenon is more obvious in a big category rather than between the categories. A future work is to mine the independent factors of the crime types, which may draw on the experience of the latent variables and feature selection.

5. Conclusions

In this paper, we propose a knotted-clues method to obtain fine-grained results of the near-repeat phenomenon both in districts and in various crime types. In the view of data interpretation, we combine correlation coefficient, hierarchical clustering and frequency patterns mining in a particular order. In districts, we refine the results to specific district rather than the near range. The accuracy results may help us identify the distribution of criminal forces in real crime networks. In police deployment, it can better coordinate the allocation of resources and strengthen cooperations. When combating crime, we can get the criminal centers or sources through near-repeat fine-grained analysis, so as to concentrate resources and improve efficiency. Through our approach, we find the associate patterns of different crime types and analyze the hierarchical relationships between the patterns. In actual actions, we should try our best to avoid the occurrence of crime types in the deeper pattern, in order to protect the lives and properties of the victims.

Acknowledgements

This research has been supported by National Natural Science Foundation of China (No.61572514), (No.61379117), (No.61379145) and the Joint Funds of CETC (No.20166141B08020101).

References

1. Comeau M, Klofas J. *Repeat and Near-repeat Burglary Victimization in Rochester, Ny.* 2014; 2014 literature review: Motivations to commit burglary and target selection.
2. Johnson SD. Repeat burglary victimisation: a tale of two theories. *J Exp Criminol.* 2008;4(3):215–240.
3. Short MB, DOrsogna MR, Brantingham PJ, Tita GE. Measuring and modeling repeat and near-repeat burglary effects. *J Quant Criminol.* 2009;25(3):325–339.
4. Bowers KJ, Johnson SD. Who commits near repeats? a test of the boost explanation. *West Criminol Rev.* 2004;5(3):12–24.
5. Adepeju M. Investigating the repeat and near-repeat patterns in sub-categories of burglary crime. *International Conference on Geocomputation.* 2017; 2017.
6. Wang Z, Liu X. Analysis of burglary hot spots and near-repeat victimization in a large Chinese city. *Int J Geo-Inf.* 2017:148.
7. Haberman CP, Ratcliffe JH. The predictive policing challenges of near repeat armed street robberies. *Policing.* 2012;6(2):151–166.
8. Ratcliffe JH, Rengert GF. Near-repeat patterns in philadelphia shootings. *Secur J.* 2008;21(1-2):58–76.
9. Sturup J, Rostami A, Gerell M, Sandholm A. Near-repeat shootings in contemporary Sweden 2011 to 2015. *Secur J.* 2017:1–20.
10. Wells W, Wu L, Ye X. Patterns of near-repeat gun assaults in houston. *J Res Crime Delinquen.* 2012;49(2):186–212.
11. Youstin TJ, Nobles MR, Ward JT, Cook CL. Assessing the generalizability of the near repeat phenomenon. *Crim Justice Behav Int J.* 2011;38(10):1042–1063.
12. Townsley M, Homel R, Chaseling J. Infectious burglaries: a test of the near repeat hypothesis. *Br J Criminol.* 2003;43(3):615–633.
13. Bernasco W. Them again?same-offender involvement in repeat and near repeat burglaries. *Eur J Criminol.* 2008;5(4):411–431.
14. Piza EL, Carter JG. Predicting Initiator and Near Repeat Events in Spatiotemporal Crime Patterns: An Analysis of Residential Burglary and Motor Vehicle Theft, Justice Quarterly.
15. Watalingam RD, Richetelli N, Pelz JB, Speir JA. Eye tracking to evaluate evidence recognition in crime scene investigations. *Forensic Sci Int.* 2017;280:64–80.
16. Aquila I, Pepe F, Nunzio CD, Ausania F, Serra A, Ricci P. Suicide case due to phosphoric acid ingestion: case report and review of literature. *J Forensic Sci.*

2014;59(6):1665–1667.
17. Aquila I, Gratteri S, Sacco MA, Ricci P. The role of forensic botany in solving a case: scientific evidence on the falsification of a crime scene. *J Forensic Sci.* 2017(3):137–140.
18. Gratteri S, Ricci P, Tarzia P, Fineschi V, Sacco MA, Aquila I. When a suicide becomes a forensic enigma: the role of hanging marks and tools of suspension. *Med Leg J.* 2017;85(3):141–144.
19. Baechler S, Glinas A, Tremblay R, Lu K, Crispino F. Smartphone and tablet applications for crime scene investigation: state of the art, typology, and assessment criteria. *J Forensic Sci.* 2017;62(4):1043–1053.
20. Johnson D. The space/time behaviour of dwelling burglars: finding near repeat patterns in serial offender data. *Appl Geogr.* 2013;41(4):139–146.
21. Johnson SD, Bowers KJ. *Near Repeats and Crime Forecasting.* Springer New York; 2014.
22. Kump P, Alonso DH, Yang Y, Candella J, Lewin J, Wernick MN. Measurement of repeat effects in chicagos criminal social network. *Appl Comput Inf.* 2016;12(2):154–160.
23. Murray AT, Grubesic TH. *Exploring Spatial Patterns of Crime Using Non-hierarchical Cluster Analysis.* Springer Netherlands; 2013.
24. Karypis George, Han E (Sam), Kumar Vipin. Chameleon: hierarchical clustering using dynamic modeling. *Computer.* 2002;32(8):68–75.
25. Liu X, Wang L, Zhang J, Yin J, Liu H. Global and local structure preservation for feature selection. *IEEE Trans Neural Networks & Learn Syst.* 2017;25(6):1083–1095.
26. Williams S. Pearson's correlation coefficient. *N Z Med J.* 1996;109(1015):38.
27. Aggarwal S, Phoghat P, Maitrey S. Hierarchical clustering- an efficient technique of data mining for handling voluminous data. 2015;129:31–36.
28. Sun D, Teng S, Zhang W, Zhu H. An algorithm to improve the effectiveness of apriori. *IEEE International Conference on Cognitive Informatics.* 2007; 2007:385–390.