Advanced Review

Big data analytics for security and criminal investigations

M.I. Pramanik,[1*] Raymond Y.K. Lau,[1] Wei T. Yue,[1] Yunming Ye[2] and Chunping Li[3]

Applications of various data analytics technologies to security and criminal investigation during the past three decades have demonstrated the inception, growth, and maturation of criminal analytics. We first identify five cutting-edge data mining technologies such as link analysis, intelligent agents, text mining, neural networks, and machine learning. Then, we explore their recent applications to the criminal analytics domain, and discuss the challenges arising from these innovative applications. We also extend our study to big data analytics which provides some state-of-the-art technologies to reshape criminal investigations. In this paper, we review the recent literature, and examine the potentials of big data analytics for security intelligence under a criminal analytics framework. We examine some common data sources, analytics methods, and applications related to two important aspects of social network analysis namely, structural analysis and positional analysis that lay the foundation of criminal analytics. Another contribution of this paper is that we also advocate a novel criminal analytics methodology that is underpinned by big data analytics. We discuss the merits and challenges of applying big data analytics to the criminal analytics domain. Finally, we highlight the future research directions of big data analytics enhanced criminal investigations.

How to cite this article:
WIREs Data Mining Knowl Discov 2017, 7:e1208. doi: 10.1002/widm.1208

INTRODUCTION

Recent technological revolutions in data and communication systems enable us to generate and share data much faster than ever before.[1] The advancement of these systems not only provides high-speed communication facilities but also facilitates the outbreak of organized crimes. For example, terrorism often involves a number of individuals or teams who have designated attack roles; these interrelated offences establish a criminal network of organized crime.[2] Now, different security organizations, police departments, and intelligence agencies such as the FBI and CIA constantly collect and analyze data, and investigate organized and individual crimes in order to develop preventive measures to foil future attacks.[3] The notion of big data and its applications have attracted a lot of attentions from security intelligence organizations in recent years because of its potential of solving complex problems in an efficient way. Big data does not mean the large size of data only; big data is defined as 'the amount of data just beyond existing tools, techniques, and technologies to store, manage, and process efficiently.'[4] Generally, big data is characterized by '3V'—Volume, Velocity, and Variety.[1] Under the security intelligence context, the continuous flow of information, without information lag, is necessary for intelligence agencies to make real-time decisions. Currently, most law enforcement and intelligence agencies often encounter extraordinary volume of data emerging from a variety of data resources; these big data must be processed and turned into useful security intelligence quickly.

Security professionals have long realized that knowledge about criminals and their networks is

*Correspondence to: mpramanik2-c@my.cityu.edu.hk

[1]Department of Information Systems, College of Business, City University of Hong Kong, Hong Kong, Hong Kong SAR

[2]Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China

[3]School of Software, Tsinghua University, Beijing, China

Conflict of interest: The authors have declared no conflicts of interest for this article.

important for crime investigation.[5] Extracting the hidden network structures among criminals, and inferring their respective roles from criminal data can help law enforcement and intelligence agencies develop effective strategies to prevent crimes from taking place. With the big data analytics, key factors for criminal network discovery such as identifying central members and detecting subgroups can be done by automatically mining social media data from Twitter and Facebook.[6] During the 1980s, different data mining techniques, machine-learning algorithms, neural networks, and intelligent agents were designed for classification, prediction, and profiling of human behavior to meet the objectives of tracking and deterring criminals. The applications of these techniques have demonstrated that automatic analysis of crime trends and criminals' behavioral patterns without requiring the constant intervention of humans for examining numerous criminal attributes is feasible. Increasingly criminal activities in the world will become more digitized in nature, and so most criminal monitoring and detection agencies are eager to apply data mining techniques to identify scams. Moreover, to address the challenge of data analytics for security and criminal investigation, researchers have explored different data mining techniques in the past few decades. However, a thorough study of data analytics for security and criminal detection is missing in the existing literature. One contribution of this paper is the systematic evaluation of state-of-the-art data mining technologies including intelligent agents, link analysis, text mining, machine learning (ML), and neural networks. Another contribution is the critical analysis of big data analytics for security and criminal investigations. After examining some important aspects of security intelligence discovery from criminal networks, we propose a framework to manage big data under such a context. From a practical perspective, this paper introduces security managers, law enforcement, and intelligence agencies, fraud detection specialists, and information security analysts to the latest data mining techniques and shows how these techniques can be applied to enhance crime investigations.

DATA MINING

Data mining is a rapidly growing field positioned at the intersection of several sub-fields such as statistics, database research, high-performance computing, ML, and so on. In data mining systems, the mining procedure is the fusion of statistical modeling, database storage, and Artificial Intelligence (AI) techniques.[7] The main objective of data mining is learning from structured or unstructured data, and turning data to actionable knowledge. Statisticians have developed different theories and models to extract knowledge from data. These models enable us to analyze the links among variables for developing prediction models, quantifying effects, or suggesting casual paths.[8]

AI is one of the important sub-fields that lays the solid foundation of data mining methods. A large number of AI algorithms have already been developed to enable automated learning from data. These algorithms are the foundations of building different predictive models for detecting criminal activities, criminal behavioral profiling, and clustering of criminal data.[9] Thus, security organizations, police departments, and intelligence agencies now rely on different data mining techniques to detect and deter crime and terrorism. Several breakthrough applications have already emerged in which link analysis, intelligent agent, text mining, neural networks, and ML are being used for criminal data analysis to prevent potential criminal activities.

DATA MINING TECHNIQUES

As the number of organized crimes continues to rise in recent years, law enforcement, and security agencies require new and advanced technologies to fight this battle. In this section, we identify five cutting-edge data mining technologies—link analysis, intelligent agents, text mining, neural networks, and ML—that have been used to combat crime. These technologies have developed over time to incorporate different tools and methods. We summarize these five techniques in Table 1 in terms of their tools, applications, purposes, and challenges.

Link Analysis

Link analysis is a data mining technique, which reveals the structure of data by representing it as a set of interconnected, linked object or entities.[10] Link analysis starts with data that are represented as an interconnected network. By extrapolating the hidden relational patterns from the entities and their relationships, investigators, and analysts can discover useful links among entities. This technique can correlate massive amounts of data about entities in regard to fraud, terrorism, narcotics, and others.[11] This data mining technique is the first level of analysis by which networks of people, locations, groups, vehicles, contact addresses, bank accounts, and other tangible entities can be explored, assembled, detected, and analyzed (see Figure 1). Linkage data is

TABLE 1 | Five Data Mining Techniques

Technique	Identity	Opportunities	Challenges	Tools	Purposes (existing and potential)
Link ANALYSIS	Discover knowledge from connections, relationships, and critical links	Characterize relationship, Identify group, subgroup, and group leaders.	Heterogeneous data processing, and Intelligence capability.	Analyst's notebook, Crime link, ATAC, Crime Workbench, ORION, NETMAP, and VisualLink	G, D, A, C, and E
Intelligent agent	Computing entity that performs user delegated tasks autonomously	Use to develop expert system and Real time Bio-surveillance system	Contradiction elimination and learning mechanism	FIAS, InfoGist, and Doppelgaenger	T, D, F, and E
Text mining	Work with unstructured textual information	Discover knowledge from textual data	Handling the velocity and veracity of big textual data	Perilog, Autonomy, Clairvoyance, ClearForest, Klarity, iCrossReader, Lextek, Leximancer, Quenza, VantagePoint, and Readware	C, D, and E
Neural network	Mimic the cognitive neurological functions of human brain	Discover non-linear functions from noisy and incomplete dataset	Maturity time, privacy and human rights, low-face validity	Attrasoft, BrianMaker, NeuroSolutions, SPSS neural connection, STATISTICIA Neural Networks, COGNOS 4Thought, and Proforma	B, C, D, F, and E
Machine learning	Discover patterns from learning experience	Provide automated process of statistical operations	Skewed data, analytical uncertainty, dimensions of data	AC2, Business Miner, C5.0, CART, SAS, Quadstone, ANGOSS, and Polyanalyst	D, T, F, and E

A, Fraud and money laundering; B, Bio-terrorism; C, Clustering, classification and categorization of modus operandi; D, Data extraction and prediction; E, Etcetera (feature extraction, searching, browsing, visualization); F, Forensic investigation; G, General linking analysis for visualizing related crimes or criminal networks; H, Hunting criminal gang through tracking.

visualized as a graph with linked nodes where nodes represent suspects of interest for the investigators, and the link indicates a relationship or transaction between suspects and criminal artifact.

Link analysis has been used in different studies to analyze social networks over the last two decades.[12] Criminal link analysis tools visualize relationships among suspects or offenders usually do not bother too much about the exact mathematical density of entities.[13] Their objective is mainly to depict who does what to whom, and with what frequency. Jennifer Schroeder et al. develop a prototype system named CrimeLink Explorer that enables automated link analysis.[14] They also incorporate several techniques in their proposed system: the co-occurrence analysis approach and a heuristic approach for the identification of associations between crime entities, and a shortest path algorithm for association path search. In CrimeLink Explorer the heuristic approach supports to incorporate investigative domain knowledge into a link analysis approach for measuring association strength automatically.[11] A criminal link analysis approach is proposed in Ref 15, where the system uses the time-based relationship, event similarity, time-related proximity, and document distributional proximity to identify the events of a terrorist attack incident. For a criminal network, the characteristic of a criminal is extracted via some ties and links. The social capital associated with a network is also one of the key factors that facilitates the outbreaks of crimes. By using social network analysis (SNA), a link analysis tool is introduced in Ref 5. In SNA, the link analysis approach is used to extract degree, betweenness, and closeness centralities[16] to analyze criminal network. Though link analysis is the best approach for the structural data, it can also be applied to analyze unstructured data by incorporating some text mining methods.[17]

On the basis of existing theories and practice, we find that link analysis can explore associations

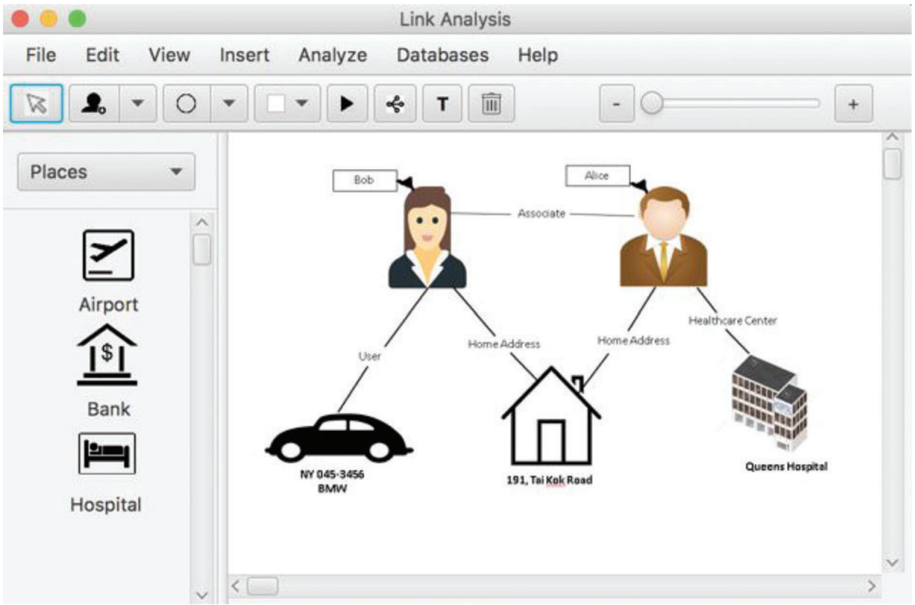


FIGURE 1 | Link analysis for criminal investigation.

among large numbers of entities with multiple attributes. Different individual instruments have already been developed by using these software packages. For instance, Analyst's notebook (www.i2.co.uk/home.html) assists investigators and analysts by providing faster, more informed decision-making across, and inside organizations. It provides a centralized, aggregated view of information from different sources. Recently, a company named i2 has developed a software that can be used on analysts' notebooks to assist FBI investigations. Another link analysis tool is crime link (www.crimelink.com) which is being used to assist the law enforcement investigators and analysts. ATAC (www.bairsoftware.com/atac.htm), Automated Tactical Analysis of Crime, is a criminal data analysis tool that can extrapolate potential criminal patterns and trends. Crime Workbench (CWB; www.memex.com/cwbover.html) is another mature link analysis tool for intelligence management. This tool offers comprehensive searching capabilities using the Memex information engine.

Intelligent Agent

Over the past two decades, the concept of agents has emerged as a new software paradigm, and it has become important in both AI and mainstream computer science. Under a criminal investigation context, agents perform the function of software detectives, monitoring, identifying, and extracting information for analysis, development, and real-time response. Agents can perform tasks delegated by their developers, and so they are also called task specific autonomous computing systems.[18] By using software agent, a security model is developed in Ref 19 where game theory is applied to develop security games between a defender and an attacker. Here, the proposed security games are bi-level models[20] that consider an attacker's ability to gather information about the defense strategy before planning an attack. For monitoring bio-terrorist attack, an agent based real-time bio-surveillance system is developed in another study.[21] A number of statistical methods are used for developing the proposed surveillance system. Moreover, in order to prevent natural hazard, Jaber et al.[22] proposed an automated decision support system incorporating both tele-geoprocessing approach and intelligent software concept. The software agents are used in corporative information systems where they collect and manipulate information obtained from relevant sources to answer the queries posed by other users or agents.[23]

By using intelligent agent technology, the FBI, and IRS (Internal Revenue Service) have developed some expert systems where younger agents automatically receive helpful advice from experienced agents who gained experience from previous solutions.[24] COPLINK agent is an example of such expert system that incorporates intelligent software agent for delivering alert messages through a number of communication channels including e-mail, and instant messaging.[25] Šišlák et al. develop an air traffic control system where agent technology supports security analysts by controlling the flow of air traffic.[26] A verification service system is developed in Ref 27, where software agents are used as integrated security agents. Furthermore, a framework of multi-agent based network security system is developed in another study.[28] There are numerous software agent tools, which are widely used in security investigation context. For example, FinCEN is used to detect money laundering through the FinCEN Artificial Intelligence System (FIAS).[29] InfoGist is an information retrieval software agent tool that can search and aggregate a set of webpages on the basis of some defined keywords. Doppelgaenger is such an intelligent tool, developed by the MIT media lab. This agent can evaluate criminal actions, monitor criminal behavior, and create new rules dynamically and then update these rules by herself. The agent also can alert investigators when a large irregularity is identified.

Text Mining

Text mining, also known as text data mining[30] or knowledge discovery[31] from unstructured datasets; it refers to the process of extracting relevant information and knowledge from textual documents. Text mining explores the key concepts or themes in documents rather than any specific keywords, and the taxonomies help the investigators and analysts to find relevant information in multiple linked documents.[32] Text mining tools and techniques offer appropriate links among many relevant textual information for discovering new knowledge to determine the appropriate course of actions. In Ref 32, authors proposed a text mining method to explore criminal networks from the collected textual documents. The proposed system primarily discovers useful knowledge for criminal study, and then visualize the extracted criminal network for investigation. Another similar research work is conducted in Ref 33, where relationship among criminals are extracted from the news articles about crimes. For discovering new knowledge from unstructured documents of police records, a comparative study was performed in another work[34] where usability of emergent self-organizing map (ESOM) and multi-dimensional scaling (MDS) were exploited as text exploration instruments to assist police investigations. Moreover, in Ref 35, Lee introduced an information extraction method from unstructured police records. He used matching phrases or sentences to predefined templates. This method mainly incorporates Natural Language Processing (NLP) techniques for identifying the required entities (e.g., individuals, profession, places, and time) from collected textual data. It then compares the phrases or sentences for extracting associations and patterns within a criminal network. Using text mining approach, it is possible to identify deceptive persons based on linguistic features. For instance, in Ref 36 text mining and data mining approaches were applied to determine the presence of false items in a textual content. For capturing sentiment feature, a rule-based system was developed by Yin, where contextual features were identified by comparing textual posts to a window of neighboring textual posts.[37] They also incorporated SVM in the text mining approach to classify harassing posts on online social media.

NASA proposed a text mining tool called Perilog, which can retrieve, and organize contextual relevant data from any sequence of terms.[38] This tool was originally developed to support the FAA's Aviation Safety Reporting System (ASRS).[39] This text mining tool was used to investigate the dominant cause of airline crashes. Autonomy (www.autonomy.com) is another text mining tool which provides a unified view of disparate data sources. Autonomy offers a user profiling system which can automatically identify a user's expertise through analyzing work patterns, records, and e-mail contents. Autonomy is used by UK police forces to categorize and tag stored criminal data. Police officers also use this tool to serve as a central police information repository. Copernic is another text mining tool that combines intelligent agent technology for context-aware retrieval, and text mining for theme aggregation; this tool is an effective forensic instrument for investigators and analysts. Clairvoyance (www.claritech.com) is a text analysis tool that applies NLP techniques to interpret and extract information from textual data such as written documents, newspaper reports, and other evidential documents.[40] Similarly, ClearForest (www.clearforest.com) is a text mining tool that can discover forensic knowledge from unstructured information repositories. It allows investigators and analysts to quickly extract and insightful information from textual files.

Artificial Neural Networks

An artificial neural network (ANN) is modeled as an aggregation of hundreds of thousands of basic computational units that mimic the function of neurons in a human brain. In ANNs, the functions of artificial neurons like learning, remembering, and decision-making are designed by some software systems that can mimic the cognitive neurological functions of the human brain.[41] Through executing the learning process and using programmable memories, ANNs can predict new trends based on existing samples, and so they are being used to predict potential criminal patterns based on observations of current criminal activities.[42] In order to provide possible psychological and behavioral profile of an unknown offender, a neural network approach can provide an investigative support tool. For example, a ANN-based psychological criminal profiling model is developed in Ref 43; the ANN model can conduct a sophisticated link analysis than traditional database oriented approach. ANNs have been used for entity extraction, where ANN-based algorithms enable investigators to identify useful entities from textual data such as police narrative reports. An ANN-based name entity extraction technique was developed from the COPLINK project.[44] In recent study,[45] an ANN-based system was developed for shortest path computation enabling the extraction of associations in a criminal network.

The ANN-base shortest path algorithms were applied to identify the strongest association paths between entities within a criminal network. There are several common properties, which are frequently visible in criminal data such as subjectivity, impression, noise, and incompleteness; these properties are most amenable to neural networks because they are highly fault-tolerant, and when properly trained, are capable of providing better solutions from degraded and erroneous raw data.[46] Although the complex properties of criminal data impose constraints on conventional security and criminal investigation techniques, ANNs can offer a significant contribution to criminal network analytics as they can properly integrate the elusive qualities of human reasoning with the compulsive thoroughness, precise logic, and perfect memory of computer.[47,48] By using neural network concept, an event initiator model was developed in Ref 49 where past behavior demonstrating the preference of offenders was used to infer both time and venue of future crimes. For extracting potential pattern of serial crime with a high degree of accuracy, Dahbur and Muscarello used an ANN tool for the classification of criminal patterns.[50] They designed a hybrid system by using neural networks and rule based heuristic techniques. Though neural networks offer significant support in terms of clustering, classifying, and data summarizing, there are still some challenges in criminal investigation and forensic study. When processing a large volume of data, ANNs may generate false alarms, which may lead to the wrong arrestment of suspects.[51] This type of problem is found in all application domains. Hence, new ANN algorithms should be developed such that the false alarms can be minimized, and law enforcers can effectively apply ANN to assist criminal investigation.

Over the past two decades, a large number of ANN tools have been developed and applied to different areas. These tools have very intuitive interfaces such that users find it easy to apply them to a variety of applications. Some ANN tools can automatically adjust their internal structures with respect to a dynamic environment.[52] Attrasoft (www.attrasoft.com) image finder is an image and facial recognition ANN tool that can handle unlimited numbers of suspect photographs at a rate of 600 images per minute.[53] BrainMaker (www.calsci.com) is a back propagation neural network tool that has the ability to handle more extensive automated training and tuning. Moreover, Ward Systems (www.wardsystems.com) STATISTICIA Neural Networks, (www.statsoftinc.com), and Pro-Forma (www.proformacorp.com) are also supplementary neural network tools which discover knowledge from the internal features of large scale criminal data.

Machine Learning

For a ML approach, a computer system learns like what human does based on past experience and different classes of tasks, and then acquired knowledge is used to make decisions and predictions.[54] ML techniques are used to adapt to dynamic and new circumstances and to detect and extrapolate internal and external structures as well.[55] In recent years, a large number of research work have appeared in the security and criminal investigation literature, which describes the application of ML techniques to identify potential anomalies (e.g., financial fraud), and criminal profiling by means of different inference engines.[50,56] Moreover, there has been a dramatic increase in the application of ML methods, tools, and techniques to facilitate diagnostic and prognostic investigation in forensic science.[57] ML approaches offer several new possibilities and find a place in the criminal investigation and forensic science. Based on partial criminal data, ML algorithms can predict vital crime patterns from among a large number of criminal incidents. These data driven tools enable investigators to better understand the patterns of crimes, leading to more specific attribution of previous crimes and the apprehension of suspects.[58] For detecting specific patterns of crimes that are committed by same offenders or gangsters, a ML algorithm is proposed in Ref 59; the algorithm can look for similarities between crimes in a growing pattern learning fashion from a database, and it tries to identify the behavioral codes that the offender follows.

When more criminal data are fed to this algorithm, behavioral codes become more fine-grained.

ML techniques are also used as an entity extractor that can identify useful entities from the police narrative reports.[60] Moreover, Yin et al. demonstrated that ML approach was more feasible and achieved better document classification results with respect to local features, sentiment features, and contextual feature.[37] They also found that supervised learning approach could accurately identify harassing posts in chat rooms and discussion forums. In some expert systems, both ML and neural network techniques are adopted to support police operations. For instance, in research work[46] the AICMS prototype system was proposed to address the operational needs of police who encountered an increasing number of criminal activities. The ACIMS system is a rule-based system with the support by machine-learning and neural network techniques. Another rule-based expert system was developed in Ref 61, where proposed system used the fuzzy set theory to improve the effectiveness and the quality of the data analysis phase of criminal investigation. Furthermore, ML approaches deliver quick responses through analyzing various dimensions of active data in criminal investigation.[62,63] As criminals can intentionally alter their methods of operations, this is another great challenge of ML methods to criminal investigation, which should be continuously adaptive in responding to the evolving crime patterns. For criminal investigation, ML will perform as an adaptive process rather than just a technology solution.

AC2 (www.alice-soft.com) is a ML software tool that uses decision tree learning, and it provides graphical interfaces for both data preparation and decision tree building. This tool can test all possible groupings of attributes, and hence to segment and extract the desired patterns that converge on a selected variable (eg., criminals versus suspects). ANGOSS (www.angoss.com) is a predictive analytics and ML software suite that has some decision tree components with industrial strength for discovering and visualizing links among variables in which the relationships among criminals are identified. Moreover, Prudsys (www.prudsys.de), Quadstone (www.quadstone.com), and SAS (www.sas.com) are also ML software tools than can be used to analyze criminal data for profiling, automated segmentation, and modeling.

Table 2 summaries 30 relevant studies with respect to six dimensions: technology, solution methods, data type, key contribution, scalability, and impact. The dimension of technology refers to the collection of techniques applied to the corresponding study. Though most of the studies use one of five data mining techniques, some studies employ hybrid techniques. Solution refers to the specific methods or theories that are applied to solve the research problems, and data type refers to the type of experimental data used. Research impact is an important aspect for the evaluation of success of any studies.

BIG DATA FOR CRIMINAL INVESTIGATION

Big data analytics include numerous state-of-the-art technologies that reshape security intelligence, which concerns the discovery of vital security knowledge from large amount of data. While the accessibility of big data creates different opportunities for police and intelligence agencies, it also brings some challenges to practitioners and researchers. Existing big data analytics is concerned with three important challenges storage, management, and processing.[4] Based on the characteristics of collected data, different methods are being applied to discover knowledge about criminal networks. When analysis models are developed based on a single data source, then analytical results may provide limited and biased knowledge. On the other hand, data from multiple sources provide a holistic view of the crime domain and allow more accurate and effective prediction of criminal patterns. Unfortunately, integrating heterogeneous data from multiple sources to discover criminal network is not a trivial task. This prompts for the exploration of advanced methods, platforms, tools, applications, and frameworks for effective data management in the context of security intelligence.

Big Data Management Framework

In this paper, we propose a framework to manage big data under the context of criminal investigation. For big data analytics, some modules are common for every task of a problem domain. For instance, big data extraction, big data transformation, big data integration, big data analysis are among the common modules of a big data analytics system. On the other hand, data sources, methods, and applications are mainly domain-specific. In the proposed framework, we adopt the R-P (relational-positional) model[76] that classifies the activities of criminal investigation under two broad perspectives namely, relation and position. There are four major steps in our proposed framework (see Figure 2). For the first step, data are collected from multiple sources, and then various big data tools and techniques are applied to transform data from a raw format to a suitable format for

TABLE 2 | A Summary of Previous Studies in Data Mining Enabled Security and Criminal Investigation

Study	Technology	Solution Method(s)	Data	Key Contribution	Scalability	Impacts
11	Link analysis	Shortest path algorithm, and heuristic method	Police Records	Develop a criminal investigation prototype system named 'CrimeLink Explorer'	High	Facilitating crime investigation process
64	Neural network	Named-entity extraction method	Police narrative reports	Develop an automatic entity and link extraction technique	Low	Facilitating automatic crime investigation
65	Link analysis	Semantic analysis and Topic analysis	Weblogs data	Social network analysis framework	High	Improve the weblog social network analysis
33	Text mining	Key word extraction algorithm	Crime news	Presenting an integrated term-relationship mining method for crime investigation	High	Offer network mining through criminal news
66	Link analysis	Encryption method	N/A	Develop a privacy protocol	High	Secure information sharing for social media analytics
15	Link analysis	Topic detection	News data(CNN)	Develop an event evaluation system	High	Facilitating large scale event evaluation
67	Machine learning	NLP approach	Wikipedia(English)	Incorporate ML technique for AVD Increase the efficiency of AVD expert system	High	Improve Wiki data authenticity and reliability
5	Link analysis	SNA method	N/A	Presents a tentative protocol for crime data handling and coding	N/A	Enable thorough understanding of criminal behavior
36	Data and text mining method	Logistic regression and decision tree approach	Criminal data	Develop a text-based deception detection technique	Low	Enable a system to detect deception
60	Machine learning	NLP approach	Corpus data	Develop an online reporting system	Low	Increase effectivity and efficiency of crime investigation
68	Text mining	NLP and fuzzy matching	News data	Develop an efficient and effective surveillance system for controlling potential violative market activity	High	Increases efficiency of review and investigation system
37	Text mining and Machine learning	Supervised learning approach	Three datasets (Kongregate, Slashdot and MySpace)	Introduce contextual and similarity features in harassment Detection	N/A	Control harassment in online community
59	Machine learning	Supervised learning method	housebroken data (USA)	Develop a crime pattern detection algorithm	High	Good impact for time management to find crime pattern
56	Machine learning	Bayesian network (BN) model	Structured criminal information	Develop a learning based decision-aid tools	Low	Support police investigations
3	Text mining	Multi-entity Bayesian network model	Heterogeneous criminal data	Discovering crime pattern in large scale datasets	High	Reduce and control organized crime
69	Intelligent Agent	Intelligent agent theory	N/A (conceptual study)	Offer a solid framework for anti-money laundering system practice.	High	Reduce and prevent financial crime

(*continued overleaf*)

TABLE 2 | Continued

Study	Technology	Solution Method(s)	Data	Key Contribution	Scalability	Impacts
70	Intelligent Agent	Monge's detection method[71]	Police records	Develop deceptive criminal detection system	High	Increase efficiency and accuracy in criminal investigation
21	Intelligent agent	Statistical modeling	Respiratory health data	Develop a statistical Process Control system for Early Detection of Bioterrorism	High	Improve public healthcare system
19	Software agent	Game theory	Police records(arrest)	Model security games between a defender and an attacker	High	Improve transportation security and effectivity
72	Intelligent agent and Machine learning	RIPPER learning method	Calling record	Develop a distributed intelligent agent approach for intrusion detection	Low	Improve operating system and network
32	Text mining and data mining	NER model	Textual documents (pdf, e-mail)	Develop criminal community discovery method	High	Improve criminal investigation system
73	Neural network	Adaptive Network models	Standard spreadsheet (structured)	Develop an advanced discriminant system	low	Improve auditing mechanism
45	Link analysis	Shortest-path algorithms, priority-first-search (PFS)	Crime reports	Develop a link analysis technique to reveal strong associations among entities	Low	Discovering new era for investigation.
47	Neural Network and Text mining	Self-organization map (SOM) algorithm and Point pattern analysis method	Orion database	Develop a text mining approach to expand upon the spatial analytical capabilities in criminal investigation	High	Discovering complex behavior of the crime geography
43	Neural network	Inductive and deductive approaches	N/A	Incorporate neural network for psychological criminal profiling	High	Improve criminal investigation process
46	Machine learning and Neural network	Rule base approach	Police reports and housebroken data (Hong Kong)	Develop a prototype system to support police investigation	High	Improving the operation of crime investigation and prevention
74	Machine learning	Unsupervised learning algorithm	Crime and forensic data	Develop a novel data mining approach to assist crime scene investigator(CSI) performance	High	Improve automated investigation mechanism
75	Text mining	Centrality measure	Crime news	Examines the social organization of a hacker community from a network perspective	High	Control hacking activities through revealing hackers' community network
61	Machine learning (fuzzy system)	Fuzzy inference system	Crime data	Develop an expert system for network forensics	High	Enhancing intelligence-based approaches for crime investigation
50	Neural network	Kohonen neural networks, heuristic processing	Crime data	Develop a hybrid Classification system	High	Data itself presents crime pattern (for serial offenders)

subsequent analysis. For the third step, various analytical methods are applied to discover criminal knowledge from the pre-processed data. Finally, the automatically discovered criminal knowledge (e.g., criminal network) is applied to support criminal investigations.

Data Resources

For studying criminal networks, security intelligence agencies collect data from various types of data resources. For the relational analysis, surveillance logs, telephone records, location-based social networks, financial transaction data, and crime incident reports are the main data resources.[2,3] Mostly, to disrupt criminal networks, law enforcement, and intelligence agencies utilize voice calling records, bank accounts, and transaction data.[5,77] Recently, different security intelligence analysis projects have used textual data (e.g., e-mail, SMS messages) to find associations and relations among suspected entities.[78] For the positional analysis, different social media data such as Facebook, Twitter, LinkedIn, and Blogs have been examined. As positional analysis generally only examines the connection pattern among network members, location-specific pattern is not the focus of investigation.[2]

Big Data Transformation, Platform, and Tools

At this stage, all collected data are raw data with heterogeneous format. Accordingly, the raw big data should be transformed to a suitable format for further analysis. Many techniques are available to preprocess raw data. In the criminal investigation domain, various kinds of data can be fed to a big data analytics platform no matter the data are structured or unstructured format. Hadoop is the most remarkable platform for big data analytics. Hadoop is an open source distributed data processing platform that mainly belongs to the 'NOSQL' approaches, which manipulate data without using the classical SQL approach. Hadoop can handle voluminous datasets through the distribution of data to numerous servers (nodes), each of which is accountable for executing a specific task and then synthesizing these intermediate results for the final solution.[79] Though Hadoop is the widely used and effective platform for big data analytics, it is somewhat challenging to install, configure, and administer.[80] Moreover, it is also difficult to find individuals with technique skills for Hadoop. Therefore, organizations may not be ready to embrace to the Hadoop platform. There are a number of surrounding big data ecosystems with additional platforms and

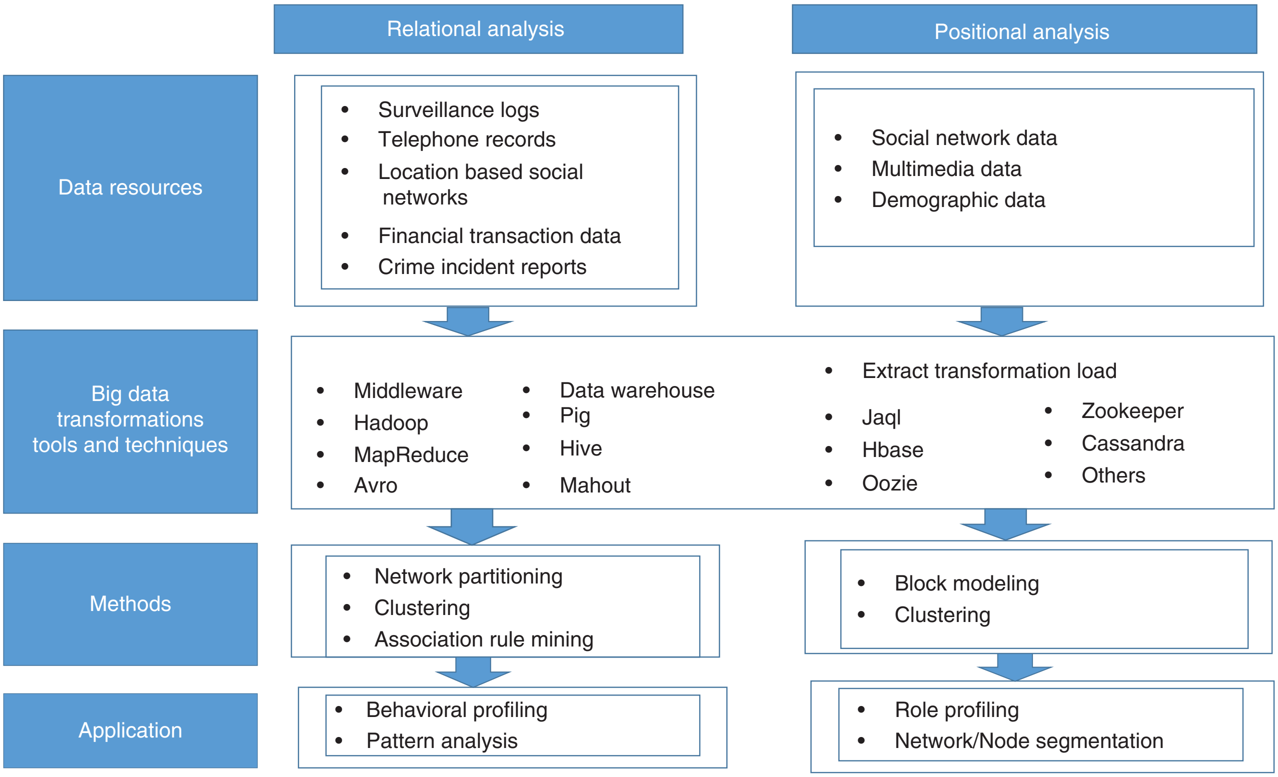


FIGURE 2 | A big data management framework for security and criminal investigation.

TABLE 3 | Tools and Platforms for Big Data Analytics in Security Intelligence

Big data Platforms/Tools	Description
The Hadoop distributed File System (HDFS)	HDFS allows the underlying storage for the Hadoop cluster. It splits the datasets into smaller parts and distributes across different servers.[82]
PIG	Pig programming language is configured to integrate heterogeneous data. PIG comprises two key modules: (1) Language itself called PigLatin, and (2) runtime module, which executes language code.
Hive	Hive is a runtime Hadoop base architecture that leverages Structure Query Language (SQL) with the Hadoop platform.[83]
MapReduce	MapReduce is a programming model that allows distributed processing across many parallel nodes on the similar system. MapReduce provides the interface for the distribution of sub-tasks and the aggregating of outputs.[84]
HBase	HBase is an Apache open-source project which is a distributed, highly scalable, column-oriented database management system.
Cassandra	Cassandra is also a distributed database enabler for managing a large volume of data spread out across many utility servers. Cassandra also offers consistent service without any failure.[85]
Avro	Avro provides record abstraction and data serialization service.[86] It has more features like versioning, version control.
Zookeeper	ZooKeeper enables big data infrastructure since it can serve to coordinate parallel computation of distributed applications.[87]
Lucene	The Lucene project is applied to text analytics/searches and has been incorporated into different open source projects. It provides numerous opportunities including full text indexing and library search for use within a Java application.[88]
Oozie	Oozie is developed from the ground up for large-scale Hadoop workflow. All architectural attributes of Oozie deliver scalability, multitenancy, and effective coordination among sub-systems.[89]
Mahout	Mahout is another Apache project whose focus is to produce free applications of distributed and scalable artificial intelligence algorithms that support big data analytics on the Hadoop platform.

tools.[81] These supporting tools and platforms are summarized in Table 3.

Methods

Security intelligence refers to developing insights from data for criminal network investigation. Data mining techniques can help achieve such a goal through extracting and detecting patterns of criminal networks or forecasting criminal behavior from large-scale datasets. Different data mining methods such as clustering, association rule mining, block modeling, network partitioning, and classification are employed to study organized crimes.[2,3] In this paper, we classify SNA into two classes. One of them is relational analysis where network partitioning, clustering and association rule mining approaches are more suitable because they can estimate different kinds of centralities more effectively. In addition, clustering methods are applied to associate a person with an organization and/or vehicle in criminal investigation.[64] Sequential clustering techniques have been used to discover the preference of computer criminals by Brown and Gunderson.[49] Recently, some neural network models such as Self-Organizing Maps (SOMs) have been applied to this domain. Association rule is a rule that implies a particular link between sets of objects within a dataset, in the form of 'if precedent then incident.'[90] Another class of SNA is positional analysis where block modeling and clustering approaches are more suitable because these two approaches can effectively extract connection patterns among nodes in a network. To discover the overall structure of a network, the key approach is block modeling that includes two steps: network partition and interaction pattern identification.[91] However, some appropriate data mining techniques should be selected based on the underlying data characteristics and business problems.[90]

Applications

Applications for Relational Analysis Purposes

Behavioral profiling is the capacity to extract patterns of criminal activities, to predict the probable time and place of crime to take place, and to identify the different members of a criminal network.[24] Behavioral profiling has been widely used in marketing intelligence to provide *personalization*, that is, making the right offer to the right person at the right place and at the right time.[92,93] By the same token, behavior profiling can be applied to launch the right investigation against the right suspect at the right time, and before the criminal commits a crime. However, behavioral profiling is becoming increasingly challenging because of the 3Vs of big data. For instance, identifying a suspect group is one challenge, and extracting deviations from the norm is another challenge. To design behavioral profiling, it is necessary to analyze huge amount of telephone call data and bank transaction records with a variety of formats.[24] Pattern analyses mainly focuses on three important factors: interaction pattern, relationship pattern, and financial transaction pattern.[94] Pattern analysis defines some standards to identify deviations from groups.

Applications for Positional Analysis Purposes

Role analysis mainly gains insight from the following questions:[5]

> What types of roles does a specific individual or group appear to be playing within a criminal network?

Role analysis faces numerous challenges, especially when a criminal network is dynamic such as members leaving or joining the network; they change their duties and responsibilities, relations may form or dissolve among members, groups may merge or split, and so on.[2] In the literature, role analyses are divided into two models: inductive and deductive. The inductive model relies on generalizing behavioral patterns through statistical analysis of criminal data. In contrast, the deductive model does not rely on generalities from sample groups. The inductive model uses inductive logic that infers from specific individuals to a group, whereas the deductive model uses deductive logic that reasons from a general group pattern back to specific individuals.[95] One model is important for group-role profiling, and another is for individual-role profiling. Chen and Loh developed a system to establish criminal role profiling through combining linguistic rules based on pattern matching and lexical lookup in texts.[42,96] More recently, Westphal proposed some advanced data analytical approaches that apply not only to textual data, but also to consider graphical images and video data.[97] However, to carry out an automatic analysis of multimedia data, it is essential to develop rich computer-based representations of criminal information that is extracted based on positional analysis of perpetrators.

Moreover, it is essential to identify similar groups of criminal networks and their nodes (members) based on the analogous conversation and transaction data in inter or intra networks for the development of effective security measures. The segmentation techniques can be applied to identify various criminal groups who have similar interests. McCue proposed a behavioral segmentation approach for identifying communities within a criminal network.[98] There are different approaches (e.g., hierarchical clustering, network partitioning and matrix permutation approaches) which are widely used for segmenting network.[99] Segmentation is also related to various decision-making processes, and so it is becoming increasingly challenging under the big data environment. Therefore, new computational methods must be developed to cope with the 3Vs of big data.

THE BIG DATA ANALYTICS METHODOLOGY

A plethora of data analytic methodologies have been developed in the security intelligence discipline such as Van der Hulst's methodology,[5] the AMPA methodology,[100] Cross Industry Process for Data Mining (CRISP-DM) methodology, and others. The CRISP-DM methodology is practical and recommended for use in crime prevention and detection.[24] CRISP-DM is an iterative data mining methodology that has been adopted by the Northamptonshire police project.[101] It is also reported that CRISP-DM is the most widely used methodology in several business areas.[102] Figure 3 shows the main stages of our proposed crime investigation methodology which is underpinned by CRISP-DM. Before starting the project, the big data analytics team should develop a feasibility statement in which the project significance, and the trade-offs among costs, benefits, security, and scalability are discussed. After the project feasibility has been examined, the team can proceed to the remaining stages. Step 1 is the business requirement exploration stage where several questions are addressed. The project motivations and project goals are defined at this stage.

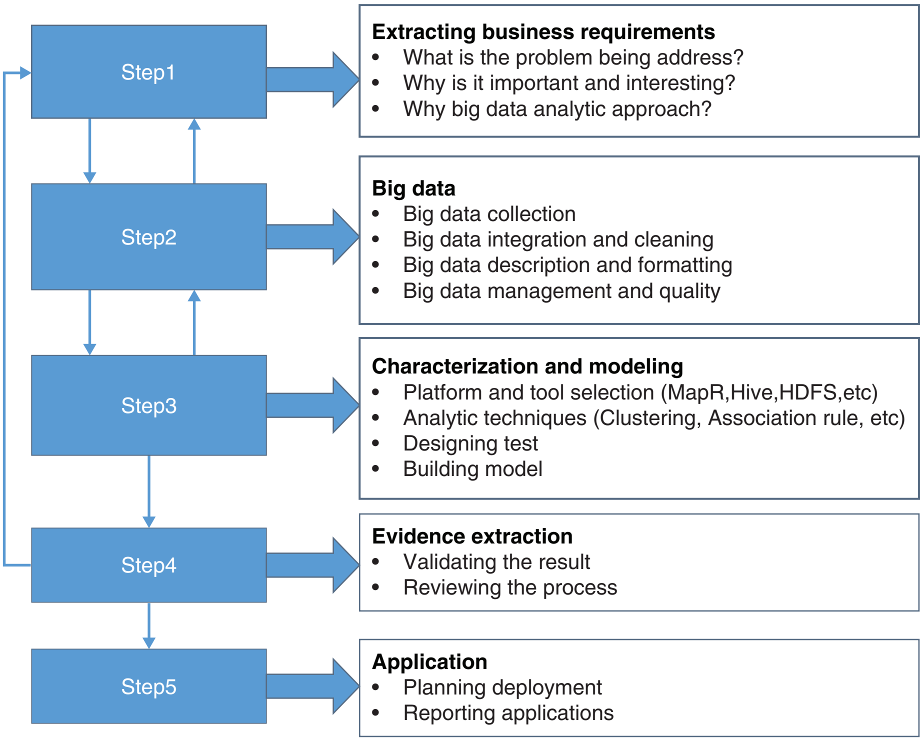
Big data involve three inherent challenges namely, volume, velocity, and variety. As a result, the cost and complexity of big data analytics project are considerably higher than that of traditional analytics projects. Moreover, the project team should study the related approaches of some previous and contemporary projects at this stage. When the business understanding is clear, the team can proceed to step 2 in which the big data sources and the various dimensions of data are examined. During this stage, big data are first collected, described, and transformed to a format suitable for sequent analysis. Then, the team should develop the basic understanding about the structure and contents of the datasets. After the data have been characterized and cleaned-up, they are prepared for subsequent analysis. Although exploring and preparing big data is time-consuming, it is a fundamental step of a data mining methodology.[103] Moreover, apart from data cleaning, data aggregating and the production of training and test data should be conducted at this stage. Then, the team should proceed to step 3 of the proposed methodology if the quality of the pre-processed data is above the desired threshold; otherwise, it must go back to step 1 for further refinement.

In step 3, different big data analytic platforms and tools are chosen to analyze the big data. This step is more challenging, and so much more complicated than other steps. Different big data analytics tools are deployed at this stage. The deployment of big data analytics tools and platforms tends to be complicated, and it usually involves a team of experts for its execution.[79] At this stage, various big data analytics methods are applied to gain insights from the big data. Unlike routine analytical methods, big data analytics methods can scale up with high volume of heterogeneous data. Finally, the analytical models are designed for the big data through a series of iterative what-if analyses. The analytical models and their characteristics are tested and validated in step 4. During the evidence extraction stage, the big data enabled models are refined to improve their accuracy and scalability, as well as their ability to meet the original stated business objectives (e.g., crime investigation).[104] For the final application step, the discovered actionable knowledge is applied to support relevant activities and processes such as potential crime prediction.

In the next section, some real-world big data analytics projects for security intelligence are discussed. We collected these examples from numerous sources including organizations' official Websites.

APPLICATIONS AND CHALLENGES

According to the White House report on big data,[105] top administrators of security intelligence agencies agreed to launch a big data enabled safety and security initiative back to 2012. They have introduced two pilot projects named Neptune and Cerberus, respectively. Neptune mainly performed the security and privacy preservation processes. Its information architecture is designed in such a way that data confidentiality is strongly protected. For strengthening data security, Neptune which belongs to the Sensitive



Extracting business requirements
- What is the problem being address?
- Why is it important and interesting?
- Why big data analytic approach?

Big data
- Big data collection
- Big data integration and cleaning
- Big data description and formatting
- Big data management and quality

Characterization and modeling
- Platform and tool selection (MapR,Hive,HDFS,etc)
- Analytic techniques (Clustering, Association rule, etc)
- Designing test
- Building model

Evidence extraction
- Validating the result
- Reviewing the process

Application
- Planning deployment
- Reporting applications

FIGURE 3 | A methodology of big data analytics enhanced crime investigation.

but Unclassified (SBU) domain, tags the information through assigning both tag names and values of all consumed data. The pilot project of Neptune can ingest, tag, and transfer collected raw data to external systems such as the Electronic System for Travel Authorization (ESTA) system, Student and Exchange Visitor Information System (SEVIS), and Alien Flight Student Program (AFSP) component database systems. Cerberus includes protected cloud computing in big data settings on a Top Secret/Sensitive Compartmented Information (TS/SCI) network to allow real-time computing, analysis, and it has been used by the Department of Homeland Security (DSH) via the Neptune platform. This project is administrated by the Intelligence and Analysis (I&A) office and supervised by the Common Vetting Task Force (CVTF).

UK police and the US defense department have been using CWB developed by a British company named Memex technology limited. They use large-scale textual databases to enhance the criminal data analysis process. They have employed an enhanced search engine with a database intelligent management system. By using big textual data analytics, they cluster and categorize criminal networks and criminals' modus operandi. The LEAP network, developed by the technology community and law enforcement agencies is an information network for analyzing large volumes of heterogeneous data which are shared across security agencies, geographies, and machines. This network provides users with a unique capability to combat different offences that transcend geography and time such as human trafficking, narcotics trafficking, and organized retail crime.

Operation Virtual Shield (OVS) is another example of a big data analytics project to enhance security intelligence discovery. This project is financed by the Homeland Security grants with an amount of $217 million. In Chicago's OVS, it includes at least 2,250 cameras, where 250 of them have biometric equipment and technology. It also enables investigators to identify deceptive movement and harmful activity, and extract events' locations. Moreover, OVS includes facial recognition, which is a computer application for recognizing and validating a person from digital images or video data.[105]

Recently, the Apache project has developed another big data security platform named Metron that can detect cyber anomalies through analyzing big data. The Metron platform enables an organization to store and manipulate huge volume of data. This platform not only identifies cyber anomalies but also enables organizations to constantly monitor their network traffic and activities.

A data mining project was conducted by Wolverhampton University and West Midlands Police (UK), where they used a SOM to predict and prevent sex- and homicide-related crime. The USA, UK, Canada, and South Africa police have used ANNs to identify fraud and money laundering. Their ANN application was empowered by the HNC falcon system.[106] Auto trackXP, DARPA, COPLINK, and DolphinSearch have leveraged data mining techniques to combat different terrorist activities such as bioterrorism, drug enforcement, fraud, and money laundering. The aforementioned technologies have also been adopted by the department of Homeland Security, USA. On the other hand, the UK police department has applied another text mining technology named Copernic for forensic investigation.

Big data analytics for security intelligence should support some key features that are common for analyzing security data. The big data platforms are assessed by some key features such as availability, scalability, privacy, security enablement, ease of access, real-time output, and ability to process diverse levels of granularity.[107,108] As the nature of criminal data has evolved, so are the data analytic approaches that should be scaled up to carry out the more complex and dynamic analytics to address the 3Vs challenges of big data. Criminal datasets have become increasingly more popular represented by unstructured multimedia formats. Heterogeneous nature of the collected criminal data is a challenging issue that makes the investigation process become more complicated and challenging. Recently, another characteristic of big data named veracity have received a lot of attention by researchers and practitioners. Big data sources must be reliable and big data analytics methods should be effective for extracting useful and relevant knowledge from among a mixture of high quality and low quality data. Under the criminal investigation context, veracity is a more challenging issue for the following reasons. First, investigative actions and law-enforcement decisions heavily depend on having the right information (high quality information). Second, the security and well-being of citizens quite depend on the correct and timely law-enforcement actions. To address the above challenges, we identify some general research guidelines to facilitate big data analytics projects for security and criminal investigations.

FUTURE RESEARCH DIRECTIONS

In this study, we identify the data sources, methods, big data platforms, tools, and applications from different security perspectives. We also describe some challenging issues of big data analytics in the context of security and criminal investigation perspectives. We now highlight some future research directions for big data enhanced criminal investigation.

1. Studying the characteristics of different data and the strategies to select the most appropriate data sources for achieving particular investigation goals. As the volume of data is increasing continuously, classical techniques may not be efficient enough to process all available data in a timely manner.[4,105] Thus, the selection of proper data sources is necessary for managing security intelligence.
2. Selection of appropriate analytical models. For big data analytics, there are numerous data analytical methods in which some methods are more suitable for certain investigation purposes and datasets.[1] Consistent with the HACE theorem in Ref 109, big data also introduce some new challenges such as non-uniform data circulation and distributed manipulation with a large number of variables that cannot be coped with by existing analytical methods. Therefore, the project team needs to examine the characteristics of different analytical models, and match the most appropriate methods with the specific nature of the investigation domains.
3. Matching between data mining techniques and methodology. After selecting the analytical methods, project teams should select a suitable methodology to guide the whole investigation work. Some techniques are applicable to some methodologies. For example, the clustering technique is more suitable for the CRISP-DM methodology, but less suitable for the AMPA methodology.[104] Therefore, proper matching between techniques and methodologies are required for effective and efficient investigative actions.
4. Exploring proper data integration methods to tackle complicated investigation problems. Most existing studies on security and criminal investigation use data from a single data source. However, more complicated investigation problems require aggregating data from multiple sources.[3]
5. Tackling the variety of data formats from multiple data sources. Owing to various structures, quality, granularity, and objectives of investigations, data collection and analysis methods tend to be different.[110] Accordingly, project teams should continuously refine their frameworks, analytics techniques, and tools to meet the evolving characteristics of investigations, and the emerging formats of criminal data generating from possibly new data sources.

CONCLUSIONS

The purpose of this review article is to provide researchers and practitioners with a retrospective view of several methodologies and technologies, particularly big data analytics methodology for enhancing security and criminal investigations. We identify five major technologies namely, link analysis, intelligent agents, text mining, ANNs, and ML which have been widely used in various domains for developing the technical foundations of an automated security and criminal investigation system. Under the big data environment, opportunities exist for tapping into big data to enhance security and criminal investigations. However, new methodologies and analytical techniques should be explored to address the fundamental challenges of big data, and hence to leverage big data to facilitate criminal investigations. In short, big data analytics has the potential to transform the way that law enforcement and security intelligence agencies extract vital knowledge (e.g., criminal networks) from multiple data sources in real-time to support their investigations. Such real-time knowledge could help security intelligence agencies to develop comprehensive strategies to prevent and respond to organized crimes such as terrorist attacks and human trafficking. Accordingly, we propose a big data analytics methodology to facilitate security intelligence agencies for criminal investigations. With the continuous advancement of big data analytics techniques, platforms, tools, and methodologies, we will see the widespread utilization of big data across law enforcement and security intelligence agencies in coming few years. The big data revolution leads to the vision that the weapons against crimes, extremism, and terrorism are no longer bullets or bombs, but big data enhanced criminal analytics.

ACKNOWLEDGMENTS

This work is supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Projects: CityU 11502115 and CityU 11525716), the NSFC Basic Research Program (Project: 71671155), and the Shenzhen Municipal Science and Technology R&D Funding - Basic Research Program (Project No. JCYJ20160229165300897).

REFERENCES

1. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big data. The management revolution. *Harvard Bus Rev* 2012, 90:61–67.
2. Xu J, Chen H. Criminal network analysis and visualization. *Commun ACM* 2005, 48:100–107.
3. Chibelushi C, Sharp B, Shah H. ASKARI: a crime text mining approach. *Digital Crime Forensic Sci Cyberspace* 2006, 30:155.
4. Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. In: *System sciences (HICSS), 2013 46th Hawaii International Conference*, IEEE, 17 January, 2013, 995–1004.
5. Van der Hulst RC. Introduction to social network analysis (SNA) as an investigative tool. *Trends Organ Crime* 2009, 12:101–121.
6. Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. *IEEE Internet Comput* 2013, 17:62–69.
7. Han J, Pei J, Kamber M. *Data Mining: Concepts and Techniques*. Amsterdam, Netherlands: Elsevier; 2011.
8. Tufféry S. Data mining and statistics for decision making.
9. Oatley G, Ewart B, Zeleznikow J. Decision support systems for police: lessons from the application of data mining techniques to "soft" forensic evidence. *Artif Intell Law* 2006, 14:35–100.
10. Getoor L. Link mining: a new data mining challenge. *ACM SIGKDD Explor Newslett* 2003, 5:84–89.
11. Schroeder J, Xu J, Chen H, Chau M. Automated criminal link analysis based on domain knowledge. *J Am Soc Inf Sci Technol* 2007, 58:842–855.
12. Gilbert E, Karahalios K. Predicting tie strength with social media. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, ACM, 4 April, 2009, 211–220.
13. Klerks P. The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. *Connections* 2001, 24:53–65.
14. Schroeder J, Xu J, Chen H. Crimelink explorer: Using domain knowledge to facilitate automated crime association analysis. In: *International Conference on Intelligence and Security Informatics*, 2 June, 2003, 168–180. Berlin, Heidelberg: Springer.
15. Yang CC, Shi X, Wei CP. Tracing the event evolution of terror attacks from on-line news. In: *International Conference on Intelligence and Security Informatics*, 23 May, 2006, 343–354. Berlin, Heidelberg: Springer.
16. Kim Y, Choi TY, Yan T, Dooley K. Structural investigation of supply networks: a social network analysis approach. *J Oper Manage* 2011, 29:194–211.
17. Zhao Z, Feng S, Wang Q, Huang JZ, Williams GJ, Fan J. Topic oriented community detection through social objects and link analysis in social networks. *Knowl Syst* 2012, 26:164–173.
18. Brenner W, Zarnekow R, Wittig H. *Intelligent Software Agents: Foundations and Applications*. Berlin, Germany: Springer Science & Business Media; 2012.
19. Jain M, Tsai J, Pita J, Kiekintveld C, Rathi S, Tambe M, Ordónez F. Software assistants for randomized patrol planning for the lax airport police and the federal air marshal service. *Interfaces* 2010, 40:267–290.
20. Bard JF. *Practical Bilevel Optimization: Algorithms and Applications*. Berlin, Germany: Springer Science & Business Media; 2013:9.
21. Fricker RD Jr, Rolka H. Protecting against biological terrorism: statistical issues in electronic biosurveillance. *Chance* 2006, 19:4–14.
22. Jaber A, Guarnieri F, Wybo JL. Intelligent software agents for forest fire prevention and fighting. *Safety Sci* 2001, 39:3–17.
23. Taylor M, Haggerty J, Gresty D, Almond P, Berry T. Forensic investigation of social networking applications. *Netw Security* 2014, 2014:9–16.
24. Mena J. *Investigative Data Mining for Security and Criminal Detection*. Oxford, United Kingdom: Butterworth-Heinemann; 2003.
25. Lin C, Hu PJ, Chen H. Technology implementation management in law enforcement: COPLINK system usability and user acceptance evaluations. *Soc Sci Comput Rev* 2004, 22:24–36.
26. Šišlák D, Pěchouček M, Volf P, Pavlíček D, Samek J, Mařík V, Losiewicz P. AGENTFLY: towards multi-agent technology in free flight air traffic control. In: *Defence Industry Applications of Autonomous Agents and Multi-Agent Systems*, 2007, 73–96. Basel: Birkhäuser.
27. He Q, Sycara K, Su Z. A solution to open standard of PKI. In: *Australasian Conference on Information Security and Privacy*, 13 July, 1998, 99–110. Berlin, Heidelberg: Springer.
28. Hegazy IM, Al-Arif T, Fayed ZT, Faheem HM. A multi-agent based system for intrusion detection. *IEEE Potentials* 2003, 22:28–31.
29. Goldberg HG, Wong RW. Restructuring transactional data for link analysis in the FinCEN AI system. In: *AAAI Fall Symposium*, January 1998, 38–46.
30. Aggarwal CC, Zhai C, eds. *Mining Text Data*. Berlin, Germany: Springer Science & Business Media; 2012.
31. Zhong N, Li Y, Wu ST. Effective pattern discovery for text mining. *IEEE Trans Knowl Data Eng* 2012, 24:30–44.
32. Al-Zaidy R, Fung B, Youssef AM. Towards discovering criminal communities from textual data. In:

Proceedings of the 2011 ACM Symposium on Applied Computing, ACM, 21 March, 2011, 172–177.

33. Tseng YH, Ho ZP, Yang KS, Chen CC. Mining term networks from text collections for crime investigation. *Expert Syst Appl* 2012, 39:10082–10090.
34. Poelmans J, Van Hulle MM, Viaene S, Elzinga P, Dedene G. Text mining with emergent self organizing maps and multi-dimensional scaling: a comparative study on domestic violence. *Appl Soft Comput* 2011, 11:3870–3876.
35. Lee R. Automatic information extraction from documents: A tool for intelligence and law enforcement analysts. In: *Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis*, 23 October, 1998. Menlo Park, CA: AAAI Press.
36. Fuller CM, Biros DP, Delen D. An investigation of data and text mining methods for real world deception detection. *Expert Syst Appl* 2011, 38:8392–8398.
37. Yin D, Xue Z, Hong L, Davison BD, Kontostathis A, Edwards L. Detection of harassment on web 2.0. *Proc Content Anal Web* 2009, 2:1–7.
38. McGreevy MW. Using Perilog to explore "Decision Making at NASA".
39. Connell L. NASA Aviation Safety Reporting System (ASRS).
40. Bovbjerg KM. Personal development under market conditions: NLP and the emergence of an ethics of sensitivity based on the idea of the hidden potential of the individual. *J Contemp Relig* 2011, 26:189–205.
41. Haykin S, Network N. A comprehensive foundation. *Neural Netw* 2004, 2:41.
42. Chen H, Chung W, Xu JJ, Wang G, Qin Y, Chau M. Crime data mining: a general framework and some examples. *Computer* 2004, 37:50–56.
43. Strano M. A neural network applied to criminal psychological profiling: an Italian initiative. *Int J Offender Ther Comp Criminol* 2004, 48:495–503.
44. Chen H, Chung W, Qin Y, Chau M, Xu JJ, Wang G, Zheng R, Atabakhsh H. Crime data mining: an overview and case studies. In: *Proceedings of the 2003 Annual National Conference on Digital Government Research*, Digital Government Society of North America, 18 May, 2003, 1–5.
45. Xu JJ, Chen H. Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. *Decis Support Syst* 2004, 38:473–487.
46. Brahan JW, Lam KP, Chan H, Leung W. AICAMS: artificial intelligence crime analysis and management system. *Knowl Syst* 1998, 11:355–361.
47. Helbich M, Hagenauer J, Leitner M, Edwards R. Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach. *Cartogr Geogr Inf Sci* 2013, 40:326–336.
48. Maren AJ, Harston CT. *Pap RM*. Handbook of Neural Computing Applications: Academic Press; 2014.
49. Brown DE, Gunderson LF. Using clustering to discover the preferences of computer criminals. *IEEE Trans Syst Man Cybern A Syst Hum* 2001, 31:311–318.
50. Dahbur K, Muscarello T. Classification system for serial criminal patterns. *Artif Intell Law* 2003, 11:251–269.
51. Linda O, Vollmer T, Manic M. Neural network based intrusion detection system for critical infrastructures. In: *Neural Networks. IJCNN 2009. International Joint Conference*, IEEE 14 June, 2009, 1827–1834).
52. Elder JF, Abbott DW. A comparison of leading data mining tools. In: *Fourth International Conference on Knowledge Discovery and Data Mining*, 28 August, 1998.
53. Veltkamp RC, Tanase M. Content-based image retrieval systems: a survey.
54. Mitchell TM. *Machine Learning*. Burr Ridge, IL: McGraw Hill; 1997.
55. Russell SJ, Norvig P, Canny JF, Malik JM, Edwards DD. *Artificial Intelligence: A Modern Approach*, vol. 2. Upper Saddle River, NJ: Prentice Hall; 2003.
56. Baumgartner K, Ferrari S, Palermo G. Constructing Bayesian networks for criminal profiling from limited data. *Knowl Syst* 2008, 21:563–572.
57. Franke K, Srihari SN. Computational forensics: an overview. In: *International Workshop on Computational Forensics*, 7 August 2008, 1–10. Berlin, Heidelberg: Springer.
58. Nath SV. Crime pattern detection using data mining. In: 2006 *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, WI-IAT 2006 Workshops, IEEE, December 2006, 41–44.
59. Wang T, Rudin C, Wagner D, Sevieri R. Learning to detect patterns of crime. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 23 September, 2013, 515–530. Berlin, Heidelberg: Springer.
60. Ku CH, Iriberri A, Leroy G. Crime information extraction from police and witness narrative reports. In: *2008 I.E. Conference on Technologies for Homeland Security*, IEEE, 12 May, 2008, 193–198.
61. Stoffel K, Cotofrei P, Han D. Fuzzy methods for forensic data analysis. In: *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, IEEE, 7 December, 2010, 23–28.
62. Gupta P, Sharma A, Jindal R. Scalable machine-learning algorithms for big data analytics: a comprehensive review. *WIREs Data Min Knowl Discov* 2016, 6:194–214.
63. Van Halteren H, Baayen H, Tweedie F, Haverkort M, Neijt A. New machine learning methods demonstrate

the existence of a human stylome. *J Quant Linguist* 2005, 12:65–77.

64. Chau M, Xu JJ, Chen H. Extracting meaningful entities from police narrative reports. In: *Proceedings of the 2002 Annual National Conference on Digital Government Research*, Digital Government Society of North America, 19 May, 2002, 1–5.
65. Yang CC, Ng TD. Terrorism and crime related weblog social network: link, content analysis and information visualization. In: *Intelligence and Security Informatics*, IEEE, 23 May, 2007 May 23, 55–58.
66. Kerschbaum F, Schaad A. Privacy-preserving social network analysis for criminal investigations. In: *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, ACM, 27 October, 2008, 9–14.
67. Smets K, Goethals B, Verdonk B. Automatic vandalism detection in Wikipedia: towards a machine learning approach. In: *AAAI workshop on wikipedia and artificial intelligence: an evolving synergy*, 13 July, 2008, 43–48.
68. Goldberg HG, Kirkland JD, Lee D, Shyr P, Thakker D. The NASD Securities Observation, New Analysis and Regulation System (SONAR). In: *IAAI*, 12 August, 2003, 11–18.
69. Gao S, Xu D. Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering. *Expert Syst Appl* 2009, 36:1493–1504.
70. Wang GA, Chen H, Xu JJ, Atabakhsh H. Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Trans Syst Man Cybern A Syst Hum* 2006, 36:988–999.
71. Monge AE. Adaptive detection of approximately duplicate database records and the database integration approach to information discovery. Doctoral dissertation, *University of California, San Diego*, 1997.
72. Helmer GG, Wong JS, Honavar V, Miller L. Intelligent agents for intrusion detection. In: *Information Technology Conference*, IEEE, 1 September, 1998, 121–124.
73. Fanning K, Cogger KO, Srivastava R. Detection of management fraud: a neural network approach. *ISAFM* 1995, 4:113–126.
74. Adderley R, Townsley M, Bond J. Use of data mining techniques to model crime scene investigator performance. *Knowl Syst* 2007, 20:170–176.
75. Lu Y, Luo X, Polgar M, Cao Y. Social network analysis of a criminal hacker community. *J Comput Inform Syst* 2010, 51:31–41.
76. Burt RS. Models of network structure. *Annu Rev Sociol* 1980, 6:79–141.
77. Pramanik MI, Zhang W, Lau RY, Li C. A framework for criminal network analysis using big data. In: *13th International Conference on e-Business Engineering (ICEBE)*, IEEE, 4 November, 2016, 17–23.
78. Popp R, Armour T, Numrych K. Countering terrorism through information technology. *Commun ACM* 2004, 47:36–43.
79. Zikopoulos P, Eaton C. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw-Hill Osborne Media; 2011. Chapter 3.
80. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014, 2:3.
81. Zikopoulos P, Parasuraman K, Deutsch T, Giles J, Corrigan D. Harness the Power of big data. The IBM Big Data Platform; McGraw Hill Professional 2012.
82. Borthakur D. HDFS architecture guide. Hadoop Apache Project; 2008: 53.
83. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R. Hive: a warehousing solution over a map-reduce framework. *Proc VLDB Endowment* 2009, 2:1626–1629.
84. Fernández A, del Río S, López V, Bawakid A, del Jesús MJ, Benítez JM, Herrera F. Big data with cloud computing: an insight on the computing environment, MapReduce, and programming frameworks. *WIREs Data Min Knowl Discov* 2014, 4:380–409.
85. Lakshman A, Malik P. Cassandra: a decentralized structured storage system. *ACM SIGOPS Operat Syst Rev* 2010, 44:35–40.
86. Floratou A, Patel JM, Shekita EJ, Tata S. Column-oriented storage techniques for MapReduce. *Proc VLDB Endowment* 2011, 4:419–429.
87. Hunt P, Konar M, Junqueira FP, Reed B. ZooKeeper: Wait-free Coordination for Internet-scale Systems. In: *USENIX annual Technical Conference*, 23 June, 2010; 8: 9.
88. McCandless M, Hatcher E, Gospodnetic O. *Lucene in Action: Covers Apache Lucene 3.0*. Greenwich, CT: Manning Publications Co; 2010.
89. Islam M, Huang AK, Battisha M, Chiang M, Srinivasan S, Peters C, Neumann A, Abdelnur A. Oozie: towards a scalable workflow management system for Hadoop. In: *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*, ACM, 20 May, 2012, 4.
90. Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications–a decade review from 2000 to 2011. *Expert Syst Appl* 2012, 39:11303–11311.
91. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *J Assoc Inf Sci Technol* 2007, 58:1019–1031.
92. Abbasoğlu MA, Gedik B, Ferhatosmanoğlu H. Aggregate profile clustering for telco analytics. *Proc VLDB Endowment* 2013, 6:1234–1237.
93. Fan S, Lau RY, Zhao JL. Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Res* 2015, 2:28–32.

94. Sarvari H, Abozinadah E, Mbaziira A, McCoy D. Constructing and analyzing criminal networks. In: *Security and Privacy Workshops (SPW)*, IEEE, 2014, 84–91.

95. Rogers M. The role of criminal profiling in the computer forensics process. *Comput Secur* 2003, 22:292–298.

96. Loh S, Wives LK, de Oliveira JP. Concept-based knowledge discovery in texts extracted from the web. *ACM SIGKDD Explor Newslett* 2000, 2:29–39.

97. Westphal C. *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. Boca Raton, FL: CRC Press; 2008.

98. McCue C. *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*. Oxford, United Kingdom: Butterworth-Heinemann; 2014.

99. Drewes B. Integration of text and data mining. *WIT Trans Inform Commun Technol* 2002, 10:28.

100. McCue C. *Data Mining and Predictive Analysis*. Oxford: Elsevier, Butterworth-Heinemann; 2007.

101. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0 Step-by-step data mining guide.

102. Giraud-Carrier C, Povel O. Characterising data mining software. *Intell Data Anal* 2003, 7:181–192.

103. Sherman R. Data integration advisor: set the stage with data preparation. *DM Rev* 2005, 15: 54–55.

104. Ozgul F, Atzenbeck C, Celik A, Erdem Z. Incorporating data sources and methodologies for crime data mining. In: *IEEE International Conference on Intelligence and Security Informatics (ISI)*, IEEE, 2011, 176–180.

105. House W. Big data: seizing opportunities, preserving values (Report for the President), 2014. Washington, DC: Executive Office of the President. [WWW document]. Available at: http://www. whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

106. Piatetsky-Shapiro G. Knowledge discovery in databases: 10 years after. *ACM SIGKDD Explor Newslett* 2000, 1:59–61.

107. Bollier D, Firestone CM. *The promise and peril of big data*. Washington, DC: Aspen Institute, Communications and Society Program; 2010.

108. Ramírez-Gallego S, García S, Mouriño-Talín H, Martínez-Rego D, Bolón-Canedo V, Alonso-Betanzos A, Benítez JM, Herrera F. Data discretization: taxonomy and big data challenge. *WIREs Data Min Knowl Discov* 2016, 6:5–21.

109. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. *IEEE Trans Knowl Data Eng* 2014, 26:97–107.

110. Sagiroglu S, Sinanc D. Big data: a review. In: *International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, 2013, 42–47.