Automatic Crime Detector: A Framework for Criminal Pattern Detection in Big Data Era

Md Ileas Pramanik  
City University of Hong Kong, mpramanik2-c@my.cityu.edu.hk

Raymond Y.K. Lau  
City University of Hong Kong, raylau@cityu.edu.hk

Md Kamal Hossain Chowdhury  
Comilla University, kamal.cou@gmail.com

Follow this and additional works at: http://aisel.aisnet.org/pacis2016

Recommended Citation
http://aisel.aisnet.org/pacis2016/311

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2016 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Automatic Crime Detector: A Framework for Criminal Pattern Detection in Big Data Era

Md Ileas Pramanik, Department of Information Systems, City University of Hong Kong, mpramanik2-c@my.cityu.edu.hk
Raymond Y.K.Lau, Department of Information Systems, City University of Hong Kong, raylau@cityu.edu.hk
Md Kamal Hossain Chowdhury, Department of Computer Science and Engineering, Comilla University, Bangladesh kamal.cou@gmail.com

Abstract

In the era of “big data”, a huge number of people, devices, and sensors are connected via digital networks, and there is tremendous amount of data generated from their interactions every day. Effective processing and analysis of big data could reveal valuable knowledge that enable us to deal with emerging problems in a timely manner. However, rarely we can find big data analytics models and methods for crime forensics discussed in existing literature. In this paper, we illustrate a novel big data analytics framework that leverages heterogeneous big data resources for criminal pattern detection. The proposed framework can uncover the inherent structural properties of criminal networks which are essential for both crime investigation and the development of operational strategies to disrupt criminal networks. The structural analysis functionality generated by our proposed system could significantly improve the efficiency and accuracy of network analysis tasks. The proposed framework consists of two important analytical approaches, namely structure analysis, and network mapping. Based on the proposed framework, we have developed a prototype system called Automatic Crime Detector (ACD) that incorporates several big data analytic methods. Our empirical evaluation shows that the proposed framework is effective for criminal network discovery.

Keywords: Big data, Crime, Network analysis.
1. Introduction

Historically finding solutions for preventing criminal activities are the main task of the criminal justice and law enforcement specialist. In big data era, huge number of people, criminals, devices, and sensors are connected via digital networks and cross plays among these entities generate enormous valuable information that facilitate the law enforcement officers and detectives to speed up the procedure of resolving crimes. In this research, we will take an interdisciplinary approach among computer science, social science, and criminal justice to develop a big data paradigm that can help solve crimes easier and faster. We will present some relevant terminologies that are used in big data context in criminal justice and law enforcement departments.

Terrorism activities are classified into two broad categories. One is individual crimes where single suspect is involved (e.g. rape) and another is organized crime where a group of suspects are involved (e.g. armed robberies, gang related incidents). Since the twin tower attacks on September 11, 2001; terrorist network analysis for organized crime has received a lot of attention by academics for preventing future attacks (Latora, V., & Marchiori, M. 2004). Organized crime such as drug trafficking, organized attacks, fraud, and armed robberies, often requires conspiracy and chain actions (Chen et.al. 2004).

Police department and detective agencies have long realized that pattern of criminal networks is important to crime investigation (McAndrew 1999). A clear picture of network structures, group operations, and individual roles can help police department and detective agencies to take proper measures for preventing criminal activities. Currently, most law enforcement and intelligence agencies are met with enormous quantities of data from different big data resources that must be processed and turned into useful information. Although law enforcement agencies are increasingly acquiring data fusion and data mining to support their crime analytic capabilities, they still have lack of tools and technical supports to gain big data opportunities (Schroeder, J. et.al.2003). Generally, big data is characterized by “3V”- Volume, Velocity, and Variety (A. McAfee and Erik B. 2012). Volume presents huge amount of data generated from different resources. Velocity indicates the data generation speed that is very faster than the ability of existing data processing tools. Variety is associated with different formats of data such as structured, semi-structured and unstructured. Recently, a new dimension “veracity” has been included in big data to ensure the credibility and accuracy of data (Lu, R. et.al 2014).

In this research, veracity is an important dimension because fabricated data will extract fictitious criminal pattern, which will misguide law enforcement agencies. Therefore, veracity of data should be checked at early stage of analytic process. In this article we exploit a new opportunity of big data, especially devoting attention toward efficient and accurate extraction of criminal pattern in big data era. First, we present a general architecture of big data analytics. Then we review the existing criminal network analysis tools in datamining context. We also present a framework for offering a clear view of criminal network pattern from large-scale data such as social media data, newspaper reports, police narrative reports to help law enforcement and intelligent agencies. Then we evaluate our system through experiment and in the end, we conclude the paper.

2. Big data architecture

Based on application of big data a number of architecture are proposed from different research work (Biesdorf, S., 2013). For describing privacy analytics an architecture of big data is formulated in (Lu, R.et.al.2014).In this section, we will describe a general architecture of big data analytics in criminal data domains. A general architecture of big data is shown in figure-1, which mainly consists of three important phases:

- **PHASE 1** Big data sources
- **PHASE 2** Big data processing, modeling , and
- **PHASE 3** Value presentation
Figure 1. *General architecture of big data*

**Phase-1** is connected with data collection, acquisition, and filtering by using correct metadata and processes. In this phase, data from various sources are aggregated and transformed to complement meaning to the data and the information value allows organizations to achieve competitive benefits.

**Phase-2** is concerned to apply different analytics and predictive models to find the pattern and relationship among large-scale datasets. Based on organizational requirements the levels of intelligence are varied in big data processing and modeling. This phase focuses on the current and future rather than traditional historical reality to grab big data challenges.

**Phase-3** mainly organizes and maps the data to the desired model whilst interprets the meaning of the newly discovered information. Though structured datasets are naturally accommodated to the relational datasets, semi-structured and unstructured datasets are not like so. For accommodating, the unstructured datasets there are a large number of applications that focus on providing access to these data sources via NoSQL.

3. **Literature Review**

3.1 **Data analytics in Criminal data domains**

The main ideas, used in criminal data mining, are categorized into Association and sequential rule mining, Classification and prediction, and Data extraction and clustering techniques (Chen H.et.al 2004). Association rule mining technique has been widely used for detecting intrusion detection from users’ interaction records. Moreover, investigator can apply this approach to intruders’ profiles to perceive future network attacks (W.Lee, et.al.1999). Like association rule mining, sequential rule mining mainly find the frequency occurring sequence events over some predefined transaction in different times (Ayres, J. et.al 2002). This approach can identify intrusion pattern among time-stamped data. Similarly, classification approach finds the similar characteristics among different crime entities and organizes them into predefined classes. Often used to predict crime trends, classification and prediction technique reduce time to catch crime entities. Data extraction identifies specific pattern from heterogeneous datasets. It has been used to automatically identify individual address, characteristics, location from the police narrative unstructured records (Chau M. et.al 2002). Moreover, some statistical base approaches are used to associate different objects such as criminals, suspects, and organizations in crime records (Hauck, R. et.al 2002, Lau, R. Y.et.al.2014). By using clustering methods the financial Crime Enforcement Network AI Systems (T.Senator et.al 1995) identify similar transaction to detect money laundering and other financial crimes.

3.2 **Existing Criminal Network Analysis Tools**

Klerks (2001) classifies existing criminal network analysis tools are broadly classified into three generations: I) *First Generation: Manual Approach*. Representative of the first generation is the Anacpapa Chart (Harper et.al 1975). In this approach, an analyst a) construct an association matrix, b) examining data files to identify the links between criminals, and c) Draw a link chart can be drawn based on association matrix for visualization purposes. II)*Second Generation: Graph-Based Approach*. Second-generation tools are developed by COPLINK that can automatically produce graphical
representation of criminal networks (Hauck, R. V. et al. 2002). Most current criminal network analysis tools such as Analyst’s Notebook, Netmap, and Watson are belong to this generation (Xu, J. & Chen H. 2005). Law enforcement agencies in USA and Netherlands are widely using Analyst’s Notebook (Klerks 2001). III) Third Generation: Structural Analysis Approach. Third generation approach provides more develop analytical functionality to assist discover structural properties of criminal networks such as central members, cells, interaction patterns, and overall structure of the network (Chen, H., Chung et al. 2004).

3.3 Social Network Analysis

Social network analysis (SNA) technique has used to detect terrorist cells, discovering their pattern of interaction, identifying central and peripheral individuals, finding leaders, followers and gatekeepers, and uncovering network organization and structure (Wasserman et al. 1994; McAndrew 1999). Baker et al. (1993) and Saether et al. (2001) used SNA for mapping evidence in fraud and conspiracy cases. In SNA method, networks are represented as a graph, which contains a number of nodes (network members) and links (relationship). In structural analysis of criminal network, centrality measures and position-role analysis are two important factors to identify key members who play important roles in a network.

3.3.1 Centrality Measures

Freeman (1979) presented a review of key centrality concepts where he offered three most popular centrality measures: degree, betweenness, and closeness (see figure-2(a)). Freeman’s development was partly motivated by structural properties of the center of a star graph.

<table>
<thead>
<tr>
<th>Centrality Measures</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree of nodes</td>
<td>the degree centrality of vertex ( C_d(u) = \sum_{v \neq u} w(v, u) ) where ( w(v, u) ) is a binary variable indicating whether a link exist between nodes ( u ) and ( v ) and ( n ) is the total number of nodes in a network.</td>
</tr>
<tr>
<td>Betweenness of node ( u )</td>
<td>( B(u) = \sum_{v \neq u} \sum_{w \neq u} \frac{\sigma(v, w)}{\sigma(v, w \cup u)} ) where ( \sigma(v, w) ) indicates the shortest path between other two nodes ( v ) and ( w ) passes through the node ( u ).</td>
</tr>
<tr>
<td>Closeness of node ( u )</td>
<td>( C_c(u) = \frac{1}{\sum_{v \neq u} d(v, u)} ) where, ( d(v, u) ) is shortest distance between ( u ) and ( v ).</td>
</tr>
</tbody>
</table>

(a) (b)

*Figure 2.* (a) Three most popular Centrality measure (Freeman 1979), (b) Positions in Social Network

Network partitioning is another important method to analyze criminal network where two methods are employed: matrix permutation and hierarchical clustering (Arabie et al. 1978; Wasserman et al. 1994). Matrix permutation mainly re-arrange rows and columns of a matrix such that members with same properties are organized in same group. Due to the NP hard problem in matrix permutation, many researchers give special attention to use hierarchical clustering approaches in SNA studies (Arabie et al. 1978).

3.3.2 Position and Role Analysis

Positions and Roles are important concepts in social network analysis where they are intrinsically linked. Position is a collection of network nodes who are structurally substitutable and similar in social activities, status, and links with other nodes (Wasserman et al. 1994). Moreover, two nodes from same position in the network do not need to be directly linked in positional analysis whereas it is necessarily required in relational analysis (Scott 1991, Lorrain and White 1971). Example in figure-2(b), where
nodes A, B, D are in the same position because they are connected with other nodes (C, E) in the same way.

4. Proposed Framework for Criminal Pattern Detection

To facilitate criminal pattern detection, in this section, we propose a framework that incorporates two analytical methods- a) Structure analysis and b) Network Mapping. Based on this framework we develop a prototype system, Automatic Crime Detector, which can be applied in big data context. Our proposed framework can be categorized as third generation network analysis tool. We also examine our framework in section 6 where we find out whether criminal pattern detection can be discovered effectively and efficiently by using a number data analytical process provided by the system. In figure-3, we present our proposed framework.

![Figure 3. Framework for automatic criminal pattern detection](image)

4.1 Structural Analysis

We used three most popular centrality measures in our framework. In order to identify central members in any criminal network degree of centrality, closeness centrality, and betweenness centrality are widely used. Degree centrality of a node is measured as the number of ties this node has with all other nodes. If N is the total number of nodes in a criminal network then degree centrality of node \( u \) can be measured by \( C_d(u) \) (See section 3.3.1) that can be standardized by dividing by \( N - 1 \).

\[
C_{ds}(u) = C_d(u)/(N - 1)
\]

Closeness centrality of a node is calculated by the total distance of this node from all other nodes. Standardized closeness is defined as- \( C_{cs}(u) = C_c(u)/(N - 1) \)

The number of shortest path that pass through a given node defines betweenness centrality of this node. As maximum betweenness centrality of a node is\(((N - 1) \times (N - 2))/2\), the standardized betweenness centrality is: \( C_{bs}(u) = 2 \times C_b(u)/(N - 1)(N - 2) \).

Beside the Wiener index and the BRS compactness measure, we measure the general level of connectedness, defined by density (D), of the network and the centrality for the whole network, denoted by \( C_{network} \). The density is done as follows:

\[
D = \frac{2 \times L}{N \times (N - 1)} \text{ where } L \text{ denotes the set of links}
\]

\( C_{network} \) is calculated as follows:

\[
C_{network} = \sum_i(C_{max} - C_i)/\text{Max possible value} \quad (1)
\]

Equation (1) can be applied to find degree, closeness, and betweenness centrality of the overall network. \( C_{max} \) is the largest value obtained in the network and ‘max value possible’ for the numerator given the total number of nodes. For the star topology the total network C-measures is one.

In our framework, we employ Dijkstra’s(1959) for measuring Wiener index and the BRS compactness measures in a cell or subgroup. In Dijkstra’s algorithm overall time complexity is \( O(n^3) \). By using this algorithm, we get closeness without spending any extra time. We can easily extract dependency matrix
from the network. The dependency matrix benefits an analyst by providing a clear picture where it presents how much a particular node is dependent on others and presents how much others depend on that particular node. The dependency matrix is useful for role and position analysis of a network as well. Figure-4 (I) presents the examples of three sample networks and table-1 presents the weight of their different characteristics. A sample crime network is shown in figure-4 (II) and dependency matrix is shown in table-2. In the evaluation section we will describe the significance of different values in different characters.

Figure 4. (I) Three sample networks (1,2,3); (II) A sample Criminal Network

<table>
<thead>
<tr>
<th></th>
<th>Network 3</th>
<th>Network 2</th>
<th>Network 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>(C_{bs}(a) = C_{bs}(b) = C_{bs}(c) = C_{bs}(d))</td>
<td>0.333</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(C_{bs}(u))</td>
<td>0.667</td>
<td>0.667</td>
<td>1</td>
</tr>
<tr>
<td>Betweenness of Whole Network, (C_{d(network)})</td>
<td>0.056</td>
<td>0.667</td>
<td>1</td>
</tr>
<tr>
<td>(C_{ds}(a) = C_{ds}(b) = C_{ds}(c) = C_{ds}(d))</td>
<td>0.75</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>(C_{ds}(u))</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Degree of Whole Network, (C_{d(network)})</td>
<td>0.333</td>
<td>0.667</td>
<td>1</td>
</tr>
<tr>
<td>(C_{cs}(a) = C_{cs}(b) = C_{cs}(c) = C_{cs}(d))</td>
<td>0.8</td>
<td>0.667</td>
<td>0.571</td>
</tr>
<tr>
<td>(C_{cs}(u))</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Closeness of Whole Network, (C_{c(network)})</td>
<td>0.467</td>
<td>0.778</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Density, and Centrality measures for networks shown in figure-4(I)

<table>
<thead>
<tr>
<th>Node</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>100</td>
<td>25</td>
<td>33</td>
<td>22</td>
<td>17</td>
<td>17</td>
<td>214</td>
</tr>
<tr>
<td>b</td>
<td>17</td>
<td>0</td>
<td>33</td>
<td>50</td>
<td>25</td>
<td>17</td>
<td>17</td>
<td>159</td>
</tr>
<tr>
<td>c</td>
<td>17</td>
<td>33</td>
<td>0</td>
<td>25</td>
<td>42</td>
<td>17</td>
<td>17</td>
<td>151</td>
</tr>
<tr>
<td>d</td>
<td>17</td>
<td>33</td>
<td>17</td>
<td>0</td>
<td>33</td>
<td>17</td>
<td>17</td>
<td>134</td>
</tr>
<tr>
<td>e</td>
<td>17</td>
<td>25</td>
<td>33</td>
<td>33</td>
<td>0</td>
<td>17</td>
<td>17</td>
<td>142</td>
</tr>
<tr>
<td>f</td>
<td>17</td>
<td>22</td>
<td>25</td>
<td>25</td>
<td>100</td>
<td>0</td>
<td>17</td>
<td>206</td>
</tr>
<tr>
<td>g</td>
<td>17</td>
<td>25</td>
<td>17</td>
<td>58</td>
<td>42</td>
<td>17</td>
<td>0</td>
<td>176</td>
</tr>
<tr>
<td>Sum</td>
<td>102</td>
<td>238</td>
<td>150</td>
<td>224</td>
<td>164</td>
<td>102</td>
<td>102</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Dependency matrix for network of figure-4(II).

4.2 Network Mapping

To visualize a criminal network into 2D graphical display, we used multidimensional scaling to assign a location to each node in a network of \(N\) nodes, given the corresponding \(N \times N\) distance matrix. We present criminal networks using a familiar node-link representation, where nodes represent members of the system and links represent the articulated “relationship” links between them. We adopted Classic Metric Multi-Dimensional (MDS) Scaling technique, which takes all-pairs-shortest path matrix as input. We selected Torgerson’s (1952) classical metric MDS algorithm that use a model of 2D projection of high dimensional space of the network using matrix algebra to determine Eigenvectors. For the exact optimization, classic MDS technique uses singular value decomposition (SVD) approach.

5. The Prototype System

We have developed a prototype system called ACD (Automatic Crime Detector) system, based on our proposed framework. The system is implemented in python and text data is formatted in XML. A
screenshot of the developed prototype system is shown in figure-5. ACD has six main menus (File, Edit, Extract, Cluster, Structure, Map, and help). First two menus- file and edit, are used for the similar purposes like typical application system. Generally, file and edit menus cover functions like, new, open, save, edit, print, cut, copy, paste, and so on.

**Figure 5. A Snapshot of the network generated by the ACD-Prototype System**

# 6. Framework Evaluation

## 6.1 Dataset and Experiment Setup

Our dataset contains 56,530 records that we collected from the official site of Los Angeles County Sheriff’s Department’s jurisdiction\(^1\). The database covers criminal records from 2005 to 2014. We choose that duration because different social media are flourished in that time. In collected datasets, each record belongs to an individual criminal and every record has 18 attributes in the original data, but we use only 8 attributes in our experimental dataset. Though sheriff’s dataset holds 30 type of crimes, we consider 18 crimes in our laboratory experiment work. All data records are clustered into 350 clusters. We evaluated our framework in different clusters. Due to space limit, we describe our experiment when it is applied for analyzing a cluster that contains 100 criminal records.

## 6.2 Relational Pattern and Central Members Identification

Structure analysis and network mapping modules present the centrality measures and dependence matrix that are used to provide the clear picture to identify how much a particular node is dependent on others and identify how other nodes depend on any particular node as well. Moreover, this type of structure analysis and network mapping facilitate crime association and pattern visualization. Table-3 shows the centrality measure and dependency matrix that are calculated by ACD system.

In centrality measure, degree and betweenness measures are very simple where the node that has more links to other nodes presents higher degree, and betweenness actually measures the dependency. When more nodes depend on a particular node to communicate with other nodes then that node holds higher betweenness. In table-3, node 8 has higher degree and betweenness than other nodes, which means node 8 is potential node in this criminal network. Betweenness measures for nodes 2, 4, 5, and 7 are zero because they are not used for communication between any other pairs. The closeness of a node to all other nodes in the network is measured by the total distance between that node and all other nodes in a cell. The central nodes (cell leader) have higher closeness value, as they have minimum total distance to all other nodes in a cell.

---

\(^1\) [http://shq.lasdnews.net/CrimeStats/CAASS/desc.html](http://shq.lasdnews.net/CrimeStats/CAASS/desc.html)
For position and role analysis, we apply the dependency matrix that provides a clearer picture than betweenness or closeness centrality. In table-3, the smallest sum of raw values of a particular node presents that the node is the most difficult to be deactivated from the network, such as node-8 that holds central position and has multiple links to communicate with other nodes. Whereas, highest sum of row values indicates that the node can be easily deactivated, such as node-4, node-2, node-5, which are in peripheral position and have limited links to connect with other nodes. In case of column, the lowest sum value means that least communication takes place through these nodes. Therefore, capture of these nodes will have low effect to disrupt the network or even cell. The largest sum value indicates that good number of communication takes place through these nodes. Thus, capture of these nodes will damage a network. In essence, the structural analysis functionality generated by the ACD system could significantly improve the efficiency and accuracy of network analysis tasks. Moreover, ACD approach is able to analyze dynamic network in large-scale datasets. Finally, the ACD system could suggest investigative leads, which would otherwise be overlooked and help to prevent crimes by disrupting criminal networks accurately and efficiently.

<table>
<thead>
<tr>
<th>Node Label</th>
<th>Centrality Measure</th>
<th>Dependency Matrix (Position and Role analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Degree</td>
<td>Betweenness</td>
</tr>
<tr>
<td>1</td>
<td>0.012</td>
<td>0.076</td>
</tr>
<tr>
<td>2</td>
<td>0.011</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.12</td>
<td>0.021</td>
</tr>
<tr>
<td>4</td>
<td>0.11</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.21</td>
<td>0.107</td>
</tr>
<tr>
<td>7</td>
<td>0.11</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.52</td>
<td>0.311</td>
</tr>
<tr>
<td>...</td>
<td>....</td>
<td>........</td>
</tr>
<tr>
<td>100</td>
<td>0.12</td>
<td>0.082</td>
</tr>
</tbody>
</table>

Table 3. Centrality measure and Dependency matrix of criminal network.

7. Conclusions and Future Work

In this research, we incorporated various multidisciplinary methods in different modules of our proposed framework for automatically extracting criminal network from various big data resources. As big data introduce several new challenges, all companies must deal with new tools and techniques in order to compete in the era of big data (J.Leon Zhao et.al.2014). Law enforcement agencies should deal with new strategic and operational challenges to process large volume heterogeneous data. In this article, we also developed a prototype system called ACD system, which has structure analysis functionality through forming cells, identify interaction pattern among cells, detect crime pattern, and identify cell leaders. Moreover, in this research work, we introduced a new measure of centrality, dependency centrality, and its application in role and position analysis of criminal network We also validated how this measure could assist police department and intelligence agencies to understand the structure of the criminal network. Therefore, our framework is able to identify the crime patterns from large-scale criminal data making the job for crime intelligence agencies easier. In future, we have a plan to incorporate ACD system into the COPLINK research that is seeking to build a united platform for information and knowledge management in crime detective domain. However, there are some limitations of the ACD system includes that crime pattern analysis can only assist the detective, not substitute them. In addition, ACD system only considered criminal-criminal relationships. To address this problem, we plan to include ‘criminal scheme network’ analysis that analyze not only people but also other important factors such as weapons, vehicles, locations, physical characteristics, property, and organizations.
Acknowledgements
The work described in this paper was partly supported by a grant from the Science Technology and Innovation Committee of Shenzhen Municipality - the Basic Research Program (Project: JCYJ20140419115614350), and the Shenzhen Research Institute of City University of Hong Kong.

References


