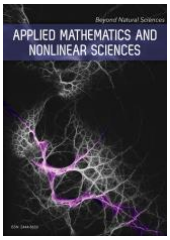

Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Research on the Application of Social Network Data Mining Technology in Crime Analysis and Prevention

Wan Duan Xiong[1,2,†], Wang Yu Chong[1,2]

1. China People's Police University (Guangzhou), Guangdong Province, Guangzhou City,399 Aoti South Road, Tianhe District, 510660.
2. China People's Police University, Hebei Province, Langfang City, 220 Xichang Road, Anci District, 065000.

Submission Info

Communicated by Z. Sabir
Received March 29, 2024
Accepted June 20, 2024
Available online July 10, 2024

Abstract

Social network data mining (SNDM) technology shows great application potential in crime analysis and prevention. This study focuses on revealing the characteristics, laws, and trends of criminal behavior through an in-depth analysis of criminal information in social networks. Using data mining techniques such as association rule mining, cluster analysis, and community discovery, the key information and organizational structure of criminal networks are successfully mined, which provides a powerful means of investigation and prevention for public security departments. It is found that important criminal clues are hidden in the user communication data in social networks, and the communication mode and hidden information between criminals can be revealed through association rule mining technology. Cluster analysis helps to identify gangs and hot spots with similar criminal behaviors, which provides important clues for further investigation. In addition, community discovery technology further reveals the internal structure and membership relationship of criminal gangs, which is helpful in deeply understanding the operation mode of criminal organizations and the spread path of criminal acts. Based on historical data and mining results, this study also constructs a crime trend prediction model, which provides timely early warning information for public security departments and helps to take measures to prevent and crack down on criminal acts in advance. On the whole, this study not only enriches the application theory of SNDM technology in the field of crime analysis but also provides new ideas and tools for actual crime investigation and prevention.

Keywords: Social Network; Data Mining; Crime Analysis; Prevention.
AMS 2020 codes: 62B05

†Corresponding author.
Email address: 1186921095@qq.com

ISSN 2444-8656
https://doi.org/10.2478/amns-2024-1832

1 Introduction

With the rapid development of Internet technology, social networks have become an indispensable part of modern society. People share their lives and exchange information on social networks, forming a huge information network. However, at the same time, social networks have gradually become a platform for some lawless elements to carry out illegal activities, which has brought new challenges to social security [1]. Traditional means of criminal investigation are inadequate in the face of massive social network data. Therefore, how to effectively use these data and mine potential criminal information has become an urgent problem to be solved.

Social network data mining (SNDM) technology, as a new data analysis method, can reveal the laws and characteristics hidden behind the data through in-depth mining of massive data. In the field of crime analysis and prevention, this technology has great application potential. Through SNDM, we can not only discover the social behavior characteristics of criminal suspects but also reveal the organizational structure and activity rules of criminal gangs, providing powerful investigation clues for public security departments [2-3].

This study aims to explore the application of SNDM technology in crime analysis and prevention. Through in-depth mining and analysis of criminal information in social networks, we can more accurately grasp the characteristics, laws, and trends of criminal behavior and provide scientific and effective means of investigation and prevention for public security departments. This not only helps to improve the efficiency of criminal investigation but also reduces the crime rate to a certain extent and improves the overall public security level of society. Therefore, this study has important theoretical and practical significance. Theoretically, it is helpful to improve and develop the theoretical system of SNDM and promote the in-depth application of this technology in the field of crime analysis. From a practical point of view, this study aims to provide a new and effective means of crime investigation and prevention for the public security department to cope with the increasingly complex crime phenomenon and maintain social harmony and stability.

2 Literature review

With the popularity of social networks and the development of big data technology, SNDM has been widely used in many fields. In the field of crime analysis and prevention, SNDM technology has also received extensive attention.

At present, the application of SNDM technology in crime analysis has achieved certain results. For example, literature [4] successfully identified some potential criminal suspects by analyzing the information exchanged by users in social networks and revealed the organizational structure among criminal gangs. This research provides strong support for the application of SNDM in criminal investigations. In addition, literature [5] uses SNDM technology to predict the trend of criminal behavior, which provides valuable early warning information for public security departments.

Some scholars have also actively explored the combination of SNDM and crime analysis. For example, literature [6-7] puts forward a criminal risk assessment model based on user behavior by analyzing user behavior in social networks, and the model has achieved good results in practical application. In addition, literature [8] also studies how to use SNDM technology to track network crime activities, which provides a new idea for combating network crime.

However, although scholars at home and abroad have made some achievements in SNDM and crime analysis, there are still some problems and challenges. First of all, the complexity and mass of social network data have brought great challenges to DM, and how to effectively process and analyze these

data has become an urgent problem [9-10]. Secondly, users' behaviors in social networks are diverse and dynamic, and how to accurately identify criminal behaviors and predict criminal trends is also an important research direction. In addition, with the continuous development of technology, criminals are constantly changing their modus operandi and strategies, so it is necessary to constantly update and improve the DM algorithm to adapt to the new forms of crime.

To sum up, SNDM technology has broad application prospects in crime analysis and prevention. However, there are still some problems and challenges to be solved in the current research. Future research can further explore how to optimize the DM algorithm, improve data processing efficiency, and strengthen interdisciplinary cooperation, to promote the further application of SNDM technology in crime analysis and prevention.

3 Research contents and methods

3.1 Research objects

This study takes the criminal information in social networks as the mining object, aiming at revealing the patterns, characteristics, and trends of criminal behavior through in-depth analysis of these data. Specifically, it will focus on two main aspects: the social behavior of criminal suspects and the organizational structure of criminal gangs.

On social networks, the social behavior of criminal suspects often hides important clues. These behaviors may include the frequency, time, and content of interaction with others, as well as their activity patterns on social networks. By deeply mining these social behavior data, we can better understand the psychological and behavioral characteristics of criminal suspects, and then provide valuable information for crime investigation and prevention [11].

To comprehensively analyze the social behavior of criminal suspects, we study and collect all kinds of data on suspects on social networks, including but not limited to published trends, comments, likes, and private messages. These data provide a multi-dimensional perspective to gain insight into the suspect's social habits, hobbies, interpersonal relationships, and potential criminal motives.

Social network not only provides a platform for individual suspects to hide and communicate but also become an important tool for criminal gangs to organize and plan activities. Therefore, an in-depth analysis of the organizational structure of criminal gangs on social networks is of great significance for revealing their operation mode and cracking down on crime [12].

This study uses SNDM technology to identify and analyze the relationship network among members of criminal gangs. This includes determining the contact information, frequency, and hierarchical relationship among gang members. In addition, we also pay attention to the role orientation of gang members on social networks, such as leaders, executors, information transmitters, etc., to reveal the division of labor and cooperation mode within the gang.

3.2 DM technology

In this study, a variety of DM technologies are used to deeply analyze the criminal information in social networks (Figure 1). In this study, DM techniques such as association rule mining, cluster analysis, and community discovery are comprehensively used to deeply analyze criminal information in social networks. These methods have high applicability in social network crime information mining

and are expected to provide valuable clues and insights for crime investigation and prevention [13-14].

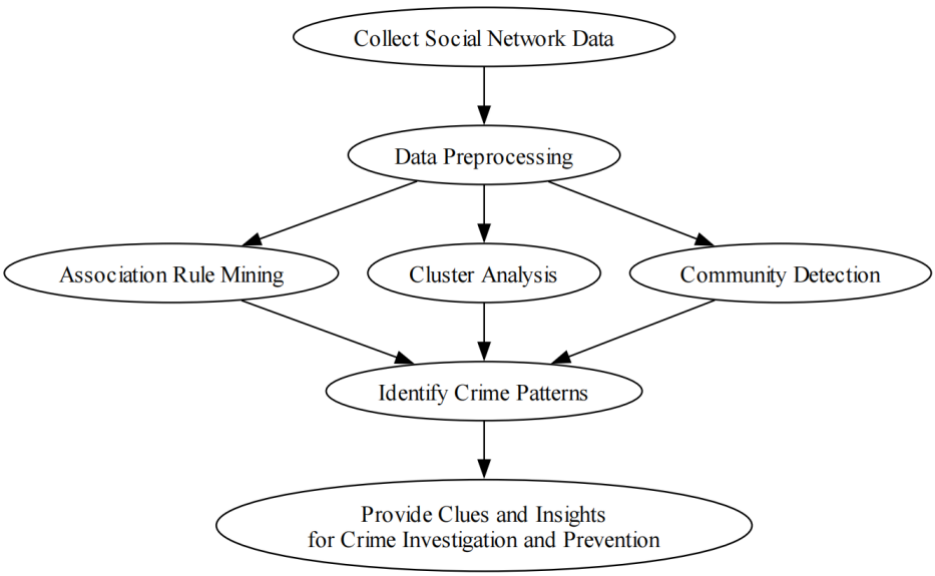

Figure 1. DM model

Association rule mining is a DM technology for discovering interesting relationships between itemsets in large data sets. In social network crime information mining, association rules mining can help to find hidden connections and behavior patterns between criminal suspects or criminal gangs.

In this paper, the Apriori algorithm is used to mine association rules. The algorithm generates association rules by finding frequent itemsets [15]. Frequent itemsets refer to the combination of items that frequently appear in data sets. In this study, items can be users, groups, keywords, and so on in social networks.

The key steps of the Apriori algorithm are as follows:

1) Scan the database, calculate the frequency of each item, collect items that meet the minimum support, and form a set of frequent 1-item sets.

2) Frequent k- itemsets are used to generate candidate $(k+1)$- itemsets, and their support degrees are calculated to find out the candidate sets that meet the minimum support degree, that is, frequent $(k+1)$- itemsets are obtained.

3) Repeat the above steps until no new frequent itemsets can be generated.

Among them, support and confidence are two important indicators to measure association rules. Support indicates the frequency of itemsets in all transactions, and the calculation formula is:

$$\text{Support}\ (X \Rightarrow Y) = \frac{\text{Freq}\ (X \cup Y)}{N} \tag{1}$$

Where $\text{Freq}\ (X \cup Y)$ is the number of transactions containing item set $X \cup Y$, and N is the total number of transactions.

Confidence indicates the conditional probability that the transaction containing X also contains Y, and the calculation formula is:

$$\text{Confidence } (X \Rightarrow Y) = \frac{\text{Support } (X \cup Y)}{\text{Support } (X)} \tag{2}$$

By setting the appropriate minimum support and minimum confidence threshold, the strong association rules are screened out, to find the hidden connections in criminal information.

Cluster analysis is an unsupervised learning method, which is used to group objects in data sets, so that the similarity of objects in the same group is high, while the similarity of objects between different groups is low. In social network crime information mining, cluster analysis can help identify criminal suspects or gangs with similar behavior patterns. In this study, the K-means clustering algorithm is used for clustering analysis. The k-means algorithm divides data into k clusters by iteratively optimizing the center point of each cluster.

The centroid is the average of all points in each cluster. For the k -th cluster, the formula for calculating the centroid C_k is as follows:

$$C_k = \left(\frac{1}{|S_k|}\sum_{x_i \in S_k} x_i\right) \tag{3}$$

Where S_k represents the set of points in the k cluster, x_i is the point in the set and $|S_k|$ is the number of points in the cluster.

Euclidean distance is used to measure the similarity between data points in K-means algorithm. For two n-dimensional vectors a, b, the Euclidean distance $d(a, b)$ between them is calculated as follows:

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{4}$$

Where a_i, b_i is the i -th component of the vector a, b, respectively.

The goal of K-means algorithm is to minimize the sum of squares of distances from all data points to the centroid of the cluster to which they belong. This objective function can be expressed as:

$$J = \sum_{k=1}^{K} \sum_{x_i \in S_k} \|x_i - C_k\|^2 \tag{5}$$

Where K is the number of clusters, S_k is the set of points in the k cluster, x_i is the point in the set, C_k is the centroid of the k cluster, and $\|x_i - C_k\|$ represents the Euclidean distance from the point x_i to the centroid C_k.

Community discovery is an important task in social network analysis, which aims to identify closely related node groups in the network. In social network crime information mining, community discovery can help reveal the organizational structure and membership relationship of criminal gangs.

In this study, community discovery algorithms based on modularity optimization, such as Louvain algorithm, are adopted. The algorithm maximizes the modularity of the whole network by iteratively merging communities. Modularity is an index to measure the quality of community division, and its calculation formula is:

$$Q = \frac{1}{2m}\sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m}\right) \delta(c_i, c_j) \tag{6}$$

Where, A_{ij} is the adjacency matrix element of the network, k_i, k_j is the degree of node i, j, m is the number of edges in the network, c_i, c_j is the community to which node i, j belongs, and $\delta(c_i, c_j)$ is 1 in $c_i = c_j$, otherwise it is 0.

3.3 Data source and processing

The data sources of this study mainly focus on the online fraud case records of the public security department and the user exchange data on social platforms. The research cooperated closely with the Anti-Fraud Center of Z Public Security Bureau and obtained detailed records of all online fraud cases in the past year. These records include the basic information of the victim, the amount cheated, the means of fraud, the contact information of the liar, the trading account and other key information. These data provide us with real cases of online fraud, as well as the modus operandi and characteristics of criminals. Given social platforms that frequently appear in online fraud cases, such as WeChat and QQ, the study negotiated with the platform operators and obtained some group chat records and personal data of users suspected of fraud. These data include text chat, pictures, link sharing, etc., which reflect the activity track and deception methods of fraudsters on social platforms.

To ensure the accuracy and validity of the data, data preprocessing is carried out to remove duplicate records: for example, the same fraud case may be reported many times at different times, and these duplicate information needs to be identified and deleted. Eliminate irrelevant information: content unrelated to fraud, such as advertisements and spam, is cleaned up. Correct the wrong data: such as the wrong reporting time and amount, etc., according to the verification results of the public security department. The report time and chat record time are converted into the format of "YYYY-MM-DD HH:MM: SS" to facilitate the subsequent time series analysis. Convert the text content in chat records into numerical vectors for text mining and sentiment analysis.

Fraud cases can be associated with specific social platform accounts through the victim's report information and chat records on social platforms. For example, the liar WeChat account provided by the victim can help us find the corresponding fraud in the WeChat chat record. Combining the case data of the public security department and the user behavior data on the social platform, a comprehensive data set is constructed, which includes the behavioral characteristics of fraudsters, the reaction of victims, the success rate of fraud, and other multi-dimensional information.

4 Experimental results and analysis

In this study, a complex criminal network was successfully excavated from social networks by using DM techniques such as association rule mining, cluster analysis, and community discovery.

Firstly, the Apriori algorithm is used to mine the association rules of users' communication data on social platforms, and frequent crime-related words and phrases and their relationships are found. These association rules reveal the communication mode and hidden information of criminals in social networks (Table 1).

Table 1. Mining results of association rules

serial number	Antecedent	Consequent	Support	Confidence	Lift
1	{Transfer, Emergency, Remittance}	{Fraud}	0.03	0.85	4.2
2	{High return, investment}	{Fraud}	0.02	0.78	3.8
3	{QR code, payment}	{Fraud}	0.025	0.8	4.0
4	{Free, collect, prize}	{Fraud}	0.018	0.75	3.7
5	{Part-time job, billing}	{Fraud}	0.022	0.77	3.8
6	{impersonation, public security law}	{Fraud}	0.015	0.9	4.5
7	{Relatives and friends, arrested, ransom}	{Fraud}	0.01	0.88	4.4
8	{Low price, snap up, mobile phone}	{Fraud}	0.012	0.7	3.5
9	{Loan, unsecured}	{Fraud}	0.017	0.73	3.6
10	{System upgrade, abnormal account}	{Fraud}	0.014	0.8	4.0

The association rules in Table 1 all show a high degree of confidence and promotion, which indicates that these rules are strongly associated. For example, in Rule 1, when the words "transfer", "emergency" and "remittance" appear in the current item, there is a high probability (confidence 0.85) that they will be associated with "fraud". This strong association can provide important clues for investigators and help them quickly identify potential fraud. There are various fraud methods, including but not limited to transfer and remittance, high-return investment, QR code payment, free prize collection, part-time brushing, etc. This reflects the complexity and variability of online fraud and requires the public and law enforcement departments to be highly vigilant and constantly update their anti-fraud knowledge.

Confidence reflects the probability of fraud in the case of a specific Antecedent. The promotion degree reveals that the correlation between antecedents and subsequent is stronger than the probability of their independent appearance, which shows that these association rules are not accidental, but statistically significant.

These association rules can not only help investigators track and crack down on criminal acts but also be used for public education and fraud prevention. By publicizing these common fraudulent means and related words, the public's anti-fraud awareness can be improved and the occurrence of fraudulent cases can be reduced.

Furthermore, a K-means clustering algorithm is used to cluster and analyze users involved in crimes, and users with similar social behaviors and criminal characteristics are gathered together. Through cluster analysis, some potential criminal gangs and crime hotspots were found, which provided important clues for further investigation (Figure 2).

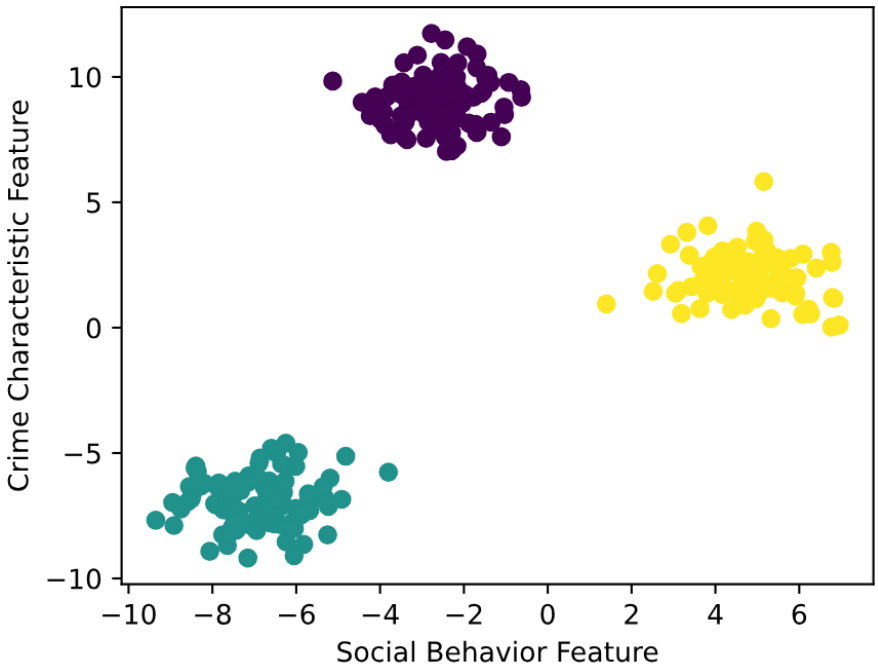

Figure 2. K-means clustering results

Finally, the Louvain algorithm is used for community discovery, and the whole criminal network is divided into several close communities. These communities represent different criminal organizations or gangs, and they connect and communicate with each other through key nodes (namely key criminal suspects). By analyzing the structural characteristics and key nodes of the community, we can have a deeper understanding of the operation mode of criminal organizations and the spread path of criminal acts. As shown in Figure 3.

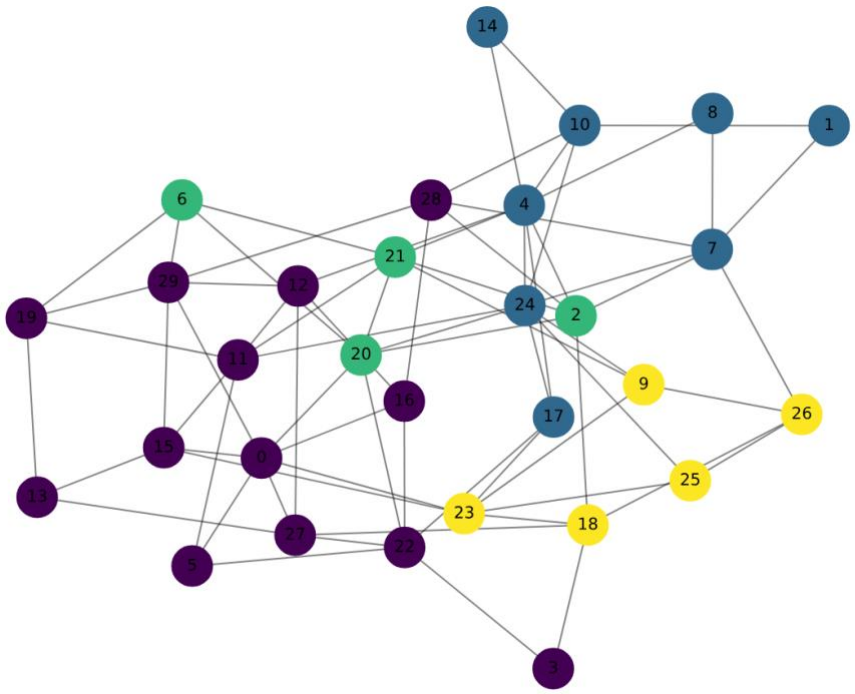

Figure 3. Community detection based on the Louvain algorithm

It can be observed that the nodes in the network are effectively divided into different communities, and each community is composed of nodes with similar characteristics. This division is based on the analysis of the network structure by the Louvain algorithm, which detects the community structure by optimizing the modularity of the network. In the diagram, each node represents an entity in a network (people involved in fraud cases, users on social platforms, etc.), and the connection between nodes indicates some association between these entities (participating in the same fraud case together, communicating frequently on social platforms, etc.).

By distinguishing colors, we can see that there are many communities in the picture. The nodes in each community have the same color, which indicates that they are classified into the same community by the algorithm. These communities may represent different fraud gangs, different groups in social networks, or user groups with similar interests and behavior patterns.

According to Figure 4, the distribution of criminal activities in different geographical areas can be observed. The darker areas in the picture indicate that criminal activities are intensive, that is, crime hot spots; While lighter areas indicate relatively less criminal activity.

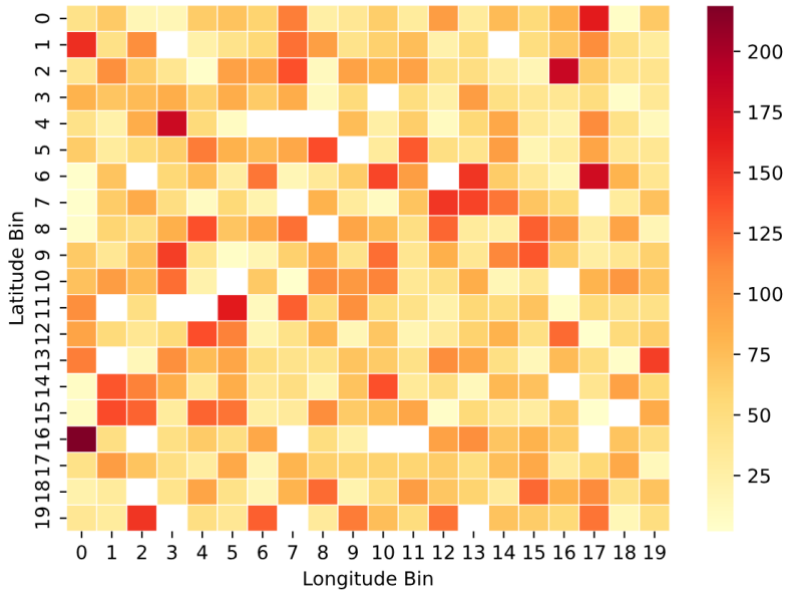

Figure 4. Hot spots of criminal activities

There are several major crime hotspots. These areas correspond to some specific locations in the city, such as commercial centers, transportation hubs, or densely populated residential areas. The darkest area is the main crime hotspot. These areas have become the key areas of criminal activities because of dense crowds, frequent economic activities, or relatively poor social security. Given these areas, the public security department should strengthen patrol and monitoring, improve the level of security, and reduce the occurrence of criminal incidents.

Some hot spots in the picture are slightly lighter in color but still obvious. Although criminal activities in these areas are not as intensive as those in major hotspots, they are still crime-prone areas. Public security departments should also pay attention to these areas, strengthen the comprehensive management of social security, and improve residents' safety awareness. Most areas in the picture are light in color, indicating that there are relatively few criminal activities in these areas. This does not mean that we can completely relax our vigilance, because criminal activities may change with the change of time and environment. Therefore, the public security department should continue to pay

attention to the public security situation in the whole region and adjust the security prevention strategy in time.

Based on historical data and mining results, a crime trend prediction model is constructed. The model can comprehensively consider many factors, such as the evolution trend of criminal networks and the changes in the social behavior of criminal suspects, and predict the future crime trend. By comparing with the actual criminal cases, it is found that the model has high prediction accuracy. This provides timely early warning information for public security departments and helps them to take measures to prevent and crack down on criminal acts in advance. Figure 5 shows the historical data on the number of online fraud crimes in the past few years and the forecast trend for the next few months.

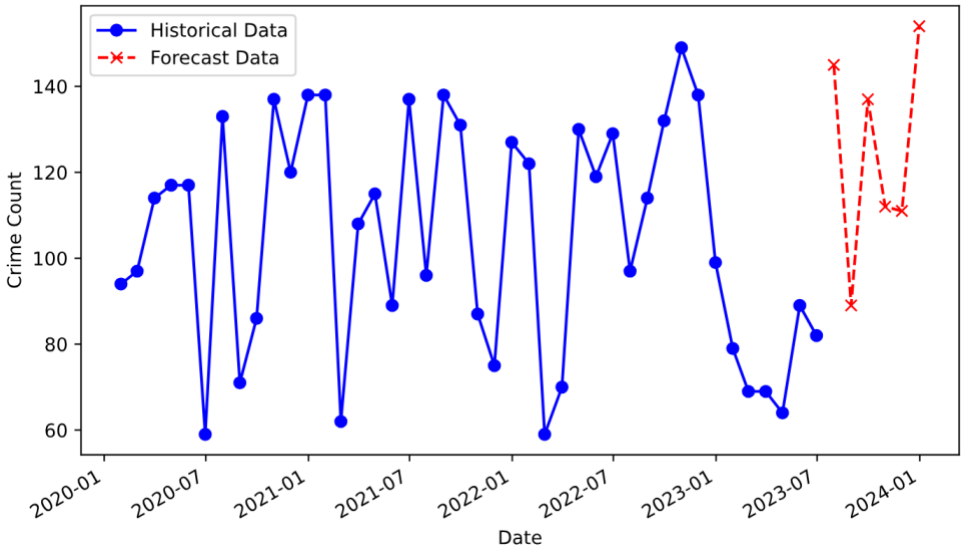

Figure 5. Historical crime data and forecast trend

From the historical data, the number of online fraud crimes shows certain fluctuation, but there is no obvious long-term upward or downward trend. This shows that in the past few years, online fraud activities may have been affected by many factors, including but not limited to the changes in law enforcement, the improvement of public awareness of prevention, and the adjustment of criminals' strategies. Some peaks and valleys can be observed, which may correspond to specific time nodes, such as holidays, major social events, or economic fluctuations. These periods are often the high incidence of online fraud because criminals usually use these opportunities to commit fraud.

According to the forecast model, the number of online fraud crimes may increase slightly in the next few months, but the overall trend is relatively stable. This means that although there may be a small increase during the forecast period, it is unlikely that there will be a large surge. It should be noted that this forecast is only based on the currently available data and a simple forecast model. The crime of online fraud is influenced by many complicated factors, including social environment, technological progress, legal policies, and so on. Therefore, to predict the future trend more accurately, it is necessary to continuously collect new data and improve the prediction model.

5 Conclusion

This study deeply discusses the application of SNDM technology in crime analysis and prevention and successfully excavates complex criminal networks and their behavioral characteristics from social networks through various DM technologies such as association rule mining, cluster analysis, and community discovery. The research results show that SNDM technology has important

application value in revealing the social behavior characteristics of criminal suspects, the organizational structure, and the activity rules of criminal gangs. First of all, through the association rules mining technology, we found the crime-related words and phrases hidden in social network communication and their strong correlation. These association rules provide investigators with clues to quickly identify potential fraud. Secondly, the application of cluster analysis technology reveals the user groups with similar social behaviors and criminal characteristics and further reveals the potential criminal gangs and crime hot spots. Finally, the community discovery technology successfully divides the whole criminal network into several close communities, and each community represents a different criminal organization or gang, which helps the public security department to have a deeper understanding of the operation mode of criminal organizations and the propagation path of criminal acts. This study not only improves the efficiency of criminal investigation but also provides scientific and effective means of investigation and prevention for public security departments. Based on historical data and mining results, a crime trend prediction model is constructed, which provides timely early warning information for public security departments and helps them to take measures to prevent and crack down on criminal behavior in advance.

Acknowledgment

This paper is supported by the Key Project of Public Security Theory and Soft Science Research Program (2023LL20) "Research on the Construction of Social Security Prevention and Control System from the Perspective of National Security". This study is a phased research achievement of the funded project.

Data Availability

The datasets used to support the results of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] Traynham, S., Kelley, M. L., Long, L. J., & Britt, T. W. (2019). Posttraumatic stress disorder symptoms and criminal behavior in U.S. Army populations: The mediating role of psychopathy and suicidal ideation. The American Journal of Psychology, 132(1), 85.

[2] Li, D., Zhang, Y., & Li, C. (2019). Mining public opinion on transportation systems based on social media data. Sustainability, 11(15), 4016.

[3] Wall, D. S. (2018). How big data feeds big crime. Current History (New York, N.Y.: 1941), 117(795), 29-34.

[4] Khemprasit, J., & Esichaikul, V. (2016). Design and implementation of a mobile crime analysis and monitoring system (MCAM) based on service-oriented architecture (SOA). Information Development, 32(4), 861-879.

[5] Lemov, R. (2018). An episode in the history of precrime. Historical Studies in the Natural Sciences, 48(5), 637-647.

[6] Birendra, K. C., Duarte, M., Erin, S., Jordan, S., & Peterson, M. (2018). Bonding and bridging forms of social capital in wildlife tourism microentrepreneurship: An application of social network analysis. Sustainability, 10(2), 315.

[7] Kumari, R., Jeong, J. Y., Lee, B. H., Choi, K. N., & Choi, K. (2021). Topic modelling and social network analysis of publications and patents in humanoid robot technology. Journal of Information Science, 47(5), 658-676.

[8] Hao, T., Chen, X., & Song, Y. (2020). A topic-based bibliometric analysis of two decades of research on the application of technology in classroom dialogue. Journal of Educational Computing Research, 58(7), 1311-1341.

[9] De Filippi, F., Cocina, G. G., & Martinuzzi, C. (2020). Integrating different data sources to address urban security in informal areas: The case study of Kibera, Nairobi. Sustainability, 12(6), 2437.

[10] Bennion, J. (2022). Polyamory in Paris: A social network theory application. Sexualities, 25(3), 173-197.

[11] Wen, H., Liang, K., & Li, Y. (2020). An evolutionary game analysis of internet public opinion events at universities: A case from China. Mathematical Problems in Engineering, 2020(19), 1-14.

[12] Skoric, M. M., Liu, J., & Jaidka, K. (2020). Electoral and public opinion forecasts with social media data: A meta-analysis. Information (Switzerland), 11(4), 187.

[13] Mayer, L., Lloyd, A., & Hitoshi, N. (2019). The promises and perils of using big data to regulate nonprofits. Washington Law Review, 94(3), 8-8.

[14] Wang, J., Shan, Z., Gupta, M., & Rao, H. R. (2019). A longitudinal study of unauthorized access attempts on information systems: The role of opportunity contexts. MIS Quarterly, 43(2), 601-622.

[15] Kfrerer, M. L., Martin, N. G., & Schermer, J. A. (2019). A behavior genetic analysis of the relationship between humor styles and depression. Humor - International Journal of Humor Research, 32(3), 417-431.