Security Aspects of Bigdata

Shreelakshmi C.M, Asharani M, Parashivamurthy B M

Assistant Professor, Department of Computer Science and Engineering

GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, India

Abstract—The proliferation of web-based applications and information systems, and recent trends such as cloud computing and outsourced data management, have increased the exposure of data and made security more difficult. The era of "big data" has ushered in a wealth of opportunities to advance science, improve health care, promote economic growth, reform our educational system, and create new forms of social interaction and entertainment. Yet these opportunities bring with them increasing challenges related to data security and privacy. The challenges include: a lack of effective tools and approaches for securely managing large-scale data and distributed data sets; third party data sharing; vulnerabilities in ever-expanding public databases; and technological advancements that are outpacing policy as it relates to digital security and privacy. Another major challenge is intentional or malicious data leakage. With the fabulous development of information technology, big data application prompts the development of storage, network and computer field. It also brings new security problems. This security challenge caused by big data has attracted the attention of information security and industrial community domain. This paper summarizes the characteristics of big data information security, and focuses on conclusion of security problems under the big data field and the inspirations to the development of information security technology. Finally, this paper outlooks the future and trend of big data information security.

Keywords- Bigdata, Cloud computing, Data leakage, data security, privacy.

I. INTRODUCTION

Security and privacy concerns are growing as big data becomes more and more accessible. The collection and aggregation of massive quantities of heterogeneous data are now possible. Large-scale data sharing is becoming routine among scientists, clinicians, businesses, governmental agencies, and citizens. However, the tools and technologies that are being developed to manage these massive data sets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy. We also lack adequate policies to ensure compliance with current approaches to security and privacy. Furthermore, existing technological approaches to security and privacy are increasingly being breached, whether accidentally or intentionally, thus necessitating the continual reassessment and updating of current approaches to prevent data leakage. Issues around data confidentiality and privacy are under greater focus than ever before as ubiquitous internet access exposes critical corporate data and personal information to new security threats. On one hand, data sharing across different parties and for different purposes is crucial for many applications, including homeland security, medical research, and environmental protection. The availability of ''big data'' technologies makes it possible to quickly analyze huge data sets and is thus further pushing the massive collection of data. On the other hand, the combination of multiple datasets may allow parties holding these datasets to infer sensitive information. Pervasive data gathering from multiple data sources and devices, such as smart phones and smart power meters, further exacerbates this tension. Techniques for fine-grained and context-based access control are crucial for achieving data confidentiality and privacy[1][2]. Depending on the specific use of data, e.g. operational purposes or analytical purposes, data anonymization techniques may also be applied. An important challenge in this context is represented by the insider threat, that is, data misuses by individuals who have access to data for carrying on their organizational functions, and thus possess the necessary authorizations to access proprietary or sensitive data. Protection against insider requires not only fine-grained and context-based access control but also anomaly detection systems, able to detect unusual patterns of data access, and data user surveillance systems, able to monitor user actions and habits in cyber space – for example whether a data user is active on social networks. Notice that the adoption of anomaly detection and surveillance systems entails data user privacy issues and therefore a challenge is how to reconcile data protection with data user privacy. It is important to point out that when dealing with data privacy, one has to distinguish between data subjects, that is, the users to whom the data is related, and data users, that is, the users accessing the data[16].

II. THREATS OF BIG DATA SECURITY

Today, big data has penetrated into various industries, and has become a kind of production factor which plays an important role. In the future it would be the highest point of the competition. With the development of rapid processing and analysis technology, the potential information it contained can quickly capture the valuable information in order to provide reference for decision making. However, as big data setting off a wave of productivity and consumer surplus, the challenge of information security is coming either[9].

A. Data Acquisition

The source of big data is diversity. Therefore, the first step to process big data is to collect data from source and pre-process, in order to provide uniform high quality data set to the subsequent process. As a result, due to the inundation of data acquisition, large data become more likely to be "discovered" as a sensitive target, and be more and more attention. On one hand, big data not only means the huge amounts of data, but also means more complex and more sensitive data. These data would attract more potential attackers, and become a more attractive target. On the other hand, with data assembled, the hacker could get more data in one successful attack, and reduce hacker's attack costs[10].

The confidentiality of information refers that according to a specified requirements, information can not be disclosed to unauthorized individuals, entities or processes, or provided the characteristics of its use. A large amount of data collection includes a large number of enterprises operating data, customer information, personal privacy and all kinds of behaviour records. The centralized storage of these data increases the risk of data leakage, and not abused of these data also becomes a part of the personal safety[13]. There is no clear definition to the proprietorship and right to use of sensitive data. And many analysis based on large data did not consider the individual privacy issues involved either.

The integrity of information refers to all the resources which can only be modified by authorized people or with the form of authorization. The purpose is to prevent information from being modified with unauthorized users. Due to the openness of big data, in the process of network transmission, information would be damaged, such as hackers intercepted, interruption, tampering and forgery. Encryption technology has solved the data confidentiality requirements as well as protecting data integrity. But encryption cannot solve all of the safety problems[11].

B. Storage of Data

The formation of network society creates the platform and channel of resource sharing and data exchange for the big data in the field of various industries. Network society based on cloud computation provides an open environment for big data. Network access and data flow provides the basis of rapid elasticity push of the resources and the personalized service. In recent years, from the chain reaction of user account information being stolen on the Internet, it can be seen that big data is more likely to attract hackers, and once being attacked, the volume of stolen data is huge[12].

Before big data, data storage is divided into relational database and file server. And in current big data, diversity of data type makes us unprepared. For more than 80% of the unstructured data, NoSQL has the advantages of scalability and availability and provides a preliminary solution for big data storage. But NoSQL still exist the following problems: one is that relative to the strict access control and privacy management of SQL technology; Secondly, although NoSQL software gain experience from the traditional data storage, NoSQL still exist all kinds of leak.

C . Data Mining

With the development of computer network technology and artificial intelligence, network equipment and data mining application system is more and more widely used, to provide convenient for big data automatic efficient collecting and intelligent dynamic analysis. On the one hand, big data itself exits leak. Big data itself can be a carrier of sustainable attack. Viruses and malicious software code hidden in large data is hard to find. On the other hand, the technique of attack improves. At the same time of the big data technology such as data mining and data analysis gaining value information, the attacker using these big data technology either, just as the two following aspects[3][4]. A large number of facts show that failure to properly handle big data will cause great violations to users' privacy. According to the different contents need to be protected, privacy protection can be further divided into location privacy protection, anonymous identifier protection, anonymous connections and so on. The threat People faced with is not only personal privacy leakage, but also prediction and behaviour of the people based on big data. In fact, anonymous protection cannot protect privacy very well. Research on social network also shows that user attributes can be found from the group features[4].

Currently collection, storage, management and use of user data is short of specification, and regulation [5][6]. Users can't determine their privacy information usage. In commercial scenario, user should have the right to decide how their information be used, and realize users' controllable privacy protection.

A general view about big data Is: data itself can tell everything, the data itself is a fact[7]. In fact, if not carefully screened, the data can deceive people, just as people can sometimes be deceived by their eyes.

one of the threat of big data credibility is counterfeit or deliberately manufacturing data, and the wrong data often lead to wrong conclusions. If data application scenarios is clearly, someone could deliberately manufacturing data, and create a "false scent", to induced analysts come to the conclusion that was on their side. Because of false information often hidden in a lot of information, it make impossible to identify authenticity of information, so as to make wrong judgment. Due to the production and propagation of false information in network community is becoming more and more easy, its effects should not be

underestimated and simply using information security technology to identify the authenticity of all sources is impossible[10].

III. DATA SECURITY PROTECTION TECHNIQUE

Key technologies in Security protection fields are in great demands to face the security challenges. In this section, we introduce important relevant fields.

A. Individual User

As with individual users' information in big data environment, the core and basic techniques to provide privacy protection are still in developing period. Take typical Kanonymity scheme as an example, its early version[13] and optimized version divide quasi-identifiers into groups through tuple generalization[14] and restraining method. When an equivalence class has identical value on some sensitive attribute, attackers are able to confirm its value. In response to this issue, researchers proposed 1-diversity[15] anonymity.

Current edge anonymity schemes are mainly based on adding and deleting of the edges. Edge anonymity can be effectively achieved by adding, deleting and exchanging edges randomly[16]. There are problems in such methods that noises randomly added are exiguity, and protections to anonymous edges are insufficient. An important method is to perform division and aggregation operations to super nodes such as node aggregation based anonymous method, genetic arithmetic based method and simulated annealing method based method.

B. Internet Enterprise

Information security is critical important for Internet enterprises. System security adopts techniques such as redundancy, network separation, access control, authentication and encryption [18]. Security issues are caused by openness, boundless, freedom of the networks, the key to solve such issues are making network free from them and turning network into controllable, manageable inner system. As network system is the foundation of application system, network security becomes principal issue. Ways to solve network security issues are network redundancy, system separation and access control

C. Cloud Service Provider

CSPs provide following measures to prevent security issues in cloud environment. In order to prevent CSPs from peeping users' data and program, separating power and hierarchical management are needed to control access to data in cloud. Provide different authority in accessing data to service provider and enterprise to ensure data security. Enterprise should have total authority and limit authority to CSP.

In cloud computing environment data separation mechanism prevents illegal access to data, however, we should take care of data leakage from CSPs. Mature techniques as symmetrical encryption, public key encryption are available to encrypt data and then upload data to cloud environment. In cloud environment data division is often used with data encryption i.e. encrypted data are scattered in user end and spread in several different clouds. In the way, any CSP is not able to gain complete data.

IV. THE SECURITY CHALLENGES

The Challenges Security and privacy concerns are growing as big data becomes more and more accessible. The collection and aggregation of massive quantities of heterogeneous data are now possible. Large-scale data sharing is becoming routine among scientists, clinicians, businesses, governmental agencies, and citizens. However, the tools and technologies that are being developed to manage these massive data sets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy. We also lack adequate policies to ensure compliance with current approaches to security and privacy. Furthermore, existing technological approaches to security and privacy are increasingly being breached, whether accidentally or intentionally, thus necessitating the continual reassessment and updating of current approaches to prevent data leakage[11][12].

Private businesses, hospitals, and biomedical researchers are also making tremendous investments in the collection, storage, and analysis of large-scale data and private information. While the aggregation of such data presents a security concern in itself, another concern is that these rich databases are being shared with other entities, both private and public. Private business have already established partnerships with the government, as highlighted by the Prism Program, which provides the NSA with direct access to the databases of Microsoft, Yahoo, Google, Facebook, PalTalk, YouTube, Skype, AOL, and Apple (Greenwald and MacAskill, 2013a). Hospitals are increasingly adopting Electronic Medical Record (EMR) systems to enable the aggregation of patient data within a hospital and across a hospital system (Charles et al., 2013), and biomedical researchers are tapping into EMR data and new, nontraditional data sources such as social media, sensor-derived data (home, body, environment), and consumer purchasing and mobility patterns. In addition, the National Institutes of Health and other funding agencies are strongly encouraging biomedical researchers to share their research data.

Data hackers have become more damaging in the era of big data due to the availability of large volumes of publically

available data, the ability to store massive amounts of data on portable devices such as USB drives and laptops, and the accessibility of simple tools to acquire and integrate disparate data sources. According to the Open Security Foundation's DataLossDB project (http://datalossdb.org), hacking accounts for 28% of all data breach incidents, with theft accounting for an additional 24%, fraud accounting for 12%, and web-related loss accounting for 9% of all data loss incidents. Greater than half (57%) of all data loss incidents involve external parties, but 10% involve malicious actions on the part of internal parties, and an additional 20% involve accidental actions by internal parties.

V. SECURITY METHODS FOR BIG DATA

A. Type Based Keyword Search for Security Of Big Data.

1)Big data provide many business opportunities to the information technology industry. Large scale applications of sensor networks, electronic health record systems, emails as well as social networks generate massive data each day. The volume of information collected and stored has exploded. Cloud computing system process extraordinary storage capacity and computation power and is promising to handle the big data processing system with its features. However, since the cloud data service provider system are distributed to share and process sensitive information assigned to them, a malicious data stealer might probably bring about serious privacy problems .Data encryption technology is used for boost information privacy protection. However, traditional encryption primitives (such as symmetric key encryption and public key encryption) are not capable to ensure the usability and hinder even authorized users from searching several keywords of encrypted files , it is difficult for the users to retrieve desired information from encrypted big data. So It is necessary to explore new cryptographic primitives to provide data encryption and searchability for big data era. Searchable encryption technology could fulfill the requirements to realize operability and data confidentiality, simultaneously. In this method, we provide a novel keyword search method to enable customers easily searching keywords from encryption-protection data. Moreover, the encrypted big data could be managed by different type that was assigned by data owner. Moreover, the access right can be given to others according to the user‟s willingness Researchers also explore new search patterns for searchable encryption, such as fuzzy search, subset search, rank search. The public key encryption with keyword search (PEKS) scheme was proposed in order to offers the user to retrieve files through keyword searching. Consider an electronic health record system. A user sends an encrypted file m appended with some encrypted keywords w1, w2,.... wn that are extracted from the message to the data service provider.

The data are organized in the format

PKE(pk A,m) || PEKS (pk A,w1)||.... || PEKS (pk A,wn), in which pk A is the public key of user. The user could generate a trapdoor that contains certain keyword Wi. After receiving the trapdoor, data service provider search the encrypted files and returns all files that contain Wi. Other researchers also try to extend searchable encryption scheme to multiple users.

2)System model:

We will design a secure big data storage system that supports multiple users. In this system, authorized users are able to store encrypted data and carry out keyword queries on encrypted data without decrypting all the files.

Moreover, data owners could delegate certain type of files to other users.

Data Service Provider: Data service provider is responsible to generate global parameter for the whole system. Its main responsibility is to store user‟s encrypted data, respond to user‟s retrieve request and return corresponding files[13][14].

Moreover, a new functionality is provided: re-encrypt second level ciphertext to first level ciphertext on behalf of delegatee. We should point out that our scheme provides finegrained delegation authority management. In other reencryption based searchable encryption schemes, delegatee is capable to decrypt all files that belong to the data owner when delegation right is given. However, in this system, delegator could delegate a designated type of files to delegatee for decryption so that delegatee is only able to recover part of ciphertext of data owner. **Delegator:** Delegator is usually the data owner and can issue the keyword search query. Only data owner has the right to update the encrypted file and the encrypted keyword index. The data file could be images, documents, videos, programs, etc. In addition, delegator is responsible to generate re-encryption key for delegatee.

Delegatee: Delegatee is responsible to generate its own private key and fulfill the delegation responsibility, i.e., to decrypt first level ciphertext on behalf of its delegator.

C. SECURITY ANALYSIS:

In this subsection, we discuss our type based keyword search for encrypted data from the following security requirements: data confidentiality, query privacy and query unforgeability. We assume that users‟ private keys are kept secret. Data confidentiality: The meanings of information confidentiality in our scheme are three fold. Both the first level and second level cipher texts should be protected from both data service provider and malicious eavesdropper. Moreover, the curious data server and malicious adversary could not obtain any information

about keyword from the encrypted index of keywords. In our system, the second level ciphertext and index of keywords are enciphered before uploading to the data service provider through algorithm Encrypt(m, pkRi, t, w) [15][16]. Since data owner‟s private key is kept secret, the data server could not get any information about the plaintext through illegal decryption operation without private key. The element r ∑ Z*p is chosen randomly to resist replay attack. Query privacy: The meaning of query privacy here indicates that the protection of personal information of users and information which may be recovered by malicious party from the keyword retrieve phase. In the keyword retrieval process, the user firstly generates a trapdoor for the keyword and sends it to the data server. In the whole process, curious data server could not get any privacy information about keyword w .Query unforgeability: In this system, an individual private key is utilized to encrypt keywords by each user. Various keyword trapdoor queries , generated by different users‟ secret keys Risk are distinctive. In multi-user big data system, no user can create a spurious trapdoor query on behalf of another illegal user. Thus, the query unforgeability is offered in this system. In this paper, we construct a type based searchable encryption scheme to secure big data, which also allows reencryption function. The plaintexts are generated with respect to a certain type. The security analysis shows that our scheme could provide data confidentiality, query privacy as well as query unforgeability[19].

Achieving Big Data Privacy via Hybrid Cloud

With the rapid development of electronic and communication technology, the amount of data produced by medical systems, surveillance systems or social networks has been grown exponentially, which makes it hard for many organizations to cost-effectively store and manage these big data. Cloud computing, a new business model, is attractive, provides the advantage of reduced cost through sharing of computing and storage resources. However, concerns in term of the privacy of data stored in public cloud have delayed the adoption of cloud computing for big data . On one hand, a large amount of image, such as medical systems or social networks, may contain sensitive information. On the other hand, Cloud Service Providers (CSPs), who own the infrastructures on which clients‟ data are stored, have full control of the stored data. Therefore, the data stored in public cloud may be scanned by CSPs for advertisement or other purposes. Furthermore, attackers may be able to access data stored in cloud if there is not sufficient secure mechanism provided by CSPs. Most existing solutions employ traditional cryptographic algorithms, such as AES, to encrypt data and then store encrypted data in public cloud. However, for image data, which have much larger size than text data, heavy computation overhead will be introduced by this approach. Meanwhile, for the mobile devices, which have been widely used, much battery energy will be consumed, and it will increase delay because of the limited computation resources. Therefore, the traditional cryptographic approaches are not suitable for big data privacy. In recent years, various image encryption algorithms have been proposed to speed up the process, among which the chaos-based approach with a substitution-diffusion layout appears to be a promising direction. In the substitution stage, the positions of pixels of the image are shifted via some chaotic map, and then the pixel values of the shuffled image are changed by chaotic sequences in the diffusion stage. However, the chaos system itself causes large computation overhead. Another approach is to take advantage of hybrid cloud by separating sensitive data from non-sensitive data and storing them in trusted private cloud and un-trusted public cloud respectively. However, if we adopt this approach directly, all images containing sensitive data or the ones that would not like to be seen by others have to be stored in private cloud, which would require a lot of storage in private cloud. Most users want to minimize the storage and computation in private cloud, and let public cloud do most of the storage and computation. To address the above challenge, we need to answer an important problem: How to efficiently achieve big data privacy by using hybrid cloud? Compared to using public cloud only, using hybrid cloud would have communication overhead between private and public cloud. Besides achieving data privacy, we want to reduce storage and computation in private cloud, as well as communication overhead between private and public cloud[17][15]. In addition, the delay introduced by communications between private and public cloud should be small. In this paper, we present a scheme that can efficiently achieve image data privacy in hybrid cloud. A novel random one-to-one mapping function is proposed for image encryption, which makes the pair wise affinity among jigsaws unreliable and at the same time significantly speeds up the process of substitution and diffusion. Only the random parameters of the mapping function are stored in private cloud.

System and Threat Model:

The original data come from private cloud, and are processed on servers within private cloud. If there are no sensitive data, the original data may be sent to public cloud directly. Otherwise, the original data will be processed to make no sensitive data leaked out. After being processed, most data are sent to public cloud, and a small amount of sensitive data are kept in private cloud. When a user queries the data, both private cloud and public cloud will be contacted to provide the complete query result. We consider an un-trusted public cloud who are curious and may intend to browse users‟ data. The public cloud has full

control of its hardware, software, and network. 3)Design Goals We want to protect image data privacy stored in public cloud via hybrid cloud. Specifically, we want to remove sensitive data and store them in trusted private cloud, and store the processed data (without sensitive information) in un-trusted public cloud. It would require too much storage in private cloud if we simply store the entire image with sensitive information in private cloud. Therefore, our design goal is to achieve image data privacy via hybrid cloud and at the same time reduce the following overheads: (1) the amount of data stored in private cloud, (2) the communication overhead between private and public cloud, and (3) the delay introduced by communications between private and public cloud. To promote the cloud computing as a solution for big data, we proposed an efficient scheme to address the increasing concern of data privacy in cloud for image data. Our scheme divides an image into blocks and shuffles the blocks with random start position and random stride[16][17]. Our scheme operates at the block level instead of the pixel level, which greatly speeds up the computation.

VI. SECURING BIG DATA ENVIRONMENTS WITH VORMETRIC

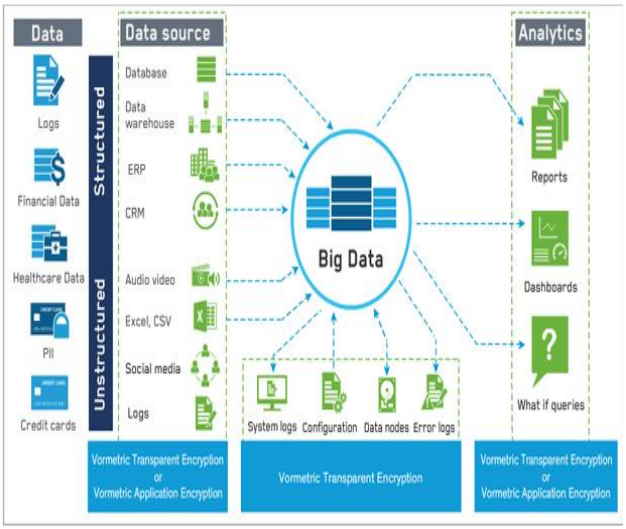

Figure: Securing Bigdata with Volumetric

Vormetric solutions for big data security enable organizations to maximize the benefits of big data analytics—while maximizing the security of their sensitive data and addressing the requirements of their compliance office. The Vormetric Data Security Platform offers the granular controls, robust encryption, and comprehensive coverage that organizations need to secure sensitive data across their big data environments—including big data sources, big data infrastructure, and big data analytic results. By delivering a single security solution that offers coverage of these areas, Vormetric enables security teams to leverage centralized controls that optimize efficiency and compliance adherence. The Vormetric Data Security Platform offers capabilities for big data encryption, key management, and access control—featuring several product offerings that share a common, extensible infrastructure. Further, the solution generates security intelligence on data access by users, processes, and applications.

VII. CONCLUSION

Big data have various challenges related to security like-computation in distributed programming, security of data storage and transaction log, input filtering from client, scalable data mining and analytics, access control and secure communication. For tackling with such security challenges we used different security methods like Type Based keyword search for security of big data, use of hybrid cloud to provide privacy in big data, securing bigdata environment with vormetric approaches.

REFERENCES

[1] Yang Yang,Xianghan Zheng"Type Based Keyword Search For Securing Big Data" in 2013 International Conference On Cloud Computing And Big Data.

[2] Xueli Huang and Xiaojiang Du"Achieving Big Data Privacy Via Hybrid Cloud" in 2014 IEEE INFOCOM workshops:2014 IEEE INFOCOM workshop on security and privacy in Big Data

[3] Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work and Think. Boston: Houghton Mifflin Harcourt , 2013 [4]Meng Xiao-Feng, Ci Xiang. Big Data Management: Concepts, Techniques and Challenges. Journal of Computer Research and Development, 2013, 50(1): 146-169 (in Chinese)

[5] Chen Mingqi, Jiang He. USA Information Network Security New Strategy Analysis in Big Data [J]. Information Network Security. 2012(8):32—35 [6]Narayanan A, Shmatikov V. How to break anonymity of the Netflix prize dataset. ArXiv Computer Science e-prints, 2006, arXiv:cs/0610105: 1-10

[7] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation.//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval(SIGIR'11), Beijing, China, 2011: 325-334

[8] Goel S., Hofman J.M., Lahaie S., Pennock D.M. and Watts D.J.. Predicting consumer behavior with Web search. National Academy of Sciences, 2010, 7 (41): 17486– 17490.

[9] http://www.wired.com/science/discoveries/magazine/16-07/pb_theory [8]Study Finds Web Sites Prying Less: Shift May Reflect Consumer Concerns[EB/OL]. http://www CNN.com, 2002-03-18

[10] A survey of data disclosing in 2010 by Verizon[EB/OL].[2012-05-10].

[11] Bessani A, Correia M, Quaresma B, et al. DEPSKY: Dependable and secure storage in a cloud-of clouds [C] //proc of the 6thConf on Computer System. New York: ACM, 2011:31-46

[12] Sweeney L..k-anonymity: a model for protecting privacy. InternationalJournal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10 (5): 557-570

[13] Sweeney L..k-Anonymity: Achieving k-Anonymity Privacy Protection using Generalization and Suppression.

[14] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1):1-52.

[15] Ying X. and Wu X.. Randomizing social networks: a spectrum preserving approach. //Proceedings of the SIAM International Conference on Data Mining (SDM'08), Georgia, USA, 2008: 739-750

[16] Lei Zou, Lei Chen and M. Tamer zsu. k-automorphism: a general framework for privacy preserving network publication. // Proceedings of the 35th International Conference on Very Large Data Bases (VLDB'2009), Lyon, France, 2009: 946-957.