Location Privacy Improvements in Telecommunication Data Management Systems

Milan Simakovic, Zoran Cica, and Dejan Drajic, *Senior Member, IEEE*

***Abstract*—In the era of digital transformation, data are among the most valuable resources. With the development of big data technologies, it is possible to store and process huge amounts of data. Data are possible to collect on every step with high granulation. Such a trend may seriously harm peoples' privacy. Corresponding laws and regulations are declared to protect data privacy. However, even when all the regulations are obeyed, privacy leakage may still happen if the implementation has some flaws. In this paper, we focus on telecommunication data sets and show how user's location information leakage may happen in already privacy-protected data. Moreover, we give a proposition on how this leakage can be prevented while preserving the same data entropy.**

***Index Terms*— data privacy, location tracking, big data.**

I. Introduction

According to [1], more than 66% of the world's population uses the internet at the end of 2021. Dynamic environment, rapid growth, and high competition on the market push companies to further enhance their products and reduce costs. Generated data may contain a huge volume of useful information that will drive such an initiative. And just like that, data becomes one of the main fuels in the industry. This phenomenon is further enhanced with the development of big data technologies [2].

To gather as much information as possible from the data, different ways of data processing are invented. Data correlated from different sources can give new insights that do not exist in separate data sets. Such great potential raises the issue of user privacy. To protect the users' privacy and limit the usage of data, GDPR (General Data Protection Regulation) [3] is defined in the European Union, CCPA (California Consumer Privacy Act) [4] in the USA (United States of America), and PIPL (Personal Information Protection Law) [5] in China. There are also laws that specify data privacy in a particular field, like HIPAA (Health Insurance Portability and Accountability Act) for healthcare in the USA [6]. These laws define data processing by protecting users' privacy and giving individuals the right to control the data collected from them. Although companies comply with all regulations, there are situations that can indirectly harm peoples' privacy. To emphasize this phenomenon, one example of such a scenario is presented in this paper. Namely, we show how the privacy of individuals may be harmed on the already privacy-protected telecommunication data set. Also, a recommendation on how this data set can be further masked to protect users from privacy breach while keeping the same information entropy is presented in the paper.

The remaining of the paper is organized as follows. Related work is discussed in section II. A brief overview of data privacy together with appropriate regulation laws is presented in section III. Section IV presents the main contribution of the paper. It discusses telecommunication data sets, shows the vulnerability of privacy-protected data sets on a sample, and proposes privacy improvement while keeping the same information entropy. Section V concludes the paper.

Milan Simakovic, Zoran Cica and Dejan Drajic are with the School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mails: milanrus@hotmail.com, zoran.cica@etf.bg.ac.rs, ddrajic@etf.bg.ac.rs).

Dejan Drajic is with the Innovation Centre of School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia.

II. Related work

Privacy in data technology, especially in big data is a hot topic over the last few years. This section gives a general overview of data privacy challenges, proposed solutions, and frameworks in the literature.

An overview of the privacy-preserving problems in big data stream mining is presented in [7]. Location privacy challenges for mobile applications are discussed in [8]. Location concerns are raised not only to the application provider but to the third parties that are able indirectly to calculate person location from gathered data. In addition, there is a raised concern for services that sell the location data to other parties. Following the people's position and their trajectory is recognized as a serious data privacy concern. To protect trajectory privacy from the data that contain GPS (Global Positioning System) location, a method for data masking that adds noise to the original data based on irregular polygons is proposed in [9]. Due to the rapid development of technology, IoT (Internet of Things) networks are becoming increasingly popular and, thus, becoming significant data generators. Such data may contain privacy-sensitive data. A privacy protection mechanism in industrial IoT based on information tree model is proposed in [10]. Location data is recognized as a huge potential for marketing and advertisement. Considering the number of users, generated amount of data is huge. A method based on big data technologies for location data mining while keeping the data privacy is proposed in [11]. This method uses a clustering algorithm and location entropy to emphasize the most active places.

Companies recognized data privacy as a serious problem and use different methods to solve these challenges. Data

 ISBN 978-86-7466-930-3

privacy models are introduced both for regular [12] and big data systems [13]. Big data models include all the data layers, i.e. collection, storage, and consumption. Data models state for schemas optimized for privacy issues is presented in [13]. In dynamic environments, data is shared both among other teams inside the company and externally. In such a huge data fluctuation, there is a high risk for data privacy issues. Considering that data privacy is protected by law, companies often decide to refrain from using the data which can significantly affect their business efficiency. Model for keeping the big data with a possibility to freely share and explore, and at the same time preserving the data privacy is presented in [14]. Data privacy mechanisms from the k-anonymity, l-diversity, and t-closeness perspectives are discussed in [15]. The advantages and challenges of these mechanisms are analyzed, and a new mechanism based on a combination of these three is proposed. Big data multilayer architecture and utilization of the "differential privacy" approach for sustainable data privacy are discussed in [16].

III. DATA PRIVACY

Data privacy, or information privacy, stands for the ability of the user to control what type of data is collected from him and how these collected data are used. Personal information, depending on the type of application, can be email, location, online search history, preferences, etc. Considering the complexity of modern applications and systems, it is necessary to gather some personal information (e.g. location) to provide the best possible experience to a client. However, applications can often gather more information than they really need and this might bring harm to people's privacy. Collected users' personal data can be used either inside the company to improve internal processes and services or to sell to other companies as a data set. Due to the lack of data security inside the company, data breaches may happen, and personal data can be stolen and used against the clients [17]. Data privacy is often mixed up with data security. While data security is protection from the 3rd party persons to access the data, data privacy is related to data collection methods and regulations that ensure that the user information is not exposed.

To bring closer control of data to the end-user, and restrict the companies in terms of data collection, usage, and surveillance capabilities, many governments around the world have created laws that regulate how data can be used, stored, and protected. Some of the most important regulations are GDPR [3], CCPA [4] and PIPL [5]. GDPR is the data privacy regulation law in European Union. GDPR gives clear instructions on how the data should be collected, transferred, stored, and protected. In addition, it gives users the right to control their personal data, i.e., the "right to be forgotten". This law regulative forces all the companies that collect users' data to have a mechanism to easily wipe all the data for clients without undue delay [3]. CCPA is the data privacy regulation law in the USA. This law stipulates that the user should be aware of what data are collected from him as well as give the company the right to sell his personal data. In addition, CCPA gives a guide to the company on how the law can be implemented [4]. PIPL is the data privacy law in PRC (People's Republic of China). Next to the regulations that prescript GDPR and CCPA, PIPL gives attention to data localization, i.e., that certain categories of personal data must be stored in PRC [5]. Next to the mentioned regulations, there are also many others that are implemented either in other countries or the ones that are industry-focused. For example, HIPAA is the regulation in the USA that governs how personal healthcare data should be handled [6].

To provide data privacy, companies use different techniques for data transformation. Among the most popular techniques are data anonymization and pseudonymization. Anonymization and pseudonymization present the process of transformation of UID (User Identifier) into data sets. Anonymized data stands for one-way encryption meaning that, once the UID is encrypted, there is no theoretical way to re-identify the user from it, neither directly nor indirectly. On the other side, pseudonymization stands for data masking techniques that can be reversed. It means that there is a way to decrypt the data. Pseudonymization is weaker in terms of data protection and should be used carefully. This technique is often used for data that are not related to the UID, but to some attributes. Some of the most popular methods for data pseudonymization are [18]:

- encryption – hiding data by encrypting it,
- shuffling – mixing data inside one column to disassociate attributes from the original user,
- suppression – removing sensitive columns from the dataset,
- redaction – completely removing parts with sensitive data.

Next to these two techniques, there are some other techniques for data privacy protection. Data generalization presents a way to change a value of some column with its range. For example, the value of column age 34 is modified to the range 30-40. Data synthetization presents a method to generate a completely new dataset from the original one using machine learning techniques. A newly generated dataset mimics the properties of original data. Although, these methods are useful in terms of data privacy, they distort information and reduce the data entropy. This is especially evident during the data aggregation.

IV. TELECOMMUNICATION DATA

Telecommunication networks represent a very important factor in the development of humanity. Due to their importance and high competition on the market, telecommunication operators gather data to further optimize services, reduce costs, and improve quality of their networks and services. Data are gathered on all network architecture levels, from physical and core network devices to the application layer. Modern telecommunication network providers typically implement centralized platform based on big data technologies to gather such amount of data.

 ISBN 978-86-7466-930-3

TABLE I
SAMPLE OF SIMPLIFIED MOBILE DATA SET

base station id	calling party	called party	billing	call type	call status	timestamp
bst_drwx_sth	aa9annch2n44	34yv5n9dwqlo	0.2	sms	sent	2022/03/15 08:45:02.000
bst_drwx_sth	i23u6bu546bx	vsf78bt489aa	0.0	call	start	2022/03/15 08:46:16.015
bst_drwx_sth	cw59coj7q33h	64hqu89bu33q	1.4	call	end	2022/03/15 08:46:55.724
…						
bst_mmss_nrt	j4kk9txbryuq	zevynom2w84t	7.2	call	end	2022/03/15 09:33:51.992
bst_mmss_nrt	i23u6bu546bx	quuhe11bcyd4	1.6	call	end	2022/03/15 09:34:08.183
bst_mmss_nrt	2jj3hb56u2bb	oppqnn4bb60u	0.2	sms	sent	2022/03/15 09:34:40.000
…						
bst_lndn_wst	i23u6bu546bx	pp22djqmxirb	1.7	call	end	2022/03/15 12:02:45.501
bst_lndn_wst	curbskoqcgg4	bsgwivg4611b	0.0	call	start	2022/03/15 12:03:10.017

Since telecommunications are always on the top of IT standards, all these systems already implement mechanisms for data privacy that are prescribed in their countries. Although, the implemented mechanisms mask the UIDs, due to the complexity and variety of data, some data privacy breaches may happen. In this section, we show one example of how the data privacy breach in terms of user tracking may happen and propose a masking mechanism that solves the problem for such a scenario while keeping the data entropy at the same level.

Mobile network operators gather data (for example, CDRs (Call Detail Record) and XDRs (Extended Detection and Response)) from base stations. This data contain information about the mobile phone device id, used frequency channel, signal strength, service type, call duration, etc. Such data helps the operators to tune the base stations, maximize the capacity of the cell, and quality of service. An example of simplified data set is shown in Table 1. In the example data set, UID is hashed due to the data privacy regulations. Assuming the anonymization is taken as the hashing mechanism, there is no way to get the UID original value.

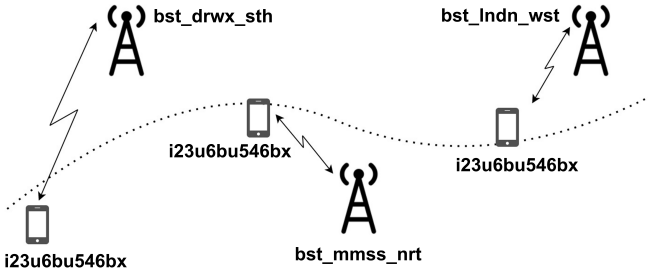

Fig. 1. User location tracking using data from base stations.

Since the same mechanism for data hashing is implemented on the data from every base station, it generates the same hashed UID for a user on every base station. This means that even though the user's UID is not known, it is possible to track user's movement as shown in Fig. 1. From the dataset shown in Table 1, it can be seen how the UID "i23u6bu546bx" is located on different base stations during the day which implies the possibility to track the location of this person. By matching the hashed UID with associated data (like base station ID, call records, etc.) it might be even possible to determine the identity of the user.

Although collected data are used for network improvement, people's privacy can be indirectly harmed in terms of location tracking. Since this information is not relevant to mobile network operators, we propose a new mechanism for data masking. The previous mechanism takes the device identifier (e.g. phone number or MAC address) and creates the hashed id by using some hashing function like shown in (1).

$$hashed\_UID = hash\_function(UID) \quad (1)$$

This mechanism creates the same hashed UID on all base stations. To improve this, we propose to create the hash UID from a combination of base station identifier and UID, as shown in (2). In this way, hashed value of UID will be different at each base station. Thus, information about user movement is removed, but all the necessary information relevant for quality-of-service monitoring are kept.

$$hashed\_UID = hash\_function(base\_station\_ID + UID) \quad (2)$$

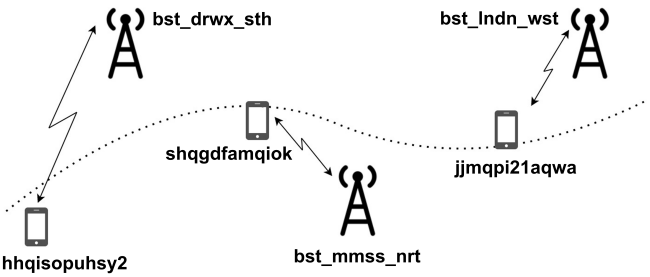

Fig. 2. Avoiding user location tracking.

The data set generated with the proposed hashed function is shown in Table 2. User that corresponds to hashed UID "i23u6bu546bx" in Table 1, has now a different UID on each base station which makes it impossible to correlate the data and track its location. Note that corresponding fields are colored in Tables 1 and 2. Fig. 2 shows the same example as Fig. 1 with a difference that now the UID is hashed with our proposed method. It can be seen in Fig. 2 that hashed UID values for the same user are different at different base stations. Thus, tracking of user movements is not possible anymore. However, if a user is connected to same base station for a

 ISBN 978-86-7466-930-3

TABLE II
SAMPLE OF SIMPLIFIED MOBILE DATA SET WITH PROPOSED HASHING MECHANISM

base station id	calling party	called party	billing	call type	call status	timestamp
bst_drwx_sth	qheuzhnq91nw	34yv5n9dwqlo	0.2	sms	sent	2022/03/15 08:45:02.000
bst_drwx_sth	hhqisopuhsy2	vsf78bt489aa	0.0	call	start	2022/03/15 08:46:16.015
bst_drwx_sth	soek28dh17h3	64hqu89bu33q	1.4	call	end	2022/03/15 08:46:55.724
…						
bst_mmss_nrt	h1ikvgwotggk	zevynom2w84t	7.2	call	end	2022/03/15 09:33:51.992
bst_mmss_nrt	shqgdfamqiok	quuhe11bcyd4	1.6	call	end	2022/03/15 09:34:08.183
bst_mmss_nrt	q11y2ms9h5b6	oppqnn4bb60u	0.2	sms	sent	2022/03/15 09:34:40.000
…						
bst_lndn_wst	jjmqpi21aqwa	pp22djqmxirb	1.7	call	end	2022/03/15 12:02:45.501
bst_lndn_wst	rxihb6ihqwhb	bsgwivg4611b	0.0	call	start	2022/03/15 12:03:10.017

period of time, hashed UID value will be the same, thus, the information about the signal quality for the session and user is preserved.

The same data privacy leak can happen not only in mobile but in other networks as well. For example, HFC (Hybrid Fiber Coaxial) network operators gather data from their cable modems. Considering that nowadays most of the clients are connected to the internet using WiFi (Wireless Fidelity), poor quality of service can be caused either by the poor signal quality on a cable modem or poor WiFi signal. The quality of the WiFi signal depends on many aspects such are the position of the cable modem, the schema of the building, types of walls, etc. Even though HFC operators are not in charge of the quality of WiFi networks, they tend to help clients to improve quality of network either by reconfiguring the WiFi (switching WiFi channel, changing channel width), relocating cable modem to some other place where it will better cover the whole apartment or by adding WiFi extenders.

Information about the cause of the poor signal quality operators find in the data gathered from the cable modems. Modern cable modems have embedded WiFi transmitters. Next to the basic information regarding the quality of the signal, data about the WiFi can be collected as well. Example of data that are gathered is the MAC (Media Access Control) address and WiFi username of the connected device (e.g., from a laptop, tablet, or mobile phone). If the device is connected to several locations that are covered by the same provider (e.g., at home, at café, store, work) the data privacy in terms of movement tracing can be breached. To solve such problem, the same hashing mechanism, (2), we propose for mobile networks can be used. Creating a different hash for UID on different cable modems, would completely remove the possibility to track the user movements.

V. CONCLUSION

Data privacy concerns are raised a few years ago and present one of the most important aspects during the development of data management systems. Depending on country, data management systems implement the appropriate privacy laws. Although the laws are obeyed, data privacy leakage may still happen if implementation is not carefully done. This paper shows an example of location data privacy violations in telecommunication data systems. In addition, we show methodology for how such violations can be solved with domain-specific knowledge. Finally, a proposal for data privacy improvement while keeping the same data entropy is given.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] Internet World Stats (site: https://www.internetworldstats.com/stats.htm), Accessed: Mar. 18, 2022.
[2] J. Dean, S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *OSDI - Operating Systems Design and Implementation*, 2004.
[3] Complete Guide to GDPR Compliance (site: https://gdpr.eu/), Accessed: Mar. 18, 2022.
[4] California Consumer Privacy Act(site: https://oag.ca.gov/privacy/ccpa), Accessed: Mar. 18, 2022.
[5] The PRC Personal Information Protection Law (site: https://www.china-briefing.com/news/the-prc-personal-information-protection-law-final-a-full-translation/), Accessed: Mar. 18, 2022.
[6] Health Insurance Portability and Accountability Act of 1996 (site: https://www.cdc.gov/phlp/publications/topic/hipaa.html), Accessed: Mar. 18, 2022.
[7] A. Cuzzocrea, "Privacy-Preserving Big Data Stream Mining: Opportunities, Challenges, Directions," in *proc. of ICDMW 2017*, New Orleans, LA, USA, Nov. 2017.
[8] M. L. Damiani, C. Cuijpers, "Privacy Challenges in Third-Party Location Services," in *proc. of MDM 2013*, Milan, Italy, June 2013.
[9] H. Liu, W. Di, "Application of Differential Privacy in Location Trajectory Big Data," in *proc. of ICITBS 2020*, Vientiane, Laos, Jan. 2020.
[10] C. Yin, J. Xi, R. Sun, J. Wang, "Location Privacy Protection Based on Differential Privacy Strategy for Big Data in Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3628 - 3636, Aug. 2018.
[11] S. Wang, R. Sinnott, S. Nepal, "Privacy-protected place of activity mining on big location data," in *proc. of Big Data 2017*, Boston, MA, USA, Dec. 2017.
[12] C. Wu, Y. Guo, "Enhanced user data privacy with pay-by-data model," in *proc. of Big Data 2013*, Silicon Valley, CA, USA, Oct. 2013.
[13] X. Feng, "The Optimization of Privacy Data Management Model In Big Data Era," in *proc. of IAEAC 2021*, Chongqing, China, Mar. 2021.

 ISBN 978-86-7466-930-3

[14] Y. Canbay, Y. Vural, S. Sagiroglu, "Privacy Preserving Big Data Publishing," in *proc. of IBIGDELFT 2018*, Ankara, Turkey, Dec. 2018.
[15] R. Mahesh, T. Meyyappan, "Anonymization technique through record elimination to preserve privacy of published data," in *proc. of PRIME 2013*, Salem, India, Feb. 2013.
[16] K. M. Shrivastva, M. A. Rizvi, S. Singh, "Big Data Privacy Based on Differential Privacy a Hope for Big Data," in *proc. of ICRCICN 2014*, Bhopal, India, Nov. 2014.
[17] What is data privacy? (site: https://www.cloudflare.com/learning/privacy/what-is-data-privacy/), Accessed: Mar. 18, 2022.
[18] Which data protection methods do you need to guarantee privacy? (site: https://www.statice.ai/post/data-protection-techniques-need-to-guarantee-privacy#:~:text=Encryption%3A%20hiding%20sensitive%20data%20using,entirety%20of%20a%20column's%20values), Accessed: Mar. 18, 2022.

 ISBN 978-86-7466-930-3