International Journal of Engineering, Science and Mathematics
Vol. 6 Issue 8, December 2017 (Special Issue)
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Survey: Privacy Preserving Data Publication in the age of Big Data in IoT Era

Pavan Kumar Vadrevu
Sri Krishna Adusumalli
Vamsi Krishna Mangalampalli

Keywords:

Privacy
Safeguard
Anonymity
Attacks
Risk.

Abstract

IoT and social networks collecting large amount of data from the people in different ways. Now it is very important to provide privacy to the generated data before available to the public for survey or research purpose. Data Privacy related research is increasing very hastily during the past 10 years. Even it has many techniques to publish the data without violating the privacy in different areas like social networks, trajectory data etc. However it is a very challenging task to protect privacy in the age of Big Data in IoT Era, because the existing techniques may not be suitable for providing the privacy to the data to be published. This paper provides a survey on several challenges and solutions from the past few years in data privacy. This paper also discusses the research directions in the age of Big Data privacy in IoT Era.

Author correspondence:

Pavan Kumar Vadrevu
Research Scholar,
Department of Computer Science and Engineering,
Centurion University of Technology and Management, Paralakhemundi, Orissa

Sri Krishna Adusumalli
Associate Professor,
Department of Information Technology,
Shri Vishnu Engineering College for Women, Bhimavaram, Andhra Pradesh, India

Vamsi Krishna Mangalampalli
Professor,
Department of Computer Science and Engineering,
Centurion University Technology and Management, Paralakhemundi, Orissa

1. Introduction

Organizations like Government, Medical and Health Care, Social Networks and e-commerce web sites provides data to the researchers or third parties to gain knowledge for their own benefit or survey purpose. Several surveys showing that collecting data increased the sense of privacy violation. Privacy is subjective and it is difficult to define, because each person understands the concept of privacy in a different way. "Privacy is the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [Alan Westin, Columbia University, 1967]. The data given by the organizations like academic, government, healthcare, and other private sectors must manage, analyze and extract the data appropriately before available to the people. The data can be in many forms such as Relational Data, Social Network Data, Transactional Data and Trajectory Data viewed as a collection of records with one or more rows and columns such as Aggregate count data or contingency table contains data based on frequency (e.g. people who smoke in a given range of Zip code) and Non aggregate data or Micro data with noise added in cells. Each entry contains unique set of attributes or identifiers such as unique identifiers (e.g. Aadhar Card Number) also known as explicit identifiers, quasi identifiers (e.g. DOB, Zip Code), sensitive identifiers (e.g. Salary, Disease). Whenever that data to be published by the publisher (organization), the data should not contain unique identifiers, because with these identifiers the privacy of an individual is revealed. So the publisher publishes the data by removing the direct identifiers from the dataset. Even though the data set may not contain direct identifiers the privacy of an individual is violated by linking quasi identifiers of the published data set with other publicly available data set (Voter Registration List).

For example the Group Insurance Commission (GIC) collected the medical records of Massachusetts state employees. GIC published the data by removing direct identifiers such as *Name*, *SSN*, *Address* or *Phone number.* The data did hold demographic information such as Gender, Date of Birth and Zip code. According to the Massachusetts voter registration list, no one else had the same combination of Date of Birth, Gender and Zip code as William Weld, who was then the governor. Therefore, William Weld medical records were simple to identify in the data provided by GIC, from this it is very clear that the organizations responsible for safeguarding the data to protect privacy of individuals. Sweeney [1] showed that 87% of the population in US can be identified uniquely with DOB, Gender and Zip Code.

Another example, The America Online (AOL) search log published anonymous logs of 21 million web search queries posed by more than 500 thousand AOL users over a period of three months. In order to protect user privacy while publishing user data, AOL had changed the raw search log data prior to its release. They excluded IP addresses, browser and other user information and published only their identifier, query, query time, the rank of the document clicked and domain of the destination URL. In addition to this AOL removes the usernames in their search logs and user identifiers were replaced with random number prior to publication. Few days later, the identity of user #4417749 had been identified by New York Times [BZ06], and drawn to Thelma Arnold, a 62-year old widow from Lilburn, GA, enlightening her entire search history and sketch of her most private interests such as landscapers in her town to dating, her dog's habits and diseases of her friends.

Sweeney proposed a k-anonymity model to address the above re-identification disclosure for better protection in data publication [1]. The basic idea of k-anonymity is that each record is indistinguishable with at least k-1 other records in the data with respect to quasi-identifier. Some others models are proposed such as *l*- diversity, *t*-Closeness, (α, k) Anonymity, Differential Privacy. All these models use any of these annonymization mechanisms to anonymize the data such as Generalization, Suppression, Swapping, Bucketization and Randomization.

In Generalization the released data can be generalized by the attribute values from the given table. Suppression produces the release data by replacing some attribute values with special symbols. Swapping produces released data attributes that can be swapped with some other values. In Bucketization the released data can be partitioned from the original data table into non overlapping groups called as buckets. Randomization is adding some noise to the original data attributes or the sanitized data could be sampled from probability distribution [11].

The aforementioned models safe guard the data privacy for Relational Data, Social Network Data, Transactional Data and Trajectory Data. These methods are not applicable directly to the data generated by IoT devices or social network, of huge data that may generate continuously. In practice, anonymizing the Big Data is much more challenging than relation data.

Today the users will accept IoT deployment only if they are happy with the infrastructure is safe and secure and mostly privacy preserving, because IoT combines several network technologies and devices like RFID tags (Radio frequency identifiers), Smart phones and sensors. Cisco estimated that by 2020 there will be more than 50 billion Internet based devices including televisions and refrigerators. Consumers are facing global privacy problems due to the usage of some devices. Recently some reports given that several privacy violations in IoT applications.

Example 1: In 2013 press released a privacy risk related to Planning Tool for Resource Integration and Synchronization Management program(PRISM), with the US NSA(United States National Security Agency) used to collect sensitive data through electronic devices from users of major services like Microsoft Outlook, Google, Facebook etc. further an Internet security reported a risk that malware attacks increased up to 58 to 60 percent from 2011 to 2012 out of this 32 percent risks from stealing the information [12]. According to FTC (US Federal Trade Commission) reports on customer privacy, the most important and top loom is Privacy by Design (PbD) to defeat privacy issues in IoT.

Example 2: Another big privacy violation risk occurred in 2015 is malware compromised blood gas analyzers to gain access to hospital network and steal confidential information from an IoT eHealth Application [13], the framework is expected to be open to patients for necessary health information and gives updates ensures protection of patient records.

The above examples motivated to work on privacy related issues in Big Data in IoT Era. This survey provides an insight to privacy related challenges and policies to be implemented for protecting data privacy of individuals in the Big Data of IoT era.

Next section provides an over view of traditional privacy preserving models, followed by the basic challenges and motivation examples in the age of Big Data in IoT Era and conclusion.

2 Privacy Preserving Data Publication Models

Several models proposed from past two decades to safeguard the privacy of data generated from different sources they are mainly k-Anonymity, *l*-Diversity, *t*-Closeness, Differential Privacy and (α, k) Anonymity models. These models address several privacy challenges along with the solutions for Relational Data, Social Network Data, Transactional Data and Trajectory Data Sources.

2.1 k-Anonymity

A data table satisfy the k-anonymity property if every distinctly occurring sequence of quasi identifiers has at least (k) occurrences in the table [2][3]. This means each record in the table is identical from at least (k-1) other records with respect to the quasi identifiers.

Consider the following tables of health records:

Age	Zip	Diagnosis
29	90146	Cancer
23	90143	Flu
24	92235	Flu
52	92257	Cancer
50	92121	Obesity
49	92208	Obesity

Table: 1

Age	Zip	Diagnosis
[21–29]	9****	Cancer
[21–28]	9****	Flu
[21–28]	9****	Flu
[48–55]	92***	Cancer
[48–55]	92***	Obesity
[48–55]	92***	Obesity

Table: 2

The table 2 is an example of a 3-anonymization of the first (quasi identifier, Age and Zip). The quasi identifier tuple (Age, Zip) uniquely identify records in the first table. The modification to the quasi identifier field in the second table ensures that all distinctive instances of the quasi identifier tuple have at least three corresponding records. The modifications include syntactic actions like simplification (mapping the specific age "52" to the more general age range "48–53" and suppression (suppressing parts of the zip field). The table is now a collection of (k-sized) equivalence classes with respect to the tuple. It ensures preventing observers from resolving past a group of (k) records using the quasi identifier tuple as a key. Result efficient and useful k-anonymizations is a computationally challenging task for k>2 [4]. The Incognito algorithm was developed for partitioning tables to just about satisfy k-anonymity [5]. K-Anonymous tables avoid identity disclosures but they do not prevent observers from learning attributes about individuals. For example, in this table, an observer can infer more precise information about participant relative risks for flu or cancer based on just background age data (a background information attack). Some k-anonymizations may result in equivalence classes with uniform distributions on the sensitive attribute. This leads to sensitive attribute disclosure for all records in those classes (a homogeneity attack).k- anonymity cannot oppose homogeneity attack and background knowledge attacks these can be addressed by *l*-diversity.

2.2 *l* -Diversity

The data in the relational table is k-anonymized it may not safeguard privacy if it lacks of diversity in sensitive attributes that is even if the group of individuals is not distinguished in published data and these groups of individuals share the sensitive value, by this sensitive information is easily identified irrespective of the individual because all the individuals share the same kind of data [15]. To solve these kind of risks in relational data sets *l*-diversity model is proposed, it requires every quasi-identifier group with at least *l* "well represented sensitive values". The *l*-diversity concept is an attempt to stop homogeneity attack. An equivalence class in a table satisfies the *l*-diversity property if the sensitive attribute has at least l well represented values for the sensitive attributes in the record class [6][3]. A table is *l*-diverse if every equivalence class is *l*-diverse. The concept of '*l* well represented' sensitive values can have different meanings. For example, it could mean that there are *l* distinct values of the sensitive attribute (distinct *l*-diversity), or that the entropy of the sensitive attribute in each class is at least *l* bits (entropy *l*-diversity). The k-anonymous table presented above satisfies 2-diversity (in the distinct *l*-diversity sense) on the sensitive attribute, “Diagnosis.” Each k-sized equivalence class has at least 2 values for its sensitive diagnosis field (measles, flu) or (cancer, obesity). Distinct *l*-diversity is identical to another k-anonymity variant, p-sensitive k-anonymity when *l* =p [7]. Finding *l*-diverse anonymizations is computationally difficult and in practice harder than finding k-anonymization for the same table [8]. The Mondrian algorithm exists for partitioning tables to approximately satisfy *l*-diversity [5]. The amount, *l* is a measure of representativeness of the distribution of the sensitive attribute in the classes. It cannot always be a raw count of the number of distinct values in use. Attribute entropy is one more measure. The aim is to prevent leaking too much information about the relative frequencies on the sensitive attribute (Table 2). But the group distributions of the sensitive property are often skewed enough compared to the overall table distribution. So observers are still able to make limited inferences about relative sensitive attribute propensities. This is a skewness attack, a generalization of the k-anonymity homogeneity attack.

2.3 *t* - Closeness

The differential privacy motivation comes from *t*-closeness to reduce information increase related to the table, identified as attribute distribution. Differential privacy minimizes increase in information from the entire table modified by a single entity or that can be deleted. This approach protects the disclosure of risks in distinction to k- anonymity and *l*- diversity. They cannot take background knowledge about the sensitive attributes. Recent work shows that *t*- closeness can be equivalent to differential privacy [9] in some data publication situation. Information increase by comparing a *t*-closeness release with re identified table by removing all quasi identifiers from the table. A *t*- closeness table splits the entire table into equivalence class table by distributing sensitive attributes for the entire table.

Age	Zip	Diagnosis
29	90146	Cancer
23	90143	Flu
24	92235	Flu
52	92257	Cancer
50	92121	Obesity
49	92208	Obesity

Table: 3

Age	Zip	Diagnosis
< 40	90***	Cancer
< 40	90***	Flu
< 40	92***	Flu
< 50	92***	Cancer
< 50	92***	Obesity
< 50	92***	Obesity

Table: 4

The distribution can be taken very carefully. A metric that can satisfy the constraints is Earth Movers Distance metric (EMD) identified by Li et al. [3] it measures how much effort it takes to optimally transform the probability distribution to the next. *t*-closeness attempts to make sub groups of indistinguishable tables from the original table. From the above tables 3 and 4 it is clear that the table 4 derived from the original table is less context dependent and more privacy preservative but implementation of t- closeness is relatively difficult for large number of data sets. From the SABRE [10] “Sensitive Attribute Bucketization and Redistribution framework for *t*-closeness” algorithm implementation of tables with approximate t-closeness is possible.

2.4 (α, k) Anonymity

To oppose inference attack by controlling frequencies of sensitive values in each equivalence class. But the general (α, k) Anonymity model does not consider the specific requirement of each sensitive value. When the frequency of sensitive values in the whole dataset is not balanced, these models will have some limitations, Wong [16] projected simple (α, k) Anonymity model. In general (α, k) Anonymity model the parameter α cannot be less than the maximum frequency of the sensitive values, which will make some high sensitive values with lower frequency cannot be effectively protected. To satisfy each tuple specific requirement, Li Zude proposed (k, l) anonymity model [17], where k indicates the anonymization level of an identifying attribute cluster and l refers to the diversity level of a sensitive attribute cluster. In the model, k and l are designed on each record and they can be defined subjectively for the corresponding individual. But its individuation constraint is oriented to tuples, not oriented to sensitive values. So it is a heavy load to define k and l on each record subjectively if the dataset are huge. Given an anonymity table T, a quasi-identifier attributes set Q and a sensitive attribute domain S. For each sensitive value s (s ∈ S), let α s be a user-specified threshold of s. T is said to be a complete (α,k) anonymization if T satisfies k-anonymity and also satisfies simple α s dissociation property for each s with respect to Q and S. Complete (α,k)-anonymity model, which requires that each sensitive value s (s ∈ S) satisfies corresponding simple (α s , k) anonymity model, is more flexible compared with general (α,k) anonymity model and simple (α,k) Anonymity model.

Date of Birth	Zip	Diagnosis
1985.*.*	901**	Flu
1989.*.*	901**	Flu
1972.*.*	922**	Cancer
1992.*.*	922**	Flu
1990.*.*	921**	Fever
1978.*.*	922**	Cancer

Table: 7

The above table satisfies complete (α, k) Anonymity, when we let both α for “Cancer” and α for “Flu” and let both α for “Cancer” and α for “fever”. We can consider complete (α, k) Anonymity model as an extension of general (α, k)-anonymity model or simple (α, k)-anonymity model. When we only set an α-threshold for one sensitive value, i.e. let α s= α (0 < α < 1) and ∀ s (s ∈{S − s}), α s=1, it becomes a simple (α, k) Anonymity model. When we only set an α-threshold for all sensitive values, i.e. let ∀ s (s ∈ S),α s = α (0 < α < 1) it becomes a general (α , k) Anonymity model. When α s =1 for every s in S, it becomes a k-anonymity model.

2.5 Differential Privacy

Information collected from the individuals can be published by the organizations in the form of relational tables. The differential privacy criterion given by Dwork[14] to safeguard the privacy of individual information generated by including indistinguishable data items other words individual privacy is safeguarded if given access to the sanitized data set and information about all but one individual say *x,* in the data set. For example consider a data set provides the average income of a person in a particular state if the person want to move to other state then querying the data set before and after the person move can enable to find that persons income. Differential privacy gives to prevent this detection. It enables a form of reasonable deniability of that person; no one can prove that data part in data set related to that person. Consider two data sets d1 and d2 they are similar apart from d2 have one row represent the person *x,* data. Every one think that d1 data set represent all entities of the data set prior to the addition of *x,* d1 and d2 represent all entries of the data set after adding data, So the data sets d1 and d2 differ by only one entry called as adjacent data set.

d1	
Sl.no	Income in Rupees
1	50,000
2	60,000
3	75,000
4	85,000
5	90,000

Table: 5

d2	
Sl.no	Income in Rupees
1	50,000
2	60,000
3	75,000
4	85,000
5	90,000
6	1,25,000

Table: 6

These data set to be differentially private, a random function can be selected to make the data set private for random output O. since d1 and d2 are adjacent, the probability M(d1)=O.

P (M (d1) =O)/P (M (d2) =O) < e^{e}

The noise that M adds given by the sensitivity of the query function, that will be applied to the data. Sensitivity can be gives as △f= max[f(d1)-f(d2)] for all adjacent data sets. If a query count is 1, since adjacent data sets can differ at most 1. This is also called as symmetric exponential distribution given by Dwork [14]. The value of e should be selected by the differential privacy designers.

The above mentioned definitions and mechanisms for privacy preserving data publishing have many more extensions and relaxations for Relational Data, Social Network Data, Transactional Data and Trajectory Data. They are broadly classified into extensions of k-anonymity, extensions of *l*-diversity and relaxations of differential privacy. All these are having different intuitions and operate on different adversarial assumptions [11].

2.6 Personalized Privacy

The balance between privacy and utility is very important. The data to be utilized by someone following the privacy policies or violating them is a key issue whenever the data is published. Linking the data with different data sets is a common attack to provide data to public. A k-anonymous table allows an adversary to gain the sensitive information of a person. A k-anonymous table misplaces substantial information from the micro data [18]. K-anonymity does not take into account personal anonymity needs. The solution is personalized anonymity means a person can state the level of privacy guard for sensitive identifiers. A personal preference can be easily solicited from an individual when supplying the data. This is called as personalized privacy.

3. Challenges of Big Data in IoT Era

The challenges of big data in IoT era is to design the protocols related to privacy and security of the data generated from either static or mobile based IoT devices [19]. New generations are coming very rapidly in the mobile and cloud technologies privacy related issues are not addressed from many years. One issue in privacy is identification of personal data during communication if the user having any IoT device like a mobile phone or a RFID tag, the data can be transmitted from the device of the user to the cloud, at that time if the cloud service providers not making policies for preserving privacy it leads to violation of user data [20]. Another challenge if the person using any smart phone connected with the Internet may disclose the geographic location information and compromises privacy. IoT user may found risks related to privacy in terms of profiling, tracking, control access, reliability, confidentiality and privacy detection.

4. Conclusion

The utility of data increased day by day. Personal and sensitive data of an individual must be privacy preserving otherwise it leads to threats and attacks. In this paper a set of challenges are identified for extending the research in Privacy Preserving Data Publishing with respect to Big Data in IoT Era. This gives a basic step to start and test some of the privacy related aspects and provide some assumptions to protect individual privacy by making new policies for organizations to satisfy the privacy needs. Based on this in future personal privacy related aspects and methodologies can be discussed along with the challenges and solutions to extend the work.

References

[1] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000. Forthcoming book titled, *The Identifiability of Data*.

[2] Sweeney, Latanya. "*k*-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): pp. 557–570.

[3] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "*t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity." *Data Engineering,2007. ICDE 2007. IEEE 23rd International Conference on. IEEE*, 2007.

[4] Bonizzoni, Paola, Gianluca Della Vedova,and Riccardo Dondi. "The *k*-anonymity problem is hard." *Fundamentals of Computation Theory*.Springer Berlin Heidelberg, 2009.

[5] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Incognito: Efficient fulldomain *k*-anonymity." *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005.

[6] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M., 2007. *l*-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), p. 3.

[7]Truta, T. M., Campan, A. and Meyer, P., 2007. *Generating microdata with* p-*sensitive* k-*anonymity property* (pp. 124–141). Springer Berlin Heidelberg.

[8] Dondi, Riccardo, Giancarlo Mauri, and Italo Zoppis. "The *l*-diversity problem: Tractability and approximability." *Theoretical Computer Science* 511 (2013): 159–171.

[9] Soria-Comas, Jordi, and Josep Domingo- Ferrer. "Differential privacy via *t*-closeness in data publishing." *Privacy, Security and Trust (PST), 2013Eleventh Annual International Conference on*. IEEE,2013.

[10] Cao, Jianneng, et al. "SABRE: a Sensitive Attribute Bucketization and Redistribution framework for *t*-closeness." *The VLDB Journal* 20.1 (2011): 59–81.

[11] Privacy- Preserving Data Publishing By Bee- Chung Chen, Daniel Kifer, Kristen LeFevre and Ashwin Machanavajjhala Vol.2, Nos 1-2(2009) 1-67 DOI: 10.1561/1900000008.

[12] InternetSecurity Threat Report, Synantec Corporation Annual Report, 2013, www4.symantec.com/mktginfo/whitepaper/ISTR/21347932_GA-internet-security-threat-report-volume-20-2015-social_v2.pdf.

[13] D. Storm. “MEDJACK: Hackers Hijacking Medical Devices to Create Backdoors in Hospital Networks.” Computerworld,8 june 2015.www.computerworld.com/article/2932371/cybercrime-hacking/medjack-hackers-hijacking-medical-devices-to-create-backdoors-in-hospital-networks.html.

[14] C.Dwork,”Differential privacy,”in ICALP,2006.

[15] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M (2006) L-diversity: Privacy beyond k-anonymity. In: Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE’06),IEEE Computer Society, Washington,Dc,USA.

[16]R.C.W. Wong, J. Li, A.W.C. Fu, and K. Wang, “(α,k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing”, Proceeding of the 12th ACM SIGKDD Conference on KDD, PA: ACM Press,Philadelphia, Aug. 2006, pp. 754-759.

[17]Zude Li, Guoqiang Zhan, Xiaojun Ye, “Towards an Anti-inference (K, l)-anonymity Model with Value Association Rules”, DEXA, Springer-Verlag Berlin Heidelberg, Krakow, Sep. 2006, pp. 883-893.

[18] Personalized privacy preservation Xiaokui Xiao ,Yufei Tao Proceedings of the 2006 ACM SIGMOD international conference on Management of data Pages 229-240

[19] Security and Privacy for Cloud-Based IoT: Challenges, Countermeasures, and Future Directions, Jun Zhou, Zhenfu Cao, Xiaolei Dong, and Athanasios V.Vasilakos IEEE Communication Magazine,January 2017.

[20] The Quest for Privacy in the Internet of Things, Pawani Porambage and Mika Ylianttila, Corinna Schmitt, Pardeep Kumar, Andrei Gurtov, Athanasios V. Vasilakos, IEEE CLOUD COMPUTING PUBLISHED BY THE IEEE COMPUTERSOCIETY 2016.