AN ENERGY EFFICIENT DATA SHARING USING INFORMATION CENTRIC APPROACH WITH INFORMATION SECURITY IN BIG DATA

ARUNUDAYA R[1], AFREEN BANU E[2]

1. ME CSE Student, Vel Tech Multitech Engineering College, Chennai
2. Assistant Professor, Vel Tech Multitech Engineering College, Chennai,

ABSTRACT-Big data strongly demands a network infrastructure having the capability to efficiently collect, process, cache, share, and deliver the data, instead of simple transmissions. The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. In this paper, we view the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. For each type of user, we discuss his privacy concerns and the methods that can be adopted to protect sensitive information. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering.

INDEX TERMS- Big data, information centric network, CCN, energy-efficiency, sensitive information, privacy-preserving data mining

I. INTRODUCTION

Big data are generally generated by and collected from geographically distributed devices, and stored in data warehouses and processed in powerful data centers with massive interconnected servers. Its applications face challenges in acquiring, storing, processing, sharing, transmitting, analysing and visualizing data with very large quantities. In this paper, we focus on the network designs for big data sharing. For IoT services, the data generated from vast amount of sensors are collected, stored, processed, visualized and delivered to the users. On the other hand, the Internet is originally designed for end-to-end communications, where the networks serve as the data transmission pipes that connect data sources, data center services for these applications. Therefore, it is time to re-consider the network infrastructure design for data sharing applications in the era of big data. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. We summarize four design requirements as follows

A. ENERGY-EFFICIENCY

The network should reduce redundant and duplicate traffic to optimize energy consumptions in data transmissions. It also should enable the data to be retrieved from the closest data copy holder. For the current implementation with data centers , the same data need to be delivered from data center, which might be far away, to a set of users one by one, which brings out large duplicate traffic overhead.

B. AVAILABILITY

The network should be enabled to provide the services to users over heterogeneous networks regardless of network scale or malfunctions.

C. HIGH-PERFORMANCE

The network should provide services with low latency and high-throughput, especially in case of the delay-sensitive applications.

D. DATA-AWARE INTELLIGENCE

The network should be aware of the characteristics of the data in transmissions for the potential in-network processing. Thus, the network computing resources can compensate for network transmission resources to make the communication model more scalable and efficient.

Information Centric Networking (ICN) approach to design network architecture for content sharing in the big data. In ICN, routers are possibly equipped with cache memories to cache data. Through the in-network caching, the current end-to-end transmission pipe for big data will break into

the many small broken pipes with mini-stops for transmissions. These mini-stops can cache the data in the transmissions for further fast data retrieval, whereas they do not bring out much longer delay. Because of in-network caching, the duplicate transmissions from the data centers to the users and further energy consumptions can be significantly reduced. The high availability can also be achieved, since the same data is not only stored in the data centers but cached in the networking nodes. Thus, the users can retrieve data from the close copy holder instead of the data center far away. Meanwhile, the desired data can be moved to the targeted users beforehand, which enable users to experience good quality on low latency and high-throughput. In the ICN, data names, rather than server IP addresses, become the handles of the requests and replies for the routers. Hence, ICN approaches can achieve data-aware intelligence based on the data names.

D.THE PROCESS OF KDD

The term ``data mining" is often treated as a synonym for another term ``knowledge discovery from data" (KDD) which highlights the goal of the mining process. To obtain useful knowledge from data, the following steps are performed in an iterative way (see Fig. 1):

- Step 1: Data pre-processing. Basic operations include data selection (to retrieve data relevant to the KDD task from the database), data cleaning (to remove noise and inconsistent data, to handle the missing data fields, etc.) and data integration (to combine data from multiple sources).
- Step 2: Data transformation. The goal is to transform data into forms appropriate for the mining task, that is, to find useful features to represent the data. Feature selection and feature transformation are basic operations.
- Step 3: Data mining. This is an essential process where intelligent methods are employed to extract data patterns.
- Step 4: Pattern evaluation and presentation. Basic operations include identifying the truly interesting patterns which represent knowledge, and presenting the mined knowledge in an easy-to-understand fashion.

FIGURE1. AnoverviewoftheKDDprocess

E. THE PRIVACY CONCERN AND PPDM

Individual's privacy may be violated due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc. The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The consideration of PPDM is two-fold. First, sensitive raw data, such as individual's ID card number and cell phone number, should not be directly used for mining. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded.

F. USER ROLE-BASED METHODOLOGY

Current models and algorithms proposed for PPDM mainly focus on how to hide that sensitive information from certain mining operations. However, as depicted in Fig. 1, the whole KDD process involves multi-phase operations. Besides the mining phase, privacy issues may also arise in the phase of data collecting or data pre-processing, even in the delivery process of the mining results. In this paper, we investigate the privacy aspects of data mining by considering the whole knowledge-discovery process. We present an overview of the many approaches which can help to make proper use of sensitive data and protect the security of sensitive information discovered by data mining. We use the term ``sensitive information" to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. The term ``sensitive data" refers to data from which sensitive information can be extracted. Throughout the paper, we consider the two terms ``privacy" and ``sensitive information" are interchangeable. Based on the stage division in KDD process (see Fig. 1), we can identify four different types of users, namely four user roles, in a typical data mining scenario (see Fig. 2):

- Data Provider: the user who owns some data that are desired by the data mining task.
- Data Collector: the users who collects data from data providers and then publish the data to the data miner.

-
-
- Data Miner: the user who performs data mining tasks on the data.
- Decision Maker: the user who makes decisions based on the data mining results in order to achieve certain goals.

FIGURE 2. A simple illustration of the application scenario with data mining at the core

II DATA MODULES

1) DATA PROVIDER

The major concern of a data provider is whether he can control the sensitivity of the data he provides to others. On one hand, the provider should be able to make his very private data, namely the data containing information that he does not want anyone else to know, inaccessible to the data collector. On the other hand, if the provider has to provide one data to the data collector, he wants to hide his sensitive information as much as possible and get enough compensation for the possible loss in privacy.

A.CONCERNS OF DATA PROVIDER

A data provider owns some data from which valuable information can be extracted. In the data mining scenario depicted in Fig.2, there are actually two types of data providers: one refers to the data provider who provides data to data collector, and the other refers to the data collector who provides data to data miner. To differentiate the privacy protecting methods adopted by different user roles, here in this section, were strict ourselves to the ordinary data provider, the one who owns are natively small amount of data which contain only information about himself. Toinvestigatethemeasuresthat the data provider can a dopt to protect privacy, we consider the following three situations:

1) If the data provider considers his data's very sensitive, that is, the data may reveal some information that he does not want anyone else to know, the provider can just refuse to provide such data. Effective access-control measures are desired by the data provider, so that he can prevent his sensitive data from being stolen by the data collector.

2) Realizing that his data are valuable to the data collector (as well as the data miner), the data provider may be willing to handover some of his private data in exchange for certain benefit, such as better services or monetary rewards.

3) If the data provider can neither prevent the access to his sensitive data no make creative deal with the data collector, the data provider can distort his data that will be fetched by the data collector, so that his true information cannot be easily disclosed.

B. APPROACHES TO PRIVACY PROTECTION

1) LIMIT THE ACCESS

A data provider provides his data to the collector in an active way or a passive way. By ``active'' we mean that the data provider voluntarily opts in a survey initiated by the data collector, 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 provider's routine activities, are recorded by the data collector, while the data provider may even have no awareness of the disclosure of his data. When the data provider provides his data creatively, he can simply ignore the collector's demand for the information that he deems very sensitive. If his data are passively provided to the data collector, the data provider can takes one measures to limit the collector's access to his sensitive data.

Current security tools can be categorized into the following three types:

1) Anti-tracking extensions. Knowing that valuable information can be extracted from the data produce by users' online activities, Internet companies have strong motivation to track the users' movements on the Internet. When browsing the Internet, a user can utilize a tracking extension to block the trackers from collecting the cookies. Popular anti-tracking extensions include Disconnect, Do Not Track Me, Ghostery, etc.

2) 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 Plus, No Script, Flash Block, etc.

3) Encryption tools. To make sure a private online communication between parties cannot be intercepted by third parties, a user can utilize encryption tools, such as Mail Cloak and Tor Chat, to encrypt his emails, instant messages, or other types of web traffic. Also, a user can encrypt his entire internet Traffic By using a VPN (virtual private network) service.

2) PROVIDE FALSE DATA

As discussed above, a data provider can take some measures to prevent data collector from accessing his sensitive data. 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 personal information. The following three methods can help an Internet user to falsify his data:

1)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 pre-tending 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 collector does not have enough knowledge to relate different sock puppets to one specific individual. As a result, the user's true activities are unknown to others and his sensitive information (e.g. political preference) cannot be easily discovered.

2) Using a fake identity to create phony information. When a network leaves dropper collects the data of a user who is utilizing this method, the leaves dropper will be interfered by them passive data created by the clone identity. Real information about of the user is buried under the manufactured phony information.

3) Using security tools to mask one's identity. When a user signs up for a web service or buys something online, he is often asked to provide information such as email address, credit card number, phone number, etc.

3) DATA COLLECTOR

The data collected from data providers may contain individuals' sensitive information. Directly releasing the data to the data miner will violate data providers' privacy, hence data modification is required. On the other hand, the data should still be useful after modification; otherwise collecting the data will be meaningless. Therefore, the major concern of data collector is to guarantee that the modified data contain no sensitive information but still preserve high utility.

A. CONCERNS OF DATA COLLECTOR

As shown in Fig. 2, a 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. 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 collecting 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).

B.APPROACHES TO PRIVACY PROTECTION

1) BASICS OF PPDP

PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes:

_Identifier (ID): Attributes that can directly and uniquely identify an individual, such as name, ID number and mobile number.

_Quasi-identifier (QID): Attributes that can be linked with external data tore-identify individual records, such as gender, age and zip code.

_Sensitive Attribute (SA): Attributes that an individual wants to conceal, such as disease and salary.

_Non-sensitive Attribute (NSA): Attributes other than ID, QID and SA.

2) PRIVACY-PRESERVING PUBLISHING OF SOCIAL NETWORK DATA

To support the analysis, the company who runs a social network application sometimes needs to publish data to a third party.

3) ATTACK MODEL

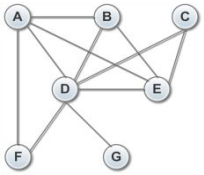

FIGURE 4. Example of mutual friend attack: (a) original network; (b) native anonymized network.

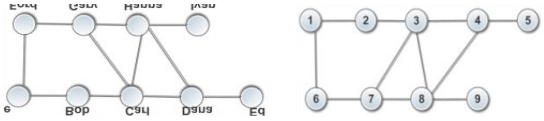

FIGURE 5. Example of friend attack: (a) original network; (b) native anonymized network.

The attack uses the cumulative degree of *n*-hop neighbours of a vertex as the regional feature, and combine sit with the simulate annealing-based graph matching method tore-identify vertices in anonymous social graphs. Sunetal introduce a relationship attack model called *mutual friend*

attack, which is based on the number of mutual friends of two connected individuals. Fig. 4 shows an example of the mutual friend attack. The original social network *G* with vertex identities is shown in Fig.4 (a), and Fig.4 (b) shows the corresponding anonymized network where all individuals' names are removed. In, Taietal. Investigate the *friendship attack* where an adversary utilizes the degrees of two vertices connected by an edge tore-identify related victims in published social network data set. Fig. 5 shows an example of friendship attack. In, another type of attack, namely *degree attack, is* explored. The motivation is that each individual in a social network is inclined to associate with not only a vertex identity but also a community identity, and the community identity reflects some sensitive information about the individual.

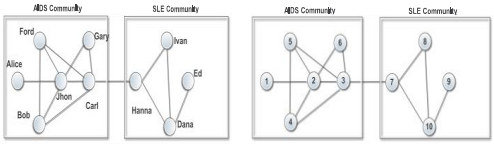

FIGURE. 6. Example of degree attack: original network; (b) naïve anonymized network.

3) DATAMINER

The data miner applies mining algorithms to the data provided by data collector, and he wishes to extract useful information from data in a Privacy-preserving manner. With the user role-based methodology proposed in this paper, we consider the data collector should take the major responsibility of protecting sensitive data, while data miner can focus on how to hide the sensitive mining results from untrusted parties.

A.CONCERNS OF DATAMINER

The privacy issues coming with the data mining operations are twofold. On one hand, if personal information can be directly observed in the data and data breach happens, privacy of the original data owner (i.e. the data provider) will be compromised. On the other hand, equipping with the many powerful data mining techniques, the data minerisable to find out various kinds of information underlying the data.

B.APPROACHES TO PRIVACY PROTECTION

Based on the distribution of data, PPDM approaches can be classified into two categories, namely approaches for centralized data mining and approaches for distributed data mining. Distributed data mining can be further categorized into data mining over horizontally partitioned data and data mining over vertically partitioned data. Based on the technique adopted for data modification, PPDM can be classified into perturbation-based, blocking-based, S wapping-based, etc.

C. PRIVACY-PRESERVING CLUSTERING

Cluster analysis is the process of grouping a set of data objects into multiple groups or clusters so that objects with in a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. Clustering methods can be categorized into partitioning methods, hierarchical methods, density-based methods, etc.

4) DECISION MAKER

As shown in Fig.2, a decision maker can get the data mining results directly from the data miner, or from some *Information Transmitter*. It is likely that the information transmitter changes the mining results intentionally or unintentionally, which may cause serious loss to the decision. In addition to investigate the privacy-protection approaches adopted by each user role, in this paper we emphasize a common type of approach, namely game theoretical approach, that can be applied to many problems involving privacy protection in data mining. By using methodologies from game theory, we can get useful implications on how each user role should behaviour in an attempt to solve his privacy problems.

A.CONCERNS OF DECISION MAKER

The data mining results provided by the data miner are of high importance to the decision maker. On the other hand, if the decision maker does not get the data mining results directly from the data miner, but from someone else which we called *information transmitter*, the decision maker should be sceptical about the credibility of the results, in case that the results have been distorted. Therefore, the privacy concerns of the decision maker are twofold: how to prevent unwanted disclosure of sensitive mining results, and how to evaluate the credibility of the received mining results.

B.APPROACHES TO PRIVACY PROTECTION

To deal with the first privacy issue proposed above, i.e. to prevent unwanted disclosure of sensitive mining results, usually the decision maker has to resort to legal measures. To handle the second issue, i.e. to determine whether the received information can be trusted, the decision maker can utilize methodologies from data provenance, credibility analysis of web information, or other related research fields. In the rest part of this section, we will first briefly review the studies on data provenance and web information credibility, and then present a preliminary discussion about how these studies can help to analyse the credibility of data mining results.

VI. GAME THEORY IN DATA PRIVACY

A. GAME THEORY PRELIMINARIES

When participating in a data mining activity, each user has his own consideration about the benefit they may obtain and the (privacy) cost he has to pay. Generally, the user would act in the way that can bring him more benefits, and one user's action may have effect on other users' interests. Therefore, it is natural to treat the data mining activity as a *game* played by multiple users, and apply game theoretical approaches to analyze the iterations among different users.

Game theory provides a formal approach to model situations where a group of agents have to choose optimum actions considering the mutual effects of other agents' decisions. The essential elements of a game are *players*, *actions*, *pay offs*, and *information*. Equilibrium is a strategy profile consisting of a best strategy for each of the players in the game. The most important equilibrium concept for the majority of games is *Nash equilibrium*.

B. PRIVATE DATA COLLECTION AND PUBLICATION

If a data collector wants to collect data from data providers who place high value on their private data, the collect or may need to negotiate with the providers about the ``price'' of the sensitive data and the level of privacy protection. 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.

C.PRIVACY PRESERVING DISTRIBUTED DATA MINING

1) SMC-BASED PRIVACY PRESERVING DISTRIBUTED DATA MINING

In a SMC scenario, a set of mutually distrustful parties, each with a private input, jointly compute a function over their inputs. However, during the execution of the protocol, a party may take one of the following actions in order to get more benefits:

_Semi-honest adversary: one follows the established protocol and correctly performs the computation but attempts to analyze others' private inputs.

_Malicious adversary: one arbitrarily deviates from the established protocol which leads to the failure of computation.

_Collusion: one colludes with several other parties to expose the private input of another party who doesn't participate in the collusion.

2) LINE ARREGRESSION AS A NON-COOPERATIVE GAME

In order to protect privacy, individuals add noise to their data, which affects the accuracy of the model. In the inter actions among individuals are modelled as an on-cooperative game, where each individual selects the variance level of the noise to minimize his cost. The cost relates to both the privacy loss incurred by the release of data and the accuracy of the estimated line regression model. It is shown that under appropriate assumption son privacy and estimation costs, there exists a unique pure Nash equilibrium at which each individual's cost is bounded.

A.DATA ANONYMIZATION

The proposed method models each tuple in the data table as a player, and computes the payoff to each player according to a concept hierarchy tree (CHT) of quasi-identifiers. The equivalent class in the anonymous table is formed by establishing a coalition among different tuples based on their payoffs. Given the afford able in formation loss, the proposed method can automatically find the most feasible value of k, while traditional methods need to fix up the value of k before the anonymization process.

B. ASSUMPTIONS OF THE GAME MODEL

We present the basic elements of some proposed game models. Most of the proposed approaches adopt the following research paradigm:

- _ define the elements of the game, namely the players, the actions and the payoffs;
- _ determine the type of the game: static or dynamic, complete information or incomplete information;
- _ solve the game to find equilibriums;
- _ analyze the equilibriums to obtain some implications for practice.

Unreasonable assumptions or too many assumptions will hurt the applicability of the game model.

C. MECHANISM DESIGN AND PRIVACY PROTECTION

1) MECHANISMS FOR TRUTHFUL DATA SHARING

A mechanism requires agents to report their preferences over the outcomes. Since the preferences are private information and agents are self-interested, it is likely that the agent would report false preferences. In many cases, the mechanism is expected to be *incentive compatible* that is, reporting one's true preferences should bring the agent larger utility than reporting false preferences. Such mechanism is also called *truthful mechanism*.

2) PRIVACY AUCTIONS

Aiming at providing support for some specific data mining task, the data collector may ask data providers to provide their sensitive data. The data provider will suffer a loss in privacy if he decides to hand over his sensitive data. In order to motivate data providers to participate in the task, the data collector needs to pay monetary incentives to data providers to compensate their privacy loss. Since different data providers assign different values to their privacy, it is natural for data collector to consider buying private data using an auction. In other words, the data provider can sell his privacy at auction.

FIGURE 7. Privacy auction. (a) Data provider makes a bid (privacy valuation v_i); (b) data collector makes a bid (price willing to pay for the data).

III. RELATED WORKS

Currently, the implementations of big data applications are circumventing data - centres, and networks act as the transmission pipes for data collection, aggregation, processing, sharing and delivery. With the explosively growing of the big data, the networks become a bottleneck for the explosive data sharing. Thus, it is necessary to redesign the network functions to build a network highway for efficient data sharing. Meanwhile, it has been observed that the network communication model has begun shifting from routing-centric to content/service/X-centric. ICN has been identified to have potential to compensate for network resources to make the communication model more scalable and efficient. As the preceding researches on ICN, TRAID, and ROFL explored methods for routing based on identifiers. Among these work, CCN mainly focuses on the opportunistic close information copy discovery and retrieval through content-name-based routing. CCN shows its promising features on the low delay and traffic with the expense of in-network cache memories. However, the flooding of interest and routing information is used, which limits the scalability of CCN. Even for the CCN with OSPF-like routing in local domains and the GP-like routing for inter-domain as in, the flooding still cannot be avoided for these routing algorithms. It also suffers from the on-path caching, which induces the inefficiency for memory utilization. In CCN, if there is no match for one Interest, this Interest will be discarded, which restricts the data reachability. CCN also lacks of data management.

III. REFERENCE ARCHITECTURE

NBRR is the core function of ICN to enable a data consumer to fast discover and retrieve the close copy of the desired data under the situation where there are several copies of the same data in the network. Besides the caching strategy to enable the efficient usage of caches in network, we identify the design requirements of NBRR as follows.

1. Any cast-capable routing: NBRR has capability to enable information consumer to is cover/retrieve the close copy of information by its name.

2. Energy Efficiency: Global flooding should be reduced in NBRR. Network deployment should be optimized on energy consumptions.

3. Aggregatability: Information is aggregately to reduce the discovery cost.

4. Management: NBRR should possibly provide the basic support for achieving efficient information management including remove, update, and synchronization.

5. Mobility support: Hosts efficiently retrieve the desired information during movement through NBRR. It is observed that the global flooding of prefix brings out much wasteful energy consumptions, because of the redundant traffic. Thus, we employ the basic thought of divide & conquer to restrict the flooding in the local area and design a reference architecture as Fig. 8.

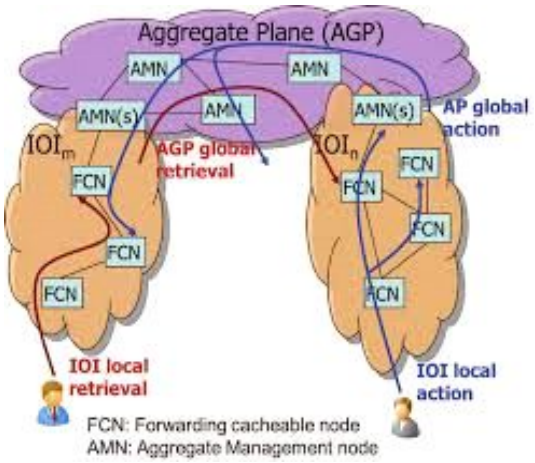

FIGURE 8. Reference architecture.

In the reference architecture, the network is divided into many suitable size IOIs for data caching and fast retrieval. An MP is utilized for global reachability and management. There are two kinds of entities, forwarding cacheable node (FCN) and aggregate management node (AMN).FCN has functions of forwarding and in-network caching, which could be routers, BSs, access points with storage. It is noticeable that FCNs hold memories for caching information in the network besides forwarding. AMN has functions of forwarding, and global discovery/management, which can be imaged as gateway, NAT, or network manager with storage.AMNs also hold memories, which are utilized to support management.

IV.CONCLUSION

Big data demands a network architecture that can achieve energy efficiency, availability, high-performance, and data-aware intelligence. It is observed that ICN approach can play a crucial role to provide the networking service in the era of big data. Therefore, we avoid the disadvantages and absorb the merits of the typical existing ICN architecture, CCN and NetInf, and design reference network architecture. We propose the naming method, packet format, and entity functions for this architecture. Then we provide the basic designs on data registration and data retrieval procedures. To enable the energy efficiency, we model the network and examine the impact on the energy consumptions for the proposed architecture from the key factor, IOI size. We observe that the energy consumption firstly decreases and then increases with IOI size increasing, and the optimized IOI size increases with average retrieval times increasing. In this paper we review the privacy issues related to data mining by using a user-role based methodology. Each user role has its own privacy concerns; hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others:

_ For data provider, his privacy-preserving objective is to effectively control the amount of sensitive data revealed to others. To achieve this goal, he can utilize security tools to limit other's access to his data, sell his data at auction to get enough compensation for privacy loss, or falsify his data to hide his true identity.

_ For data collector, his privacy-preserving objective is to release useful data to data miners without disclosing data providers' identities and sensitive information about them. To achieve this goal, he needs to develop proper privacy models to quantify the possible loss of privacy under different attacks, and apply anonymization techniques to the data.

_ For data miner, his privacy-preserving objective is to get correct data mining results while keep sensitive information undisclosed either in the process of data mining or in the mining results. To achieve this goal, he can choose a proper method to modify the data before certain mining algorithms are applied to, or utilize secure computation protocols to ensure the safety of private data and sensitive information contained in the learned model.

_ For decision maker, his privacy-preserving objective is to make a correct judgement about the credibility of the data mining results he's got. To achieve this goal, he can utilize provenance techniques to trace back the history of the received information, or build classifier to discriminate true information from false information.

To achieve the privacy-preserving goals of different users' roles, various methods from different research fields are required.

V REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[2] R. Gibbons, A Primer in Game Theory. Hertfordshire, U.K.: Harvester Wheat sheaf, 1992.

[3] D. C. Parkes, ``Iterative combinatorial auctions: Achieving economic and computational efficiency," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, PA, USA, 2001.

[4] S. Carter, ``Techniques to pollute electronic profiling," U.S. Patent 11/257614, Apr. 26, 2007. [Online]. Available: https://www.google.com/patents/US20070094738

[5] Verizon Communications Inc. (2013). 2013 Data Breach Investigations Report. [Online].

Available: http://www.verizonenterprise.com/ resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf

[6] A. Narayanan and V. Shmatikov, ``Robust de-anonymization of large sparse datasets," in Proc. IEEE Symp. Secur. Privacy (SP), May 2008, pp. 111_125.

[7] 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.

[8] R. C.-W. Wong and A. W.-C. Fu, ``Privacy-preserving data publishing: An overview," Synthesis Lectures Data Manage., vol. 2, no. 1, pp. 1_138, 2010.

[9] L. Sweeney, ``k-anonymity: A model for protecting privacy," Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557_570, 2002.

[10] R. J. Bayardo and R. Agrawal, ``Data privacy through optimal k-anonymization," in Proc. 21st Int. Conf. Data Eng. (ICDE), Apr. 2005, pp. 217_228.

ARUNUDAYA R received the BE degree in 2015 from the Anna University and pursuing ME in the department of Computer Science in Vel Tech Multitech Engineering College affiliated by Anna University, Chennai, India. Her main research interests include Big Data, Cloud Computing and Data Mining.

AFREEN BANU E is an Assistant Professor in the department of Computer Science in Vel Tech Multitech Engineering College affiliated by Anna University, Chennai, India. Her main research interests include Internet of Things, Cloud Computing and Big Data.

