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

[a] College of Computer, National University of Defense Technology, Changsha, China
[b] Department of Electronic Information and Electrical Engineering, Changsha University, 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/

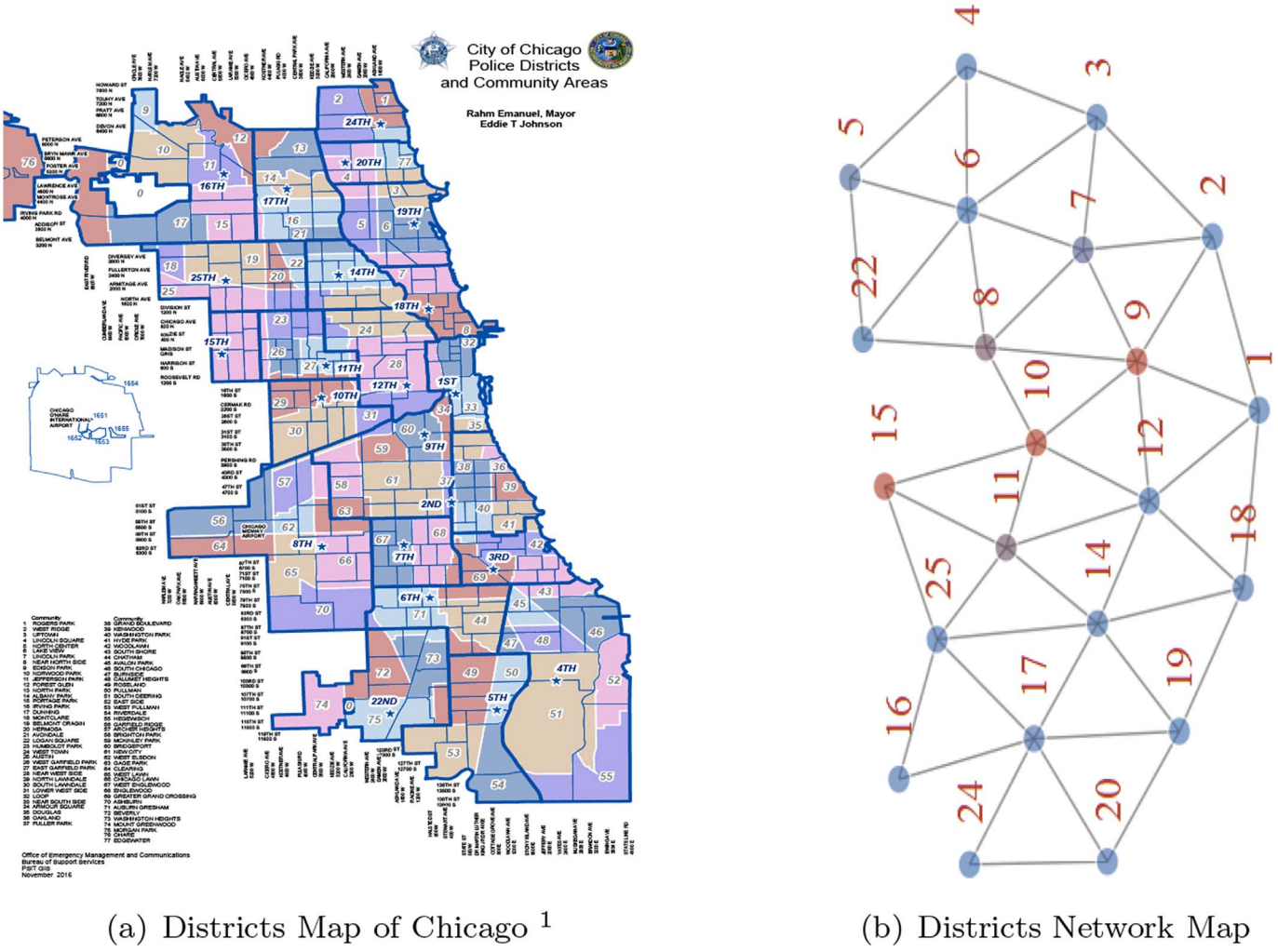

Fig. 1. Districts in Chicago. (a) Districts map of Chicago .[1] (b) Districts network map.

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

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