CRIME DATA PRIVACY

Agasthya Harekal
445 waupelani drive, state college, PA

Abstract

This paper presents crime data based privacy concerns and provides certain solutions to avoid privacy issues in this data. The paper reviews papers regarding crowdsourced data, crime based data to understand the privacy concerns and the measures presented to address the concerns.

1. Introduction

The work is based on crime dataset retrieved from dataworld website. This dataset has been defined as atlanta crime data. This data presents crime data 2009-2017. This data was made public at atlanta police dept open data website.

The data has information regarding exact crime, date of crime, location of the crime, neighbourhood and the latitude and longitude of the place of attacks.

Here, the crime serves as the sensitive information about the victim who had to face the crime. The quasi-identifiers include location variable that reveals the street address, neighbourhood variable, date, and the latitude, longitude data. From the longitude and latitude data provided we can get the exact street address where the crime has occurred. From this information anyone can trace the victim and identify the victim who is a known person. The zipcode of the person is easily identifiable. Further, in order to create data privacy scenario, we can synthesise additional columns victim's age-group, sex, race which creates a better privacy problem. Along with latitude and longitude data, these details act as quasi-identifiers to identify the victim.

2. Literature Review

In [1], researchers describe the basic definition of crowdsourcing platforms, its potential and benefits. The researchers describes privacy and its dimensions and a layered framework is used to analyse privacy risks from user and service provider perspective. The paper also describes techniques for privacy enhancement and does a systematic review of papers related to crowdsourcing and privacy. Further [1] conducts systematic review of papers that are related to crowd sourcing and privacy. Based on results, it was found there are 635 papers on privacy in crowdsourcing. Here, researchers choose major approaches such as framework, algorithm, model, survey to the concerns. Here, researchers understand that two areas are poorly researched that is gender and privacy, individual privacy perceptions. These papers were examined in terms of privacy layers, principles, concerns, enhancements.

In [2], authors discuss theoretical foundations of research on crowdsourcing, where at first it describes the emergence on crowdsourcing, types of crowdsourcing, such as ones proposed by Geiger et al, Nakatsu et al, Brahbham. Here, paper mentions that Brahbham's typology has emerged as one of the most referenced approaches. Further, the paper describes about various definitions and interpretations of privacy, and its interpretation in context of crowdsourcing. Researchers describes privacy threats in various domains. The paper describes strategies to tackle threats and challenges in crowd sourcing. Here, the paper mentions several approaches described in other papers such as mathematical models, quality control model. Further, generic principles for privacy protection is described, such as controlled linkability and composability, which means id of the crowd workers should not be linked to IDs in other contexts. Next principle is sensitive data encryption, which means that system should provide an option for crowd workers to encrypt their data such as end-to-end encryption. Next principle is security and accountability controls, which requires crowdsourcing system to exert granular data access control depending on the contextual

privacy risks in a crowdsourcing task. Final principle mentioned is transparency, which expects the requesters to inform crowd workers about usage and purpose of their data.

As part of strategy the researchers provide bi- directional and double screening as a solution which involves screening of crowd workers and their tasks from requester's perspective and also elimination of spamming tasks with certain criteria, such as implementing attention check questions or screening submissions based on crowd workers' reputation record. From the crowd workers' direction there also should be a double screening mechanism where requesters are screened and the screening of tasks posted.

Next strategy is to build templates and sandboxes for different types of crowd sourcing tasks. Different templates have to build for different crowd sourcing tasks, as they guide in successful implementation of regulations and privacy protection strategies. Further, sandboxes have to be provided for the crowd workers, to practice and become familiar with different types of crowd sourcing tasks and their threats and consequences.

The last strategy that is mentioned is to incentivise crowdsourcing platforms to more secure and humane. Also, crowdsourcing platforms can incorporate a practice of receiving reviews by external agencies and sharing their results. The paper concludes by noting that it has thrown ample light on privacy threats in crowdsourcing and the challenges in crowdsourcing.

In [3], authors discuss about crowdsourcing platforms, online services that have risks of re-identification when anonymised data sets are made public, profiling, data misuse and various issues of privacy in crowdsourcing platforms and overview of existing solutions.

In [4], researchers describes certain recommendations to keep in mind while constructing crowdsourcing platforms. Apart from that the paper discusses various privacy, security and data protection issues that were identified under various stages such as retrieval and selection stage, situational awareness stage, decision support systems. In the retrieval stage, researchers discuss the need for training of workers in crowdsourced platforms collecting geographical information(disaster platforms) and need for validation of data and assessment of accuracy at this stage. Volunteers must collect information according to international law and in accordance to rules. In the situational awareness stage, crisis management agencies must develop guidelines for general users, crisis reporters, journalists and reports must always ask for options for responders to be anonymous.

In [5], the researchers discuss about the various challenges of data collection in crowdsourcing and discuss certain solutions. Here, researchers discuss about privacy preserving crowdsourcing data collection where the threats and solutions to the threats are discussed.

In [6], data perturbation techniques are discussed for protecting privacy. Here, researchers discuss about differential privacy, local differential privacy, dimensional reduction is discussed.

In [7], the mobile crowdsourcing modes, taxonomy of applications, system framework, workflow of applications, challenges and solutions are mentioned. The researchers mention that the two modes required to mobilise users to mobile crowd platform is direct mode and WoM modes.The direct mode adopts the centralised control, in which the MCS platform accepts an upload task from a crowdsourcer, divides it into micro-tasks and co-ordinates crowd workers when they conduct the task iteratively. In contrast, WoM mode, task is released to a pool of initial workers and gets circulated in a worker-to-worker fashion. The taxonomy shows crowdsourcing applications explicitly divided into two categories human intelligence and human sensor. The human intelligence utilised human wisdom whereas human sensor utilises the concept of human-as-a-sensor and involves collecting human observations. Further researchers explain MCS system framework that is divided into modules where each module is explained int detail. Further paper describes typical workflow of application using the modules explained. The researchers explain about challenges, solutions by discussing about task design that crowdsourcer uses to describe about the task, incentives that are required to lure the workers, challenge of security and privacy that is presented, quality control issues that arise and how they can be solved. Further researchers discuss desired properties of MCS, integrated MCS schemes.

In [8], researchers discuss about types of sensitive information that are collected and gives an example. Also, researchers present solution by using an example of forgere where techniques like anonymisation are discussed. The paper also discusses dissemination technique that is used in the example and further details of the example.

Through [9], researchers want to provide strong guarantees of user privacy preservation in mobile CBCR services while allowing law enforcement authorities to share the data with third party data mining service providers.

The researchers provide mobile crime report framework that acts as a proof-of-concept crime reporting framework to test the anonymisation approach. This framework involves the concept of mobile interface, third party cloud storage provider where the crime incident is reported from the mobile interface and the law enforcement agency might store the anonymised records of reported crime incidents at the third party services. Public key is made available to users wanting to make crime reports so that all messages encrypted with the public key only be decrypted by crime report service. The mobile interface uses unstructured supplementary service data (USSD) protocol as a simple text exchange which is synchronous so crime reports can be recorded and a response returned to the user.

The user can provide responses to structured queries that are used to form the report. When the report is completed, the information is encrypted and stored in crime report database. Our crime report service handles the data transmitted from the mobile interface by logging into the database and decrypting the report using using the private key and then selects the anonymity that is required. The storage at the level of cloud storage provider is handled using a two layer encryption scheme in a manner that is similar to the solution proposed Kayern et al. In this approach cloud storage provider will double encrypt the data received from the data owner and will transmit double encryption key to the owner. The data owner shares the key with users requesting access. The user gets the anonymised data.

Since existing data anonymity algorithms are better suited to handling numeric as opposed to categorical values. So the researchers propose a new data anonymity algorithm that improves over existing algorithm by including a heuristic to minimise information loss and while maximising user privacy. This is achieved by weighting algorithm to prioritise information loss on numerical values over categorical values and second, privacy is maximised by computing both the frequency of equivalence classes as well as frequency of occurrence of sensitive attributes.

The researchers discuss in [10] about crime statistics that are available online. The researchers discuss about the benefits such as social scientists having the need to categorise data according to several dimensions, policy makers having a need to drill down updated crime statistics on a national level as well as on regional level to understand the developments of safety in different parts of the country. Citizens may use the data to underpin their decisions. Researchers also reveal some of the undesired effects of publishing crime statistics online and aspects of management of crime statistics online. Here, researchers reveal that when one publishes mean age of sex offenders each year, based on analytics of gender, profession in a city, one can expose full identity of the person violating the privacy law. Data mining technologies may expose the identity of group of individuals which may lead to stigmatisation of these groups. Here, data warehouse approach and data space approach may be used which may meet up this demand. The data extracted from different databases, are cleaned and transformed into a format that can be loaded into a data warehouse and a metadatabase maintains data about data stored in DWH and different monitors are derived from the DWH. A data space consist of set of different databases.

In [11], authors mention that 90% of adult americans are concerned about possible misuse of personal information in crime data. With regarding to criminal records made public 90% felt that state agencies should not use internet to publish conviction records, 33% for arrest records. When asked if one has ever been a victim of invasion of privacy, 38% of respondents said that they were victims of business collecting and using information, charitable, political or non profit organisation, law enforcement agencies, government tax, social service, welfare or licensing agency. Further detailed discussion and statistics is mentioned by researchers.

In [12], researchers throw light upon how criminals can use internet as a tool for their crimes, and importance of database storing these data to be secure and measures that can be taken to keep them secure. Researchers further discuss about human rights, issues of privacy involving crime data. Also, the data storage and retention issues are discussed. The researchers conclude that they hope that they are able to show results from profiling using data based on anonymised data which may require improved forms of anonymisation.

2. Gap Assessment

In [1] describes the crowd sourcing, the various privacy concerns, measures that can be taken to ensure privacy of individuals is preserved. The

paper reviews other papers of the approaches that these papers describes and provides results. The [1] does not itself provide any details of a concrete technology based solution to the privacy problem nor does it give any details of other research papers which have worked on the same. The [2] describes various types of crowdsourcing by mentioning approaches in various papers and then mentions privacy protection principles. The paper [2] describes in detail the privacy threats in various domains. This is mentioned by discussion of various examples. This is followed by discussion of privacy challenges in crowdsourcing where the paper describes generic principles for privacy protection and the solutions to the privacy concerns. However, we understand that although the describes the privacy threats using examples in detail, it does not provide specific solutions for these threats. In [3], researchers discuss about crowdsourcing platforms, their privacy issues and give us an overview about only existing solutions such as data anonymisation, data obfuscation. In [4], researchers describe certain recommendations for crowdsourcing platforms and privacy, security, data protection issues in various stages. But researchers do not mention about any methods to resolve issue of data privacy. In [5], researchers reveal methods to protect crowdsourcing from privacy threats but do not throw any light on how data collected can be published or given to third parties without infringement of privacy of the workers. In [6], researchers discuss only about existing methods to effectively publish data in a privacy preserving manner but do not disclose any new algorithm/technique of their own for this cause. In [7], researchers present modes, frameworks, challenges and solutions for mobile crowdsourcing platforms but no disclose about any new methods where disclosure of data does not hamper the privacy of individuals who participated. In [8], researchers mention about example where sensitive information is handled. Although the example uses existing solutions such as anonymisation, dissemination. In [9], researchers provide a framework for crowdsourcing of crime data and mention in detail how this mechanism works. They also mention a new data anonymity algorithm that is suited to handle categorical values. In [10], researchers make us understand the importance of making crime data public and help us understand the undesired effects of making such data public. There is no mention of any solutions to solve the privacy concerns. In [11], researchers discuss about various statistics of crime data hampering the privacy of victims/individuals and how measures should be taken in order to avoid such privacy issues. In [12], researchers discuss about issues with criminal data and discuss solutions using anonymisation.

SOLUTION

The solution to data privacy problem in Atlanta crime data is not to reveal quasi-identifiers such as date of crime, street address of the victim, neighbourhood, age, race, sex, latitude and longitude data. Some of these synthesised quasi-identifiers like age-group, race, sex were found in NYPD crime data.

As some of these variables can be used for analysis of crime data, we do processing of the data such that even if they revealed, the victim cannot be identified. Here, the date variable is converted into variable that stores month and year of the crime. Also, the longitude and longitude data were retained but they were rounded-off to last two decimal places. This avoids the street address of the victim not being revealed when looked up in internet. Further the location variable revealing the street address is removed. The synthesised variables added are such that no further processing need to be done to ensure privacy. These quasi-identifiers apart from those already in data aid in identifying the victim.

After all processing on data is done to ensure privacy, thorough analysis of the utility of the data is done. Here we understand that there is no change in the utility of the data. Although street wise analysis cannot be done, region wise analysis of crime data in the city can be done with the longitude and latitude data. This is because after processing of latitude and longitude data, the position did not change much and remained almost same. Analysis of types of crimes that occur for every region in the city using latitude and longitude data can be done. Also, similar analysis can be done for every neighbourhood.

CONCLUSION

In this work, we highlight the concerns of privacy in Atlanta City crime data. Through many researcher papers, we understand various concepts regarding crowdsourcing based data, the mechanisms involving collection, distribution of such data and concerns of privacy. Also, we understand privacy concerns of crime data which is published online. We understand the solutions presented by these papers towards the privacy

concerns. After taking cues from these papers, we try to solve the privacy concerns in the data in such a way that the utility of the data is not hampered in any way.

References

1. Alkharashi A, Renaud K. 2018. “Privacy in Crowdsouring: A Systematic Review”. Springer, Cham.11060, 387-400.
2. Huichuan Xia, Brian McKernan. 2020. “Privacy in Crowd Sourcing: A review of Threats and Challenges”.29, 263-301
3. Arik Friedman, Vijay Sivaram, Roksana Boreli. 2015.”Privacy in crowd sourced platforms”. Springer, Cham.
4. Buddhadeb Halder. 2016. “Privacy, Security and Data Protection in Crowd Sourcing Platforms: Issues and Recommendations”. SSRN.
5. Yunhui Li, Liang Chang, Long Li, Xuguang Bao, Tialong Gong.2021.”Key Research Issues and Related Technologies in Crowdsourcing Data Collection”. Hindawi. Vol 2021,13.
6. Milena Debprada Jena, Sunil Samanta Singhar, Bhabendhu kumar Mohanta, Somula Ramasubbareddy. 2021. “Ensuring Data Privacy Using Machine Learning for Responsible Data Science”. Intelligent Data Engineering and Analytics. 507-514.
7. Yufeng Wang, Xueyu Jia, Qun Jin, Jianhua Ma. 2016.“Mobile crowdsourcing: framework, challenges, and solutions”. Concurrency and Computation Practice and Experience. 29(3).
8. Lakdhar Meftah. 2019. “Towards Privacy-sensitive Mobile Crowdsourcing”. University of Lille.
9. Mark-John Burke, Anne V.D.M. Kayem. 2014.“K-Anonymity for Privacy Preserving Crime Data Publishing in Resource Constrained Environments”. 28th International Conference on Advanced Information Networking and Applications Workshops.
10. Sandra Kalidien, Sunil Choenni, Ronald Meijer. 2010.“Crime statistics online: potentials and challenges”. 131-137. Proceedings of 11 Annual Digital Governement Conference on Public Administration Online: Challenges and Opportunities. 131-137.
11. Lawerence A Greenfield.2001.“Public Attitudes Toward Uses of Criminal History Information”. Bureau of Justice Statistics.
12. Brian C Tompsett, Simon Prior. 2006. “Problems of Privacy, Security, Identity, Integrity, Legality and Confidentiality in Internet Crime investigation and evidence collection.” European E-Crime and Computer Evidence.