International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

The Role of Data Mining in Information Security

Osman Abbas[a*], Dr. Mohamed Elhafiz Mustafa[b], Dr. Siddig Balal Ibrahim[c]

[a] *PhD Candidate , Sudan University of Science and Technology, Sudan*
[b] *College of Computer Science and Information Technology, Sudan University of Science and Technology, Sudan.*
[c] *College of Business Studies, Sudan University of Science and Technology, Sudan .*
[a]*Email: osman.sd@live.com*

Abstract

Security and Privacy protection have been a public policy concern for decades. However, rapid technological changes, the rapid growth of the internet and electronic commerce, and the development of more sophisticated methods of collecting, analyzing, and using personal information have made privacy a major public and government issues. The field of data mining is gaining significance recognition to the availability of large amounts of data, easily collected and stored via computer systems. Recently, the large amount of data, gathered from various channels, contains much personal information. When personal and sensitive data are published and/or analyzed, one important question to take into account is whether the analysis violates the privacy of individuals whose data is referred to. The importance of information that can be used to increase revenue cuts costs or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data privacy is growing constantly. For this reason, many research works have focused on privacy-preserving data mining, proposing novel techniques that allow extracting knowledge while trying to protect the privacy of users. Some of these approaches aim at individual privacy while others aim at corporate privacy [1].

Keywords: Data Mining; Information security; Threats; Privacy.

1. Introduction

Data mining for cyber security applications For example, anomaly detection techniques could be used to detect unusual patterns and behaviors. Link analysis may be used to trace the viruses to the perpetrators. Classification may be used to group various cyber-attacks and then use the profiles to detect an attack when it occurs.

---

* Corresponding author.
E-mail address: osman.sd@live.com.

Prediction may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations. Data mining is also being applied for intrusion detection and auditing [2]. The conventional approach to securing computer systems against cyber threats is to design mechanisms such as firewalls, authentication tools, and virtual private networks that create a protective shield. However, these mechanisms almost always have vulnerabilities. They cannot ward attacks that are continually being adapted to exploit system weaknesses, which are often caused by careless design and implementation flaws. This has created the need for intrusion detection, security technology that complements conventional security approaches by monitoring systems and identifying computer attacks. Traditional intrusion detection methods are based on human experts extensive Knowledge of attack signatures which are character strings in a messages payload that indicate malicious content. Signatures have several limitations. They cannot detect novel attacks, because someone must manually revise the signature database beforehand for each new type of intrusion discovered. Once someone discovers a new attack and develops its signature, deploying that signature is often delayed. These Limitations have led to an increasing interest in intrusion detection techniques based on data mining [3].

Figure 1: Data Mining Applications

2. Data Mining

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations [4].

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides [5].

Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, heath care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed [6].

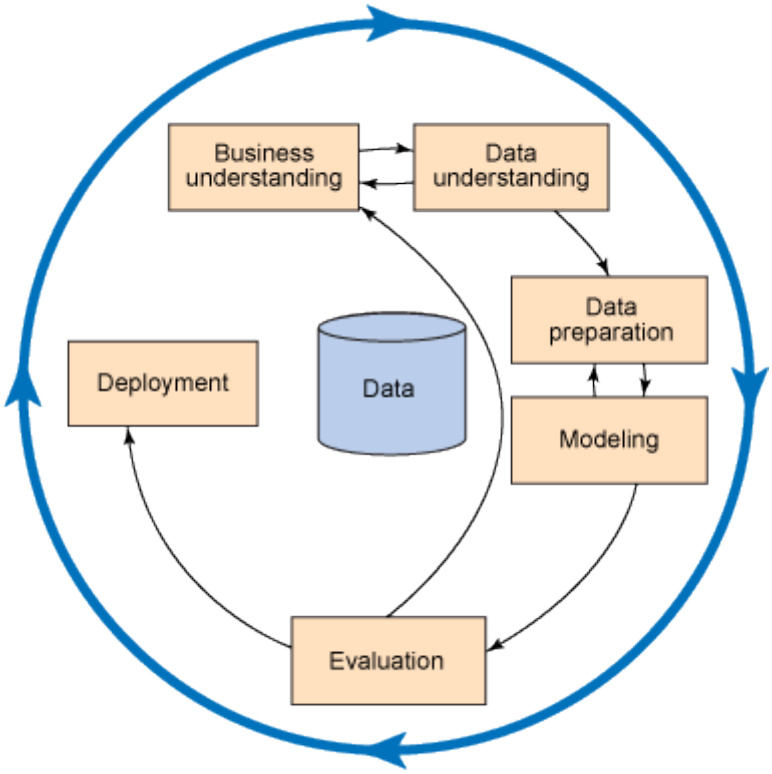

Figure 2: Data mining Process

3. Data warehouse

Data Warehouse (DW) is a system that extracts, cleans, confirms and source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making. Sophisticated OLAP and Data Mining tools are used to facilitate multinational analysis and complex business models. Inmon W.H defines the Data Warehouse as a subject oriented, integrated, time variant andnon-volatile collection of data in support of management's decision making process [7]. BI applications in enterprises provide reports for the strategic management of business by collaborating the business data and electronic data interchange. This ensures competitive intelligence and thereby helps in good decision making [8]. According to B de Ville, BI refers to the technologies and application for collecting, storing and analyzing business data that helps the enterprise to make better decisions [9].

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. The past three years have seen explosive growth, both in the number of products and services offered, and in the adoption of these technologies by industry. According to the META Group, the data warehousing market, including hardware, database software, and tools, is projected to grow from $2 billion in 1995 to $8 billion in 1998. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). This paper presents a roadmap of data warehousing technologies, focusing on the special requirements that data warehouses place on database management systems (DBMSs) [10].

Data warehousing is a phenomenon that grew from the huge amount of electronic data stored in recent years and from the urgent need to use that data to accomplish goals that go beyond the routine tasks linked to daily processing. In a typical scenario, a large corporation has many branches, and senior managers need to quantify and evaluate how each branch contributes to the global business performance. The corporate database stores detailed data on the tasks performed by branches. To meet the managers' needs, tailor-made queries can be issued to retrieve the required data. In order for this process to work, database administrators must first formulate the desired query (typically an aggregate SQL query) after closely studying database catalogs. Then the query is processed. This can take a few hours because of the huge amount of data, the query complexity, and the concurrent effects of other regular workload queries on data. Finally, a report is generated and passed to senior managers in the form of a spreadsheet. Many years ago, database designers realized that such an approach is hardly feasible, because it is very demanding in terms of time and resources, and it does not always achieve the desired results. Moreover, a mix of analytical queries with transactional routine queries inevitably slows down the system, and this does not meet the needs of users of either type of query. Today's advanced data warehousing processes separate online analytical processing (OLAP) from online transactional processing (OLTP) by creating a new information repository that integrates basic data from various sources, properly arranges data formats, and then makes data available for analysis and evaluation aimed at planning and decision-making processes [11].

4. Data mining vs Data warehouse

The terms data mining and data warehousing are often confused by both business and technical staff. The entire field of data management has experienced a phenomenal growth with the implementation of data collection software programs and the decreased cost of computer memory. The primary purpose behind both these functions is to provide the tools and methodologies to explore the patterns and meaning in large amount of data [12].

The primary differences between data mining and data warehousing are the system designs, methodology used, and the purpose. Data mining is the use of pattern recognition logic to identity trends within a sample data set and extrapolate this information against the larger data pool. Data warehousing is the process of extracting and

storing data to allow easier reporting.

Data mining is a general term used to describe a range of business processes that derive patterns from data. Typically, a statistical analysis software package is used to identify specific patterns, based on the data set and queries generated by the end user. A typical use of data mining is to create targeted marketing programs, identify financial fraud, and to flag unusual patterns in behavior as part of a security review [13].

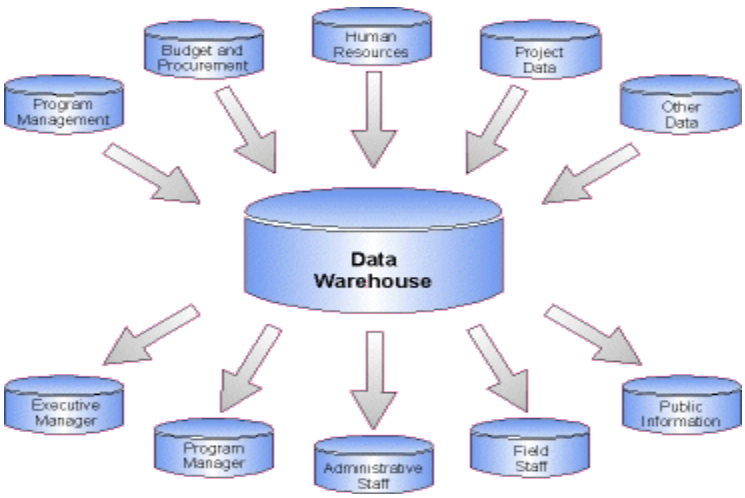

Figure 3: Data warehouse

An excellent example of data mining is the process used by telephone companies to market products to existing customers. The telephone company uses data mining software to access its database of customer information. A query is written to identify customers who have subscribed to the basic phone package and the Internet service over a specific time frame. Once this data set is selected, another query is written to determine how many of these customers took advantage of free additional phone features during a trial promotion. The results of this data mining exercise reveal patterns of behavior that can drive or help refine a marketing plan to increase the use of additional telephone services [14].

It is important to note that the primary purpose of data mining is to spot patterns in the data. The specifications used to define the sample set have a huge impact on the relevance of the output and the accuracy of the analysis. Returning to the example above, if the data set is limited to customers within a specific geographical area, the results and patterns will differ from a broader data set. Although both data mining and data warehousing work with large volumes of information, the processes used are quite different [15].

A data warehouse is a software product that is used to store large volumes of data and run specifically designed queries and reports. Business intelligence is a growing field of study that focuses on data warehousing and related functionality. These tools are designed to extract data and store it in a method designed to provide enhanced system performance. Much of the terminology in data mining and data warehousing are the same, leading to more confusion [16].

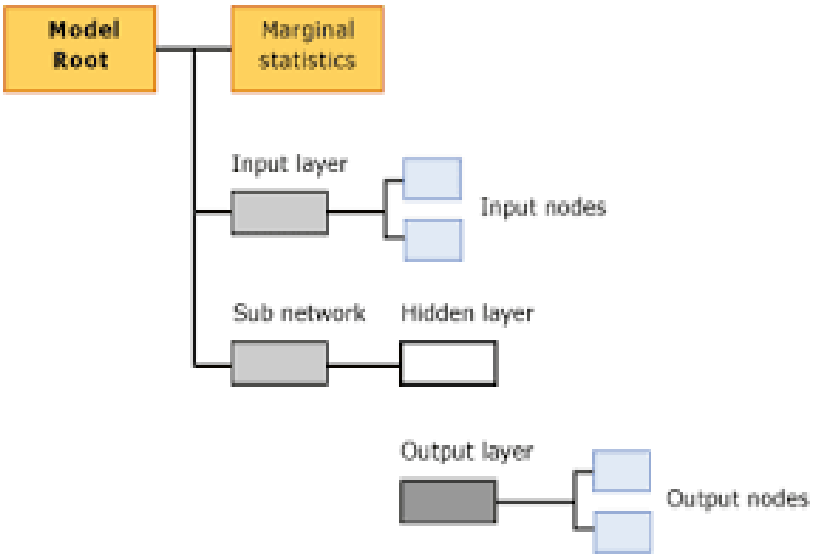

Figure 4: Data mining Structure

5. Data mining security

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences. Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered. Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive [17]. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations [18,19,20].

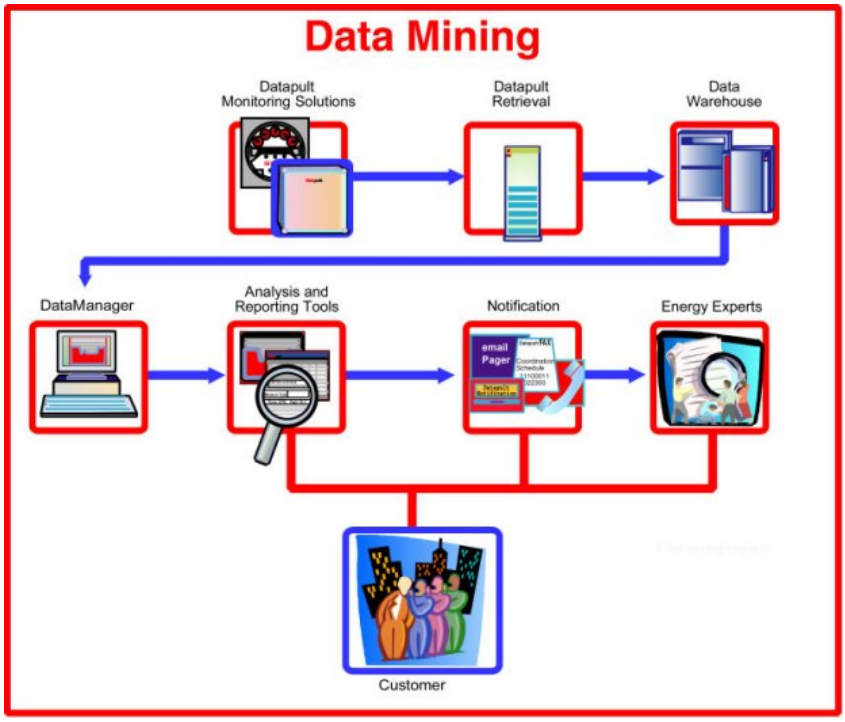

Figure 5: data Mining Security

6. Data Quality

Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data. To improve data quality, it is sometimes necessary to "clean" the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database (e.g., ensuring that "no" is represented as a 0 throughout the database, and not sometimes as a 0, sometimes as an N, etc.), accounting for missing data points, removing unneeded data fields, identifying anomalous data points (e.g., an individual whose age is shown as 142 years), and standardizing data formats (e.g., changing dates so they all include MM/DD/YYYY) [21] .

7. Privacy

As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy. Concerns about privacy focus both on actual projects proposed, as well as concerns about the potential for data mining applications to be expanded beyond their original purposes. For example, some experts suggest that anti-terrorism data mining applications might also be useful for combating other types of crime as well [22]. So far there has been little consensus about how data mining should be carried out, with several competing points of view being debated. Some observers contend that tradeoffs may need to be

made regarding privacy to ensure security. Other observers suggest that existing laws and regulations regarding privacy protections are adequate, and that these initiatives do not pose any threats to privacy. Still other observers argue that not enough is known about how data mining projects will be carried out, and that greater oversight is needed. There is also some disagreement over how privacy concerns should be addressed. Some observers suggest that technical solutions are adequate initiatives [23,24].

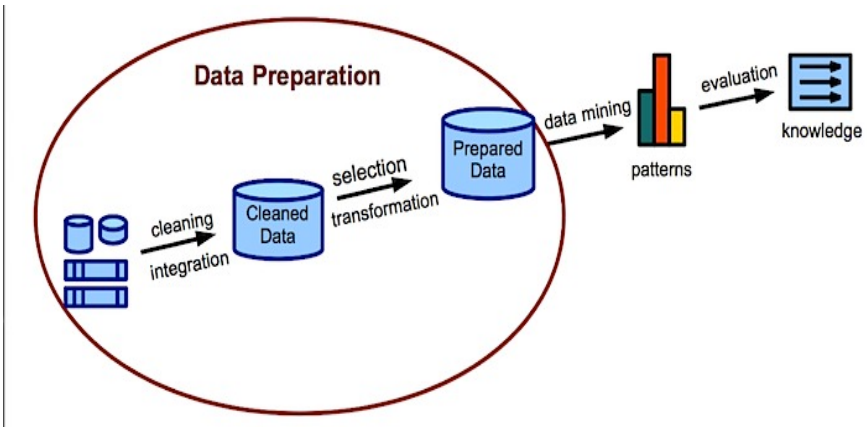

Figure 6: Data Mining Quality

Data mining has attracted significant interest especially in the past decade with its vast domain of applications. From the security perspective, data mining has been shown to be beneficial in confronting various types of attacks to computer systems. However, the same technology can be used to create potential security hazards. In addition to that, data collection and analysis efforts by government agencies and businesses raised fears about privacy, which motivated the privacy preserving data mining research. One aspect of privacy preserving data mining is that, we should be able to apply data mining algorithms without observing the confidential data values [25,26]. This challenging task is still being investigated. Another aspect is that, using data mining technology an adversary could access confidential information that could not be reached through querying tools jeopardizing the privacy of individuals. Some initial research results in privacy preserving data mining have been published. However, there are still many issues that need further investigation in the context of data mining from both privacy and security perspectives. This workshop aims to provide a meeting place for academicians to identify problems related to all aspects of privacy and security issues in data mining together with possible solutions. Researchers and practitioners working in data mining, databases, data security, and statistics are invited to submit their experience, and/or research results [27].

8. Data collection

If a data collector wants to collect data from data providers who place high value on their private data, the collector may need to negotiate with the providers about the ‘‘price’’ of the sensitive data and the level of privacy protection. In [28] the authors build a sequential game model to analyze the private data collection process. In the proposed model, a data user, who wants to buy a data set from the data collector, makes a price offer to the collector at the beginning of the game. If the data collector accepts the offer, he then announces some incentives to data providers in order to collect private data from them. Before selling the collected data to

the data user, the data collector applies anonymization technique to the data, in order to protect the privacy of data providers at certain level. Knowing that data will be anonymized, the data user asks for a privacy protection level that facilitates his most preferable balance between data quality and quantity when making his offer. The data collector also announces a specific privacy protection level to data providers. Based on the protection level and incentives offered by data collector, a data provider decides whether to provide his data. In this data collection game, the level of privacy protection has significant influence on each player's action and pay- off. Usually, the data collector and data user have different expectations on the protection level. By solving the sub game perfect Nash equilibriums of the proposed game, a consensus on the level of privacy protection can be achieved.

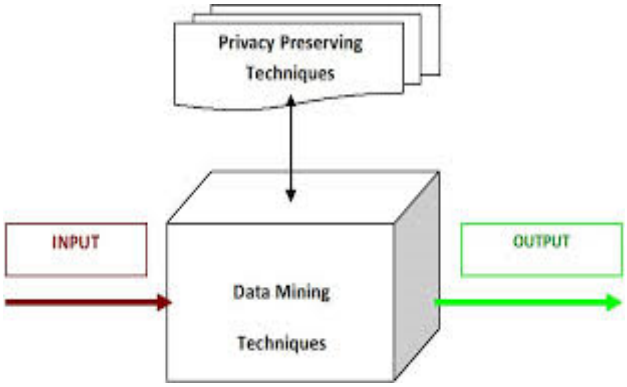

Figure 7: Data Mining Privacy

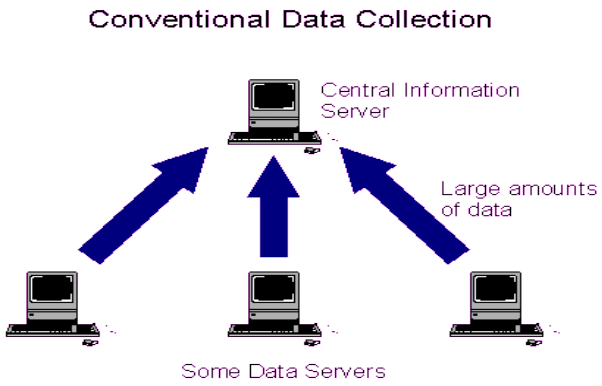

Figure 8 : Data collection

In their later work [29] the authors propose a similar game theoretical approach for aggregate query applications. They show that stable combinations of revelation level (how specific data are revealed), retention period of the collected data, price of per data item, and the incentives offered to data providers, can be found by solving the game's equilibriums. The game analysis has some implications on how to set a privacy policy to achieve maximum revenue while respecting data providers' privacy preferences. And the proposed game model can be potentially used for comparing different privacy protection approaches [30].

8. Critical Infrastructures

Attacks on critical infrastructures could cripple a nation and its economy. Infrastructure attacks include attacking the telecommunication lines, the electric, power, gas, reservoirs and water sup-plies, food supplies and other basic entities that are critical for the operation of a nation. Attacks on critical infrastructures could occur during any type of attack whether they are non-information related, information related or bio-terrorism attacks. For example, one could attack the software that runs the telecommunications industry and close down all the telecommunication lines. Similarly, software that runs the power and gas supplies could be attacked. Attacks could also occur through bombs and explosives. That is, the telecommunication lines could be physically attacked. Attacking transportation lines such as highways and railway tracks are also attacks on infrastructures. Infrastructures could also be attacked by natural disaster such as hurricanes and earth quakes. Our main interest here is the attacks on infrastructures through malicious attacks, both information related and non-information related. Our goal is to examine data mining and related data management technologies to detect and prevent such infrastructure attacks [31].

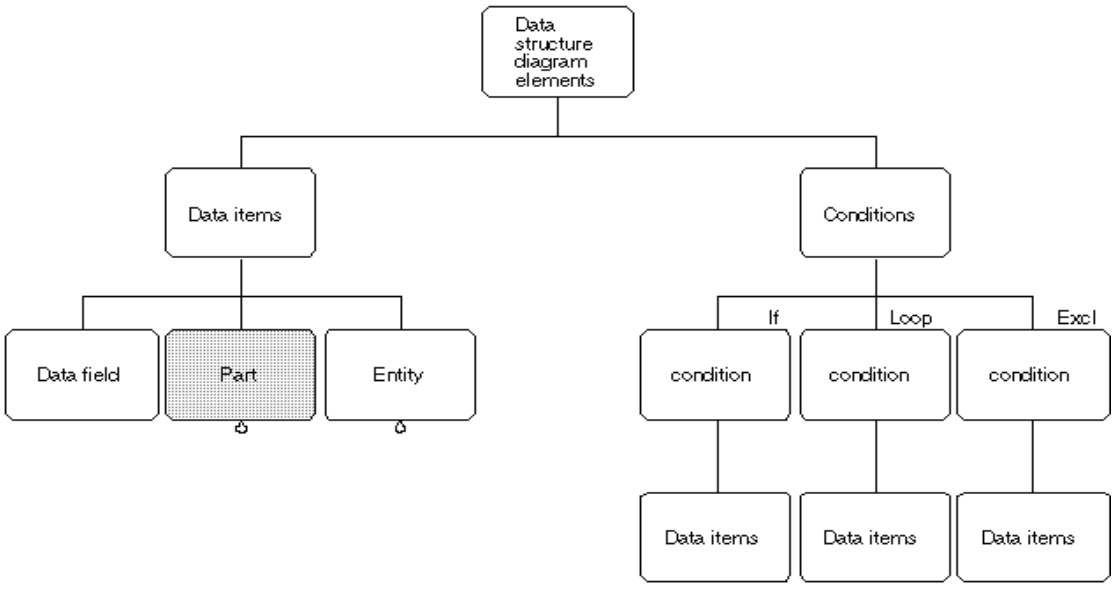

Figure 9: Data Structure

8. Data collector

Data collector collects data from data providers in order to support the subsequent data mining operations. The original data collected from data providers usually contain sensitive information about individuals. If the data collector doesn't take sufficient precautions before releasing the data to public or data miners, those sensitive information may be disclosed, even though this is not the collector's original intention. For example, on October 2, 2006, the U.S. online movie rental service Netflix14 released a data set containing movie ratings of 500,000

subscribers to the public for a challenging competition called ''the Netflix Prize". The goal of the competition was to improve the accuracy of personalized movie recommendations. The released data set was supposed to be privacy-safe, since each data record only contained a subscriber ID (irrelevant with the subscriber's real identity), the movie info, the rating, and the date on which the subscriber rated the movie. However, soon after the release, two researchers [32] from University of Texas found that with a little bit of auxiliary information about an individual subscriber, e.g. 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, an adversary can easily identify the individual's record (if the record is present in the data set). From above example we can see that, it is necessary for the data collector to modify the original data before releas- ing them to others, so that sensitive information about data providers can neither be found in the modified data nor be inferred by anyone with malicious intent. Generally, the modification will cause a loss in data utility. The data collector should also make sure that sufficient utility of the data can be retained after the modification, otherwise collect- ing the data will be a wasted effort. The data modification process adopted by data collector, with the goal of preserving privacy and utility simultaneously, is usually called privacy preserving data publishing (PPDP). Extensive approaches to PPDP have been proposed in last decade. Reference [33] have systematically summarized and evaluated different approaches in their frequently cited survey. Also, Wong and Fu have made a detailed review of studies on PPDP in their monograph [34]. To differentiate with their work, in this paper we mainly focus on how PPDP is realized in two emerging applications, namely social networks and location-based services. To make our review more self-contained, in next subsection we will first briefly introduce some basic PPDP, e.g. the privacy model, typical anonymization operations, information metrics, etc, and then we will review studies on social networks and location-based services respectively.

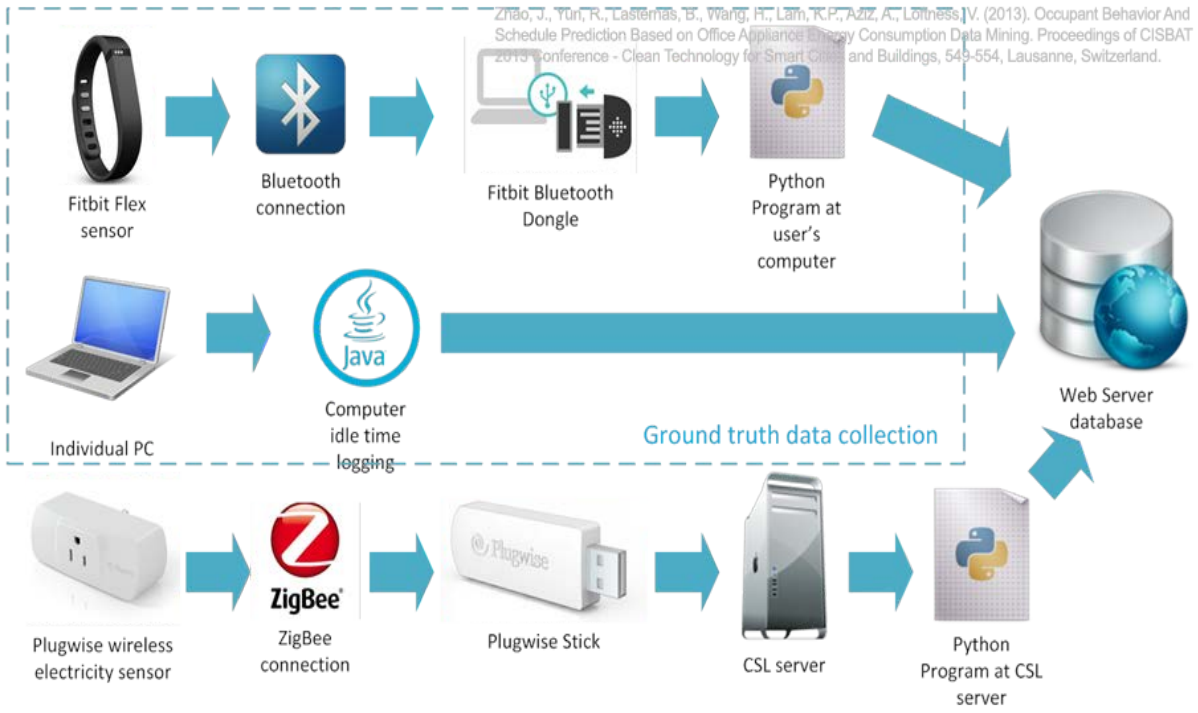

Figure 10: Data sources

8. Identity Theft

We are hearing a lot these days about credit card fraud and identity theft. In the case of credit card fraud, an attacker obtains a person's credit card and uses it to make unauthorized purchases. By the time the owner of the card becomes aware of the fraud, it may be too late to reverse the damage or apprehend the culprit. A similar problem occurs with telephone calling cards. In fact this type of attack has happened to me personally. Perhaps while I was making phone calls using my calling card at airports someone noticed the dial tones and reproduced them to make free calls. This was my company calling card. Fortunately our telephone company detected the problem and informed my company. The problem was dealt with immediately. A more serious theft is identity theft. Here one assumes the identity of another person by acquiring key personal information such as social security number, and uses that information to carry out transactions under the other person's name. Even a single such transaction, such as selling a house and depositing the income in a fraudulent bank account, can have devastating consequences for the victim. By the time the owner finds out it will be far too late. It is very likely that the owner may have lost millions of dollars due to the identity theft. We need to explore the use of data mining both for credit card fraud detection as well as for identity theft. There have been some efforts on detecting credit card fraud. We need to start working actively on detecting and preventing identity thefts [34].

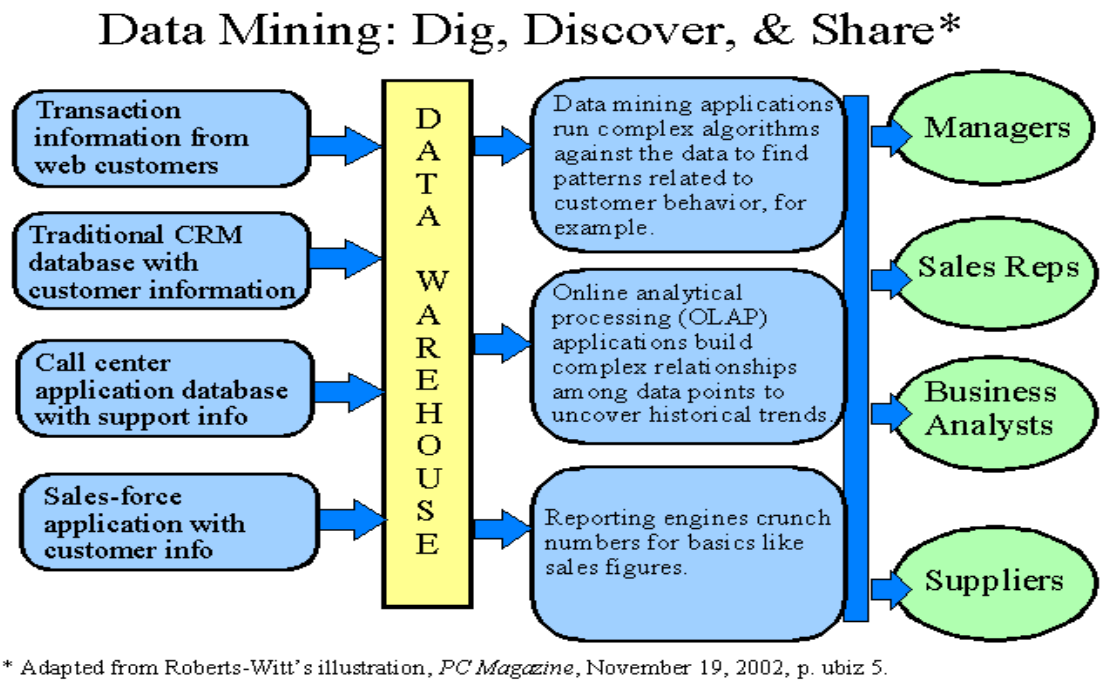

Figure 11: Data Mining Mechanism

6. Data Mining and terrorism

While data mining products can be very powerful tools, they are not self- sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel-related, rather than technology-related [35].

Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user.

Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use. However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations) [36].

Beyond these specific limitations, some researchers suggest that the circumstances surrounding our knowledge of terrorism make data mining an ill- suited tool for identifying (predicting) potential terrorists before an activity occurs. Successful "predictive data mining" requires a significant number of known instances of a particular behavior in order to develop valid predictive models. For example, data mining used to predict types of consumer behavior (i.e., the likelihood of someone shopping at a particular store, the potential of a credit card usage being fraudulent) may be based on as many as millions of previous instances of the same particular behavior. Moreover, such a robust data set can still lead to false positives. In contrast, as a CATO Institute report suggests that the relatively small number of 10 Jeff Jonas and Jim Harper, Effective Counterterrorism and the Limited Role of Predictive Data Mining, CATO Institute Policy Analysis No. 584, December 11, 2006 p. 8, [http://www.cato.org/pubs/pas/pa584.pdf]. 11 Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Third Edition (Potomac, MD: Two Crows Corporation, 1999), p. 5; Patrick Dillon, Data Mining: Transforming Business Data Into Competitive Advantage and Intellectual Capital (Atlanta GA: The Information Management Forum, 1998), pp. 5-6. 12 George Cahlink, "Data Mining Taps the Trends," Government Executive Magazine, October 1, 2000, [http://www.govexec.com/tech/articles/1000managetech. htm]. For a more detailed review of the purpose for data mining conducted by federal departments and agencies, see U.S. General Accounting Office, Data Mining: Federal (continued...) terrorist incidents or attempts each year are too few and individually unique "to enable the creation of valid predictive models "[37].

10. Privacy Protection

Limit the Access: A data source provider provides his data to the receiver in an active way or a passive way. By "active" we mean that the data source provider willingly opts in a survey initiated by the data receiver, or fill in some registration forms to create an account in a website. By "passive" we mean that the data, which are

generated by the source provider's routine activities, are recorded by the data receiver, while the data source provider may even have no awareness of the revealing of his data. When the data source provider provides his data actively, he can simply ignore the receiver's demand for the facts that he consider very sensitive. If his data are passively provided to the data receiver, the data source provider can take some measures to limit the receiver's access to his sensitive data. Also, the data source provider can utilize various security tools that are developed for Internet environment to protect his data. Many of the security tools are designed as browser extensions for ease of use [38].

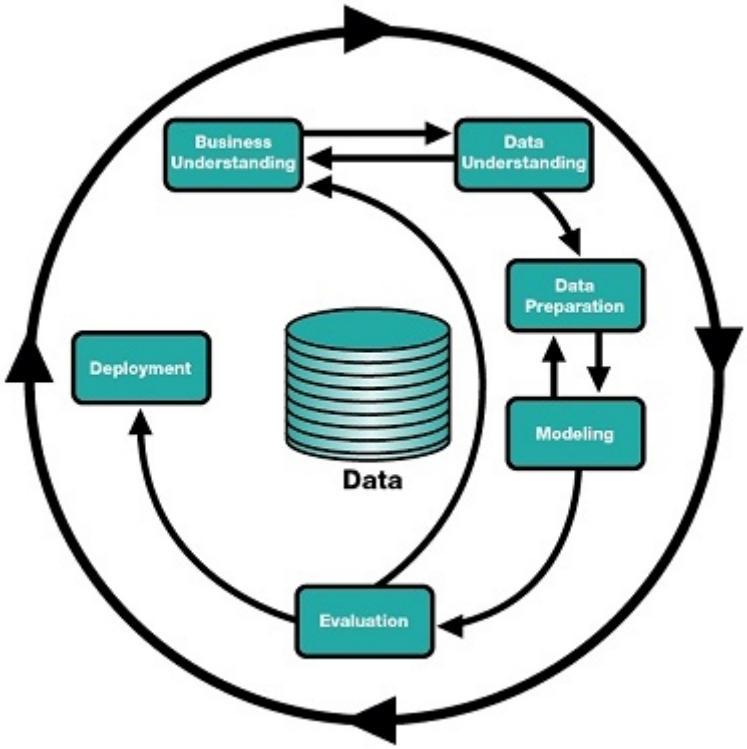

Figure 12: Role of Data Mining

Anti-tracking extensions: Knowing that valuable information can be removed from the data produced by user's online activities, Internet companies have a strong motivation to track the user's movements on the Internet. When browsing the Internet, a user can utilize an anti-tracking extension to block the trackers from collecting the cookies. Popular anti-tracking extensions include Disconnect, Do Not Track Me4, Ghostery, etc. A major technology used for anti- tracking is called Do Not Track (DNT), which enables users to opt out of tracking by websites they do not visit. A user's opt-out preference is signaled by an HTTP header field named DNT: if DNT=1, it means the user does not want to be tracked (opt out). Two U.S. researchers first created a prototype add on supporting DNT header for the Firefox web browser in 2009. Later, many web browsers have added support for DNT.DNT is not only a technology but also a policy framework for how companies that receive the signal should respond. TheW3C Tracking Protection Working Group is now trying to standardize how websites should response to user's DNT request [39]. Advertisement and script blockers: This type of browser extensions

can block advertisements on the sites, and kill scripts and widgets that send the user's data to some unknown third party. Example tools include Ad Block Plus6, NoScript7, FlashBlock8, etc. Encryption Tools: To make sure a private online communication between two parties cannot be intercepted by third parties, a user can utilize encryption tools, such as MailCloak9and TorChat10, to encrypt his emails, instant messages, or other types of web traffic. Also, a user can encrypt all of his internet traffic by using a VPN (virtual private network) service. There is no guarantee that one's sensitive data can be completely kept out of the reach of treacherous data collectors, making it a habit of clearing online traces and using security tools does can help to reduce the risk of privacy disclosure. Trade Privacy for Benefit: In some cases, the data source provider needs to make a trade-off between the loss of privacy and the benefits brought by participating in data mining. For example, by analyzing a user's demographic information and browsing history, a shopping website can offer personalized product recommendations to the user. The user's sensitive preference may be disclosed but he can enjoy a better shopping experience. Driven by some benefits, e.g. a personalized service or monetary incentives, the data source provider may be willing to provide his sensitive data to a reliable data receiver, who promises the provider's sensitive information will not be revealed to an unauthorized third-party. If the provider is able to predict how much benefit he can get, he can logically decide what kind of and how many sensitive data to provide. For example, suppose a receiver asks the data source provider to provide information about his age, gender, occupation and annual salary. And the receiver tells the data source provider how much he would pay for each data item. If the data source provider considers salary to be his sensitive information, then based on the prices offered by the receiver, he chooses one of the following actions: i) not to report his salary, if he thinks the price is too low; ii) to report a fuzzy value of his salary, e.g. "less than 10,000 dollars", if he thinks the price is just acceptable; iii) to report an accurate value of his salary, if he thinks the price is high enough. For this example we can see that, both the privacy preference of data source provider and the incentives offered by data receiver will affect the data source provider's decision on his sensitive data. On the other hand, the data receiver can make profit from the data collected from data source providers, and the profit heavily depends on the quantity and quality of the data. Hence, data source providers' privacy preferences have great influence on data receiver's profit. The profit plays an important role when data receiver decides the incentives. That is to say, data receiver's decision on incentives is related to data source provider's privacy preferences. Therefore, if the data source provider wants to obtain satisfying benefits by "selling" his data to the data receiver, he needs to consider the effect of his decision on data receiver's benefits (even the data explorer's benefits), which will in turn affects the benefits he can get from the receiver. In the data-selling scenario, both the seller (i.e. the data source provider)and the buyer (i.e. the data receiver) want to get more benefits, thus the interaction between data source provider and data receiver can be formally analyzed by using game theory. Also, the sale of data can be treated as an auction, where mechanism design theory can be applied[7]. Provide False Data: As discussed above, a data source provider can take some measures to prevent data receiver from accessing his sensitive data [39].

However, a disappointed fact that we have to admit is that no matter how hard they try, Internet users cannot completely stop the unwanted access to their persona information. So instead of trying to limit the access, the data source provider can provide false information to those treacherous data receiver. The following three methods can help an Internet user to falsify his data: Using "sock puppets" to hide one's true activities. A sock

puppet is a false online identity though which a member of an Internet community speaks while pretending to be another person, like a puppeteer manipulating a hand puppet. By using multiple sock puppets, the data produced by one individual's activities will be deemed as data belonging to different individuals, assuming that the data receiver does not have enough knowledge to relate different sock puppets to one specific individual. Using a fake identity to create phony information. In 2012, Apple Inc. was assigned a patient called "Techniques to pollute electronic profiling" which can help to protect user's privacy [40]. This patent discloses a method for polluting the information gathered by "network eavesdroppers" by making a false online identity of a principal agent, e.g. a service subscriber. A browser extension called Mask Me, which was release by the online privacy company Abine, Inc. in 2013, can help the user to create and manage aliases (or Masks) of these personal information. Users can use these aliases just like they normally do when such information is required, while the websites cannot get the real information. In this way, user's privacy is protected.

Figure 13: Data Circle

11. Conclusion

Data mining has become one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. In the context of homeland security, data mining can be a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. While data mining represents a significant advance in the type of analytical tools currently available, there are limitations to its capability. One limitation is that although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. A second limitation is that while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. Successful data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output. Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and

waste, but have grown to also be used for purposes such as measuring and improving program performance. However, some of the homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Some efforts that have attracted a higher level of congressional interest include the Terrorism .

As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of a project's outcome. One issue is data quality, which refers to the accuracy and completeness of the data being analyzed. A second issue is the interoperability of the data mining software and databases being used by different agencies. A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected. As data miners, our tasks are colliding with these concerns. In analytic customer relationship management (CRM), we often analyze customer data with the specific intent of understanding individual behavior and instituting sales campaigns based on this understanding. Researchers in economics, demographics, medicine and social sciences are trying to understand the relationships between behaviors and outcomes. Both privacy and security are politically popular areas of concern, with growing public awareness and activism in the U.S., Europe, and in many other countries. Therefore, the temptation to legislate and regulate to protect the public may outweigh the consequences of restricting both online and offline commerce. On the other hand, the burden is on business to show where federal legislation is necessary to enhance electronic commerce, with clear benefits and consumer protections. Finally, elected and public officials should be informed of the costs and consequences to consumers, businesses, and the economy of legislative or regulatory proposals to protect privacy and security [40].

References

[1] Z. Ferdousi, A. Maeda, "Unsupervised outlier detection in time series data", 22nd International Conference on Data Engineering Workshops, pp. 51-56, 2006

[2] S. A. Demurjian and J. E. Dobson, "Database Security IX Status and Prospects Edited by D. L. Spooner ISBN 0 412 72920 2, 1996, pp. 391- 399.

[3] L. Getoor, C. P. Diehl. "Link mining: a survey", ACM SIGKDD Explorations, vol. 7, pp. 3-12, 2005.

[4] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques. Morgan kaufmann, 2006.

[5] R. Agrawal and R. Srikant, "Privacy-preserving data mining,"SIGMOD Rec., vol. 29, no. 2, pp. 439– 450, 2000.

[6] Y. Lindell and B. Pinkas,"Privacy preserving data mining,"in Advances in Cryptology CRYPTO 2000. Springer, 2000, pp. 36–54.

[7] P. Bergeron, C. A. Hiller, (2002), "Competitive intelligence", in B.Cronin, Annual Review of nformation Science and Technology,zedford, N.J.: Information Today, vol. 36, chapter 8

[8] V. Ciriani, S. D. C. Di Vimercati, S. Forest, and P. Samarati,"Microdata protection,"in Secure Data Management in Decentralized Systems. Springer, 2007, pp. 291–321.

[9] R. T. Fielding and D. Singer, "Tracking preference expression (dnt).w3c working draft," 2014.[Online]. Available: http ://www.w3.org/TR/2014/WD- tracking-dnt– 20140128

[10] D. C. Parkes, "Classic mechanism design," Iterative Combinatorial Auctions: Achieving Economic and Computational Efficiency. Ph. D. dissertation, University of Pennsylvania, 2001.

[11] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preservingdata publishing: A survey of recent developments,"ACM Computing Surveys (CSUR), vol. 42, no. 4, p. 14, 2010.

[12] L. Moreau,"The foundations for provenance on the web," Foundations and Trends in Web Science, vol. 2, no. 2–3, pp. 99–241, 2010.

[13] G. Barbier, Z. Feng, P. Gundecha, and H. Liu, "Provenance data in social media," Synthesis Lectures on Data Mining and Knowledge Discovery, vol. 4, no. 1, pp. 1–84, 2013.

[14] M. Tudjman and N. Mikelic, "Information science: Science about information, misinformation and disinformation,"Proceedings of Informing Science+ Information Technology Education, pp. 1513–1527, 2003.

[15] M. J. Metzger,"Making sense of credibility on the web: Models for evaluating online information and recommendationsfor futureresearch,"Journal of the American Society for Information Science and Technology, vol. 58, no. 13, pp. 2078–2091, 2007.

[16] J. Vaidya, H. Yu, and X. Jiang, "Privacy- preserving svm classification,"Knowledge and Information Systems, vol. 14, no. 2, pp. 161–178, 2008.

[17] Z. Ferdousi, A. Maeda, "Unsupervised outlier detection in time series data", 22nd International Conference on Data Engineering Workshops, pp. 51-56, 2006

[18] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[19] Fung B., Wang K., Yu P. "Top-Down Specialization for Information and Privacy Preservation. ICDE Conference, 2005.

[20] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[21] O. Tene and J. Polenetsky, ''To track or 'do not track': Advancing transparency and individual control in

online behavioral advertising,'' Minnesota J. Law, Sci. Technol., no. 1, pp. 281–357, 2012.

[22] R. C.-W. Wong and A. W.-C. Fu, ''Privacy-preserving data publish- ing: An overview,'' Synthesis Lectures Data Manage., vol. 2, no. 1, pp. 1–138, 2010.

[23] Y. Xua, X. Qin, Z. Yang, Y. Yang, and K. Li, ''A personalized k-anonymity privacy preserving method,'' J. Inf. Comput. Sci., vol. 10, no. 1, pp. 139–155, 2013.

[24] S. Yang, L. Lijie, Z. Jianpei, and Y. Jing, ''Method for individualized privacy preservation,'' Int. J. Secur. Appl., vol. 7, no. 6, p. 109, 2013. [52] A. Halevy, A. Rajaraman, and J. Ordille, ''Data integration: The teenage years,'' in Proc. 32nd Int. Conf. Very Large Data Bases (VLDB), 2006, pp. 9–16.

[25] pp. 9–16. [53] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, ''State-of-the-art in privacy preserving data mining,'' ACM SIGMOD Rec., vol. 33, no. 1, pp. 50–57, 2004.

[26] V. S. Verykios, ''Association rule hiding methods,'' Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery, vol. 3, no. 1, pp. 28–36, 2013.

[27] K. Sathiyapriya and G. S. Sadasivam, ''A survey on privacy preserving association rule mining,'' Int. J. Data Mining Knowl. Manage. Process, vol. 3, no. 2, p. 119, 2013.

[28] R. K. Adl, M. Askari, K. Barker, and R. Safavi-Naini, ''Privacy consensus in anonymization systems via game theory,'' in Proc. 26th Annu. Data Appl. Security Privacy, 2012, pp. 74–89.

[29] R. Karimi Adl, K. Barker, and J. Denzinger, ''A negotiation game: Establishing stable privacy policies for aggregate reasoning,'' Dept. Comput. Sci., Univ. Calgary, Calgary, AB, Canada, Tech. Rep., Oct. 2012. [Online]. Available: The paper is available at http:// dspace.ucalgary.ca/jspui/bitstream/1880/49282/1/2012-1023-06.pdf

[30] A. Miyaji and M. S. Rahman, ''Privacy-preserving data mining: A game-theoretic approach,'' in Proc. 25th Data Appl. Security Privacy, 2011, pp. 186–200.

[31] X. Ge, L. Yan, J. Zhu, and W. Shi, ''Privacy-preserving distributed association rule mining based on the secret sharing technique,'' in Proc. 2nd Int. Conf. Softw. Eng. Data Mining (SEDM), Jun. 2010, pp. 345–350.

[32] A. Narayanan and V. Shmatikov, ''Robust de-anonymization of large sparse datasets,'' in Proc. IEEE Symp. Secur. Privacy (SP), May 2008, pp. 111–125.

[33] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, ''Privacy-preserving data publishing: A survey of recent developments,'' ACM Comput. Surv., vol. 42, no. 4, Jun. 2010, Art. ID 14.

[34] R. C.-W. Wong and A. W.-C. Fu, ''Privacy-preserving data publish- ing: An overview,'' Synthesis Lectures Data Manage., vol. 2, no. 1, pp. 1–138, 2010.

[35] C. Dwork, ‘‘Differential privacy,’’ in Proc. 33rd Int. Conf. Autom., Lang., Program., 2006, pp. 1–12.

[36] H. Xia, Y. Fu, J. Zhou, and Y. Fang, ‘‘Privacy-preserving SVM classifier with hyperbolic tangent kernel,’’ J. Comput. Inf. Syst., vol. 6, no. 5, pp. 1415–1420, 2010.

[37] S. Jha, L. Kruger, and P. McDaniel, ‘‘Privacy preserving clustering,’’ in Proc. 10th Eur. Symp. Res. Comput. Security (ESORICS), 2005, pp. 397–417.

[38] L. K. Fleischer and Y.-H. Lyu, ‘‘Approximately optimal auctions for selling privacy when costs are correlated with data,’’ in Proc. 13th ACM Conf. Electron. Commerce, 2012, pp. 568–585.

[39] H. Kargupta, K. Das, and K. Liu, ‘‘Multi-party, privacy-preserving distributed data mining using a game theoretic framework,’’ in Proc. 11th Eur. Conf. Principles Pract. Knowl. Discovery Databases (PKDD), 2007, pp. 523–531.

[40] R. K. Adl, M. Askari, K. Barker, and R. Safavi-Naini, ‘‘Privacy consensus in anonymization systems via game theory,’’ in Proc. 26th Annu. Data Appl. Security Privacy, 2012, pp. 74–89.