Design and Analysis of Privacy Models against Background Knowledge in Privacy - Preserving Data Publishing

by

NIDHI NITIN DESAI
201421005

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

November, 2021

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Doctor of Philosophy at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Nidhi Nitin Desai

Certificate

This is to certify that the thesis work entitled DESIGN AND ANALYSIS OF PRIVACY MODELS AGAINST BACKGROUND KNOWLEDGE IN PRIVACY - PRESERVING DATA PUBLISHING has been carried out by NIDHI NITIN DESAI for the degree of Doctor of Philosophy at *Dhirubhai Ambani Institute of Information and Communication Technology* under my supervision.

Prof. Manik Lal Das
Thesis Supervisor

Acknowledgments

Firstly, I express my gratitude to the Almighty for giving me strength and blessings during this long journey.

I am sincerely thankful to my supervisor and mentor, Prof Manik Lal Das, for providing guidance and motivation throughout this long journey. His direction, guidance, motivation, generosity and deep knowledge has inspired me to work towards reaching this goal. I am privileged to have Prof. Manik Lal Das as my mentor, who has helped me overcome my laxity and weaknesses.

I am thankful to my research and synopsis committee, Prof. Anish Mathuria, Prof. Sanjay Srivastava, Prof. Prasenjit Majumder and Prof. Sourish Dasgupta, for their valuable inputs and guidance. I am thankful to all the Professors of DA-IICT for their inspiration. I am thankful to all the staff members of DA-IICT for their support.

I am thankful to all my family members for their support and blessings. I am dedicating my thesis to my father (Mr Nitin Desai), mother (Mrs Bhavna Desai), nanima (Mrs Chandraprabha Bhatt) and fufi (Dr Chandrika Desai), who are special to me and have encouraged me to complete this work.

I am grateful to my senior researchers, Dr Sarita Agrawal, Dr Naveen Kumar, Dr Payal Chaudhari and Dr Hardik Gajera, for their help, motivation and valuable insights during discussions. I am thankful to all my lab members at DA-IICT. Special thanks to Hemantha, Shruti and Jinita for their technical help and support.

I am grateful to all my friends and fellow researchers at DA-IICT. Special thanks to Anjali madam, Archana, Miral didi, Nupur didi, Purvi didi, Purvi, Sarita didi, Sujata and Vandana didi for their support and help during the doctoral studies.

Lastly, I would like to acknowledge the contribution of one and all for their direct or indirect help in working towards the goal.

Contents

Abstract

Humongous amount of data gets collected by various online applications like social networks, cellular technologies, the healthcare sector, location - based services, and many more. The collected data can be accessed by third - party applications to study social and economic issues of society, leverage research, propose healthcare and business solutions, and even track a pandemic. As a result, online collected - data is a significant contributor in recent times. Despite the umpteen usefulness of online collected - data, it is vulnerable to privacy threats due to the presence of sensitive information of individual(s). Adding to that, the adversary has also become strong and powerful in terms of capabilities and access to knowledge. Knowledge is freely available in the public domain from sources like social profiles, social relations, previously published data and many more. As a result, privacy - preserving data publishing is a challenging research direction to venture upon. Our work mainly focuses on designing privacy models against background knowledge. Briefly, background knowledge is knowledge present with adversary used to disclose privacy of the individual(s). This makes background knowledge highly uncertain and inaccurate in nature as we cannot quantify the amount of knowledge present with the adversary. In this work, we design and analyze privacy solutions based on background knowledge. First of all, we propose an adversarial model against background knowledge and analyze existing and prominent privacy models against it. Secondly, we propose a privacy model $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private against background knowledge. The background knowledge assumption is comprehensive and realistic, which makes the proposed privacy model more strong and comprehensive in nature. The proposed privacy model has been theoretically analyzed against a strong adversary. Also, the proposed privacy

model has been evaluated experimentally and compared with existing literature. Progressively, our research work extends to Social Networks, which is an important application of privacy - preserving data publishing. Social network data has become an important resource in recent times but is prone to privacy threats. Thirdly, we propose a privacy model named Rule Anonymity against rule - based mining techniques in social networks. The rule - based mining techniques can predict unpublished sensitive information by generating rules. This makes it a challenging adversarial assumption. A rule - based anonymization technique has been proposed that incorporates the Rule Anonymity principle. We analyze the rule - based anonymization technique against a strong adversary having the capability of rule - based mining technique. The experimental evaluation of the rule - based anonymization technique shows positive results in terms of privacy when compared with existing literature. Fourth, we propose a de - anonymization technique against adversary's background knowledge. The adversary's background knowledge considers a comprehensive background knowledge that is imprecise and inaccurate in nature. We suggested distance metrics that consider imprecise and inaccurate identification and structural information. The de - anonymization technique has been implemented on a real social dataset and exhibits positive results in terms of de - anonymization accuracy. Fifth, we propose a privacy - preserving technique against comprehensive adversarial background knowledge. We have evaluated the proposed privacy model $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private on the Adult dataset and Census Income dataset and compared it with existing literature in terms of privacy. For social networks, we have used the Facebook dataset to evaluate the proposed privacy models and techniques.

List of Principal Symbols and Acronyms

α	Upper Bund for Redundancy
Δ	Degree Difference
δ	Spurious Rules
θ	Semantic Threshold
A_m	Moderate Adversary
A_s	Strong Adversary
A_w	Weak Adversary
Adv_R	Adversary with Rule Generation Capability
BI	Background Information
$D(u^k, u^{'})$	Aggregate Distance
$D_A(u^k, u^{'})$	Attribute Distance
$D_S(u^k, u^{'})$	Structural Distance
K_D	Definite Knowledge
K_P	Personalized Knowledge
K_{Pr}	Probabilistic Knowledge
KE	Knowledge Extractor
lb	Lower Bound

PI	Public Information
R	Rules
RG	Rule Generator
SC	Social Connections
sp_{SA}	Spurious Sensitive Attribute Value
t_{conf}	Confidence Threshold
t_{sup}	Support Threshold
ub	Upper Bound
w_a	Attribute Distance Weight
w_s	Structural Distance Weight
BK	Background Knowledge
PPDP	Privacy - Preserving Data Publishing
QI	Quasiidentifiers
SA	Sensitive Attributes

List of Tables

List of Figures

CHAPTER 1

Introduction

In the digital age, data collection and sharing have gained tremendous momentum that attains the objective of analysing the data and computing the statistics to improve decision - making. With the utility of data sharing, there occurs the need to limit the disclosure of sensitive information of individuals. For example, the researcher wants to find out how many patients are suffering from novel coronavirus (COVID - 19) in a city. It will collect the data from all the hospitals, but the hospitals cannot provide the data as it will disclose the individual's identity as per the HIPAA privacy rules [36]. To address the issue, the hospitals will anonymize the data so that the usefulness of the data remains intact while individual privacy is preserved. This makes data privacy an important and challenging research direction to address.

Nowadays, Microdata tables (unaggregated information about individuals like voter registration, medical data) are increasingly published by organizations that are useful for trend analysis and medical research. Also, there is an increase in meticulously analyzing the published data to get some interesting results or information. It will indeed help in moulding government policies, business innovations and socio - economic growth of society. On the other hand, the adversary has ease of access to tons of information available in the public domain. Consequently, the adversary can link data present in the public domain with the published data to dig in sensitive information like medical information, bank details, assets and location. For example, if we take an example of a voter list present in the public domain itself, it possesses information like Name, Address, Zip - code, Birth

- date and gender. The voter list can be linked with the published medical information, which contains attributes like Visit Date, Zip - code, Birth - date, gender, and disease diagnosis. By joining the two data tables, i.e. Voter list and Medical information, the Zip - Code, Birth - date, and gender are common to both the data tables. As a result, records comprising Zip - code, Birth - date, and gender are linked to the disease, where the disease is sensitive in nature. Thus, an adversary can narrow down the search based on Zip - code, Birth - date and gender to link an individual to disease. This linking inhibits the privacy of the individuals, which leads to critical circumstances. Therefore, Privacy - Preserving Data Publishing came into existence.

1.1 Privacy - Preserving Data Publishing

1.1.1 Opportunities in Data Publishing

In the current decade, sharing and publishing data has become inevitable across the globe. Numerous online applications like social networking applications, healthcare applications, location - based applications, and many more, collect and outsource data to different third - party applications. This collected data, when rigorously analyzed, gives insightful information which can help design government policies, health care solutions, business innovations and research directions. For example, in the current COVID - 19 pandemic, countries built various apps [86], [88], [89] to track infected individuals' locations to determine high - risk zones where stringent actions are applied to stop the transmission chain. Moreover, when analyzed by researchers [84], [85], [87], the infected individuals' data can help understand the demographics of infected individuals and their movement [90]. It will help respective governments design solutions specifically to demographic factors like age, area, co - morbidities, and many more, which help curb the infections. The above example signifies that data publishing has become a significant contributor to helping the human ecosystem. Furthermore, we dis-

cuss the opportunities of data publishing as follows:

- **Data Analytics:** Data Analytics [91], [94] applies approaches like clustering, association rules on the data to generate insightful and valuable information. The information unravels unknown correlations and patterns useful to make decisions in different domains. Further, the data can also possess different characteristics [94] like volume, velocity, and variety, known as "big data." Big data analytics [92], [94] is helpful in the current scenario, as it considers complex, large, heterogeneous, real - time, incomplete data. More is the diverseness in the data published; more accurate and efficient can be the analysis. This results in data publishing play an essential role in data analytics.

- **Social Media Data Analytics:** In the current pretext, social media has dominated the data analytics domain. The social media data [93], [95] considers social interactions like status updates, tweets, posts, comments on social network sites like Facebook, Linkedin, Twitter, Instagram, Youtube. It also considers data [93] from blogs, news articles, QA and discussions forums. Once analyzed, social media data helps enhance user experience [93] in product preference, insight about current trends and even alert administrative authorities against crimes [96]. Social media data needs to be published in large number, making data publishing indispensable in social media data analytics.

- **Machine Learning:** Machine Learning [98] has gained drastic momentum in the present data - centric era. Various machine learning methods [100] like supervised machine learning, unsupervised machine learning and semi - supervised machine learning, applied to data, will lead to accurate decisions and predictions. Deep Learning [97], a subset of machine learning, has emerged as a promising approach for predictions in the big data scenario. Deep learning models [97], [99] use artificial neural networks, a multi - layered approach, for accurate predictions. Data plays an essential role in pre-

dictions and making machines more independent of human intervention. So, for the machine learning domain, data is the fundamental requirement, making data publishing predominant.

- **Business Analytics:** Business Analytics [102] has garnered the attention of the business fraternity in the decade of big data. Valuable observations and patterns can be obtained when data is analyzed using tools and techniques [101]. It [101] helps make business decisions, analyze consumer behaviour, plan future business goals, and analyze competitors' strategy. So, Data publishing is vital in the domain of business analytics.

- **E - Governance:** E - Governance is instrumental in providing essential services [103], [104] in agriculture, tax and water supply to the people. It will help design policies for social and economic growth and provide possible solutions to handle poverty in developing countries. Many countries [103], [104] across the world have incorporated E - governance due to its numerous benefits. Nevertheless, data is the backbone for the smooth and effective functioning of e - governance, which can be only possible if data is published.

- **Research:** Data promotes leveraging research, academic as well as industry. Real - time and relevant data [106] help devise robust and realistic techniques and models in the research domain. A recent example is the covid pandemic [105], where scientific solutions are developed by analyzing the published data. Certainly, Data Publishing sees ample scope in the field of research.

- **Healthcare:** In the current decade, the healthcare domain is evolving digitally. This evolution will generate tons of data that can generate useful information. This information [107] helps in the timely treatment of the patients. However, to transform the healthcare industry, meaningful and real - time data needs to be published.

- **Disaster Management:** Disaster Management [108], [109] requires analysis of real - time data to get information about the missing people at times of earthquake, floods, avalanches. Inevitably, data publishing is instrumental in disaster management.

In a nutshell, Data publishing envisions the ultimate solution to various existing problems universally.

1.1.2 Existing Issues in Data Publishing

Despite the extensive opportunities in Data Publishing, it still comes under constant threat as the published data contains individuals' sensitive information. The examples of sensitive information in the published data are salary, medical condition, location, tax - related information. The disclosure of individuals' sensitive information in the published data results in dire consequences to their privacy in terms of dignity and trust. On the other hand, organizations (who have published data) have to face the law in case of a privacy breach. The existing threat results in individuals' having apprehensions about submitting data, which will challenge the very idea of data publishing. We discuss the existing issues in data publishing domain as follows:

- **Privacy Attacks:** Privacy attacks aim to disclose sensitive information by either linking the data with the external table like the voter's list or eliminate data based on the knowledge available. However, either of the approaches used will ultimately link an individual with its sensitive information, violating an individual's privacy. Privacy attacks on the published data not only break the trust of the individuals but also harm the reputation of the organization. Various privacy attacks like linking attack [1], background knowledge attack [2], similarity attack [3], skewness attack [3] are discussed in the literature. Nevertheless, the data publishing domain suffers from privacy attacks.

- **Data Publisher is unaware of the capability of the adversary:** In current times, the adversary has widened its access to knowledge and manipulation capabilities. The adversary can access the published data and manipulate it to disclose the privacy of individuals. However, data publisher can not change or update the privacy requirement after the data gets published. This requires the data publisher to be knowledgeable and consider a comprehensive and realistic privacy solution. So, data publisher needs to be aware of the adversarial capabilities to protect the data from privacy attacks.

- **Trust deficit across the individuals:** Instances of privacy attacks have resulted in a trust deficit atmosphere across individuals. Either the individuals do not submit data or submit inaccurate data. Ultimately, this will make data analysis less efficient and will result in a barrier to the cause of data publishing. To restore the trust of individuals, organizations need to ensure the privacy of their data by incorporating strong privacy protection solutions. This will motivate individuals to publish their data.

The issues in data publishing limit opportunities with apprehensions in publishing the data. It paves way for incorporating practical and realistic privacy solutions so that the data publishing domain can flourish. A potential solution is the Privacy - Preserving Data Publishing domain. Specifically, Privacy - Preserving Data Publishing (PPDP) [30] addresses privacy issues by publishing online data collections while preserving the privacy of an individual or group of individuals.

In the following sections, we discuss the architecture, research challenges and applications of Privacy - Preserving Data Publishing.

1.1.3 Architecture

Privacy - Preserving Data Publishing (PPDP) architecture consists of basically three components, namely, *Data Publisher*, *Data Recipient* and *Data Owner*. Figure 1.1 shows the PPDP architecture.

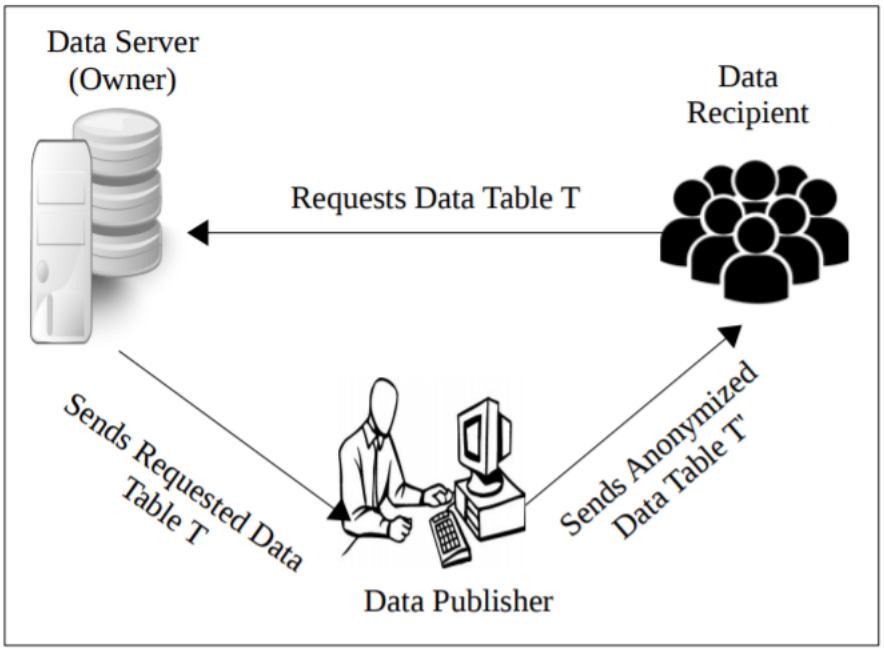

Figure 1.1: Privacy - Preserving Data Publishing Architecture

Data Recipient requests the online collected data (Data Table T) from the *Data Owner* (the data owner possesses the data server). *Data Owner* sends the online collected data (Data Table T) to the *Data Publisher*. The *Data Publisher* applies techniques to preserve privacy and send the privacy - preserved data (Anonymized Data Table T') to *Data Recipient* (here, *Data Owner* and *Data Publisher* can be the same entity).

The data (microdata table) collected by *Data Owner* consists of external identifiers, quasiidentifiers and sensitive attributes. The external identifiers (for example, Social Security Number) are the attributes that explicitly identify individuals. The quasiidentifiers are the attributes that implicitly help identify the individuals. Commonly used attributes for quasiidentifiers are Zip - code, Gender and Age with respect to a record that represents a person. Sensitive attributes are the attributes that could reveal the privacy of an individual. Some examples of sensitive attributes are disease, occupation and name. The anonymized table consists of quasiidentifiers and sensitive attributes, excluding the external identifiers. The

Data Publisher publishes data by anonymizing it to prevent individual identification and disclosure of sensitive attributes.

1.1.4 Research Challenges

The domain of Privacy - Preserving Data Publishing addresses privacy concerns while maintaining the utility. As a result, Privacy - Preserving Data Publishing has two significant research challenges: privacy and utility.

Privacy

Publishing data is a challenge as it contains sensitive information about individuals. The adversary's objective is to disclose sensitive information with the help of the available information. On the contrary, privacy models [32] protect the published data from the privacy attack against the adversary. Initially, the adversary [1] had access to tables like the voter's list. The adversary links [1] the voter's list with the published table to disclose the identity of individuals. Moving a step further, the adversary [2], [8], [10] started to have more specific information about a group of individuals. The information is in the form of individual - specific quasi-identifier information like location, age, gender. Besides, the adversary expands its scope in terms of information [3], [5], [7], [9], [10] available in the public domain like correlations, semantics and demographics. The above information will help in narrowing down to link an individual with its sensitive information. The adversary had further access to information that was not certain but probabilistic [7], [11]. The above information will only help the adversary further in linking the sensitive information with the individuals. The scope of adversary broadened and advanced to social networks [24], [25] in terms of structural information like the number of friends, number of friend's friends, structural properties of social networks. Furthermore, the adversary increased its capability to infer [7], the sensitive information not present in the published data.

The above discussion emphasizes the fact that the horizon of the adversary has

broadened in terms of access to information as well as its manipulation capabilities. Besides, an individual's information is readily available in the public domain in terms of profiles, published tables, and social relations. The ease in availability of information makes the adversary more resourceful. On the other hand, privacy models were proposed to address the privacy attacks in the Privacy - Preserving Data Publishing domain. Some prominent privacy models like k - anonymity [1], l - diversity [2], t - closeness [3] protect against privacy attacks due to information available with the adversary. In general, the information assumed are tables like voter's list, individual quasiidentifier information, access to the published data along with information related to semantics , demographics and data distribution.

In a nutshell, the adversary has become strong and progressive in terms of manipulating capabilities. Consequently, it has made published data more vulnerable to privacy threats and paved the way for more comprehensive privacy models. As a result, privacy is a significant challenge in Privacy - Preserving Data Publishing against strong and dynamic adversaries.

Utility

In the current times, published data has been a significant contributor to the decision making process. The main objective for publishing data is to mine critical observations and statistical information, which helps design policies, business innovations and analyze market trends. As a result, maintaining the usefulness of published data is essential and a challenge in privacy - preserving data publishing.

Preserving privacy and maintaining utility is equally crucial for the published data. Nevertheless, privacy and utility are not two individual research challenge but interrelated ones. In the privacy - preserving data publishing domain, researchers have discussed the privacy vs utility trade - offs versatilely. [118] dis-

cussed the curse of high dimensional data in k - anonymity which takes a toll on utility. [117] proposed utility measures for k - anonymity and l - diversity privacy models. Rastogi et al. [111] proved that no privacy solution would work against an adversary with too much information. It [111] also proposed a privacy - utility trade - off solution for bounded adversaries. There are various works [113], [114], [115] that consider utility in the privacy solutions. Further on, Brickell et al. [116] treats the privacy and utility trade - offs in terms of privacy loss and utility gain. It directly compares privacy loss and utility gain on the same privacy preserved data and concludes that utility is affected when perfect privacy is gained. Li et al. [110] treats differently by mapping the privacy - utility trade - off to risk - return trade - off, a concept from the Modern Theory portfolio. It measures privacy loss on privacy - preserved data while utility loss on the original data. Sankar et al. [112] proposed a framework for privacy - utility trade - off.

Briefly, the trade - off between privacy and utility for published data is application - specific. The data publishers adjust the privacy and utility for the published data based on different applications. Consequently, increasing privacy will decrease the usefulness of the published data. As a result, privacy and utility are two loggerheads in the field of privacy - preserving data publishing.

Concisely, the usefulness of data and the privacy of data is collectively vital in the current scenario and a potential research challenge.

1.1.5 Applications

The prominent applications which implement Privacy - Preserving Data Publishing are as follows:

Social Networks

Social network data is one of the most significant contributors to analyzing data and generating important observations. Despite its usefulness, it is prone to privacy attacks due to the presence of sensitive information.

Gross et al. [83] studies the patterns that reveal information in online social networks. The observed patterns [83] are information about user names and profile images on different social networks, information related to hobbies and interests, varied user information available on different social network platforms. The authors [83] carried out extensive experiments on more than 4000 students (users) of Carnegie Mellon University on the Facebook social networking site. They have observed that users tend to share personal information prone to privacy implications [83] like stalking, re - identification, building a digital dossier.

Smith et al. [80] highlighted the privacy issues related to big data. Nowadays, individuals [80] share geo - tagged media that has the image having the presence of multiple individuals (friends) as well as embedded location information. The individual that shares the media has control over their privacy, whereas friends of that individual can face privacy implications. Specifically, privacy implications arise when the media/data is not linked to an individual/friend and can be vulnerable to privacy issues. The authors [80] have analyzed and discussed privacy implications in the existing social networks.

Zhang et al. [81] discussed privacy concerns of an individual in the social networks in terms of identity anonymity, personal space privacy and communication privacy. Identity anonymity deals with preserving the privacy of the identity of an individual across different social network platforms. Personal space privacy considers privacy in terms of accessibility to an individual's profile across different social networks platform. Communication privacy focuses on privacy against the network operator in terms of time, location of the connection. Incorporating

privacy - preserving approaches helps protect the individual's privacy requirement [81] in terms of access control and unlinkability.

Zheleva et al. [82] discusses privacy breaches in social networks in terms of identity disclosure, attribute disclosure, social link disclosure and affiliation link disclosure. The identity disclosure occurs if a social profile is linked to an individual in a social network. The attribute disclosure results if the sensitive information of an individual is disclosed. If the sensitive relations between two individuals get disclosed, it will transpire to social link disclosure. Affiliation link disclosure links an individual to an affiliation group, where the link is sensitive. Further, the author discussed the existing privacy mechanisms to counter the privacy breach.

The above - discussed literature has brought light to the privacy breaches in social networks. To get the maximum benefits of social network data, it has also motivated to incorporate privacy - preserving mechanisms against the privacy breaches in the social networks. As a result, privacy - preserving data publishing provides solutions to the privacy concerns in social networks.

Location - based Services

Location - based services [79] have garnered attention in the recent decade, but geographic locations are considered sensitive information.

Location privacy [75] is restricting the access of an individual's current or prior location for the applications. Various location - based services access or track an individual's real - time location to provide services like nearby places, cafes, hospitals, medical stores and many more. However, location is a sensitive entity, but the complete restriction is not advisable due to its usefulness. The location can be accessible to other applications in a controlled manner such that individual privacy is preserved. The authors in [75] have suggested a privacy mechanism to replace real identities with pseudonyms to protect against privacy attacks.

Liu [77] discusses two types of location privacy namely the personal subscriber level privacy and the corporate enterprise - level privacy. The user controls the access to its location for the applications in the personal subscriber level privacy. In the corporate enterprise - level privacy, corporate managers control the individual's location in consideration of the enterprise's requirements. Location disclosure is a threat that can harm the privacy of an individual, and as a result, privacy - preserving mechanisms need to be incorporated.

Chow et al. [76] has bifurcated the location - based services to snapshot location - based services and continuous location-based services. In snapshot location - based services, the service provider takes the individual's current location to access their services. In contrast, the service provider of continuous location - based services tracks the individual's location, either periodic or continuous, to access services. Despite a prominent presence in the applications like the intelligent transport system and business analysis, location - based services violate individual privacy. Moreover, continuous location - based services are more vulnerable as individual location information can be inferred based on the location trajectory. The authors in [76] discussed the privacy mechanisms like the clustering - based approach and generalization based approach for the location privacy problem.

Niu et al. [78] addressed a privacy issue where the untrusted location - based services server can release information to third party applications. The authors in [78] suggested a dummy location selection solution that achieves k - anonymity. The research works in [79] have given insights on privacy issues and their respective solutions in the domain of location - based services. Privacy-preserving mechanisms like k - anonymity is one of the possible solutions to address the privacy attacks. Moreover, location data is a crucial stakeholder in various applications like tracking health emergency, logistics, cabs, accidents, and many more.

On the other side, privacy issues [119] in location data can cause harm to the individuals in terms of stalking, burglary. Location - based services can regain individual's trust and safety to submit location data by integrating robust and practical privacy solutions. Privacy-preserving data publishing inevitably addresses privacy concerns in location - based services.

Healthcare Applications

Healthcare applications make life simpler, but medical conditions and insurance information are sensitive information. Healthcare data [73] can be extensively obtained from electronic health records [121], insurance claims, interconnected devices, social media posts and locations, and many more.

The authors in [73] have divided privacy violation into consequentialist concerns and deontological concerns. In consequentialist concerns [73], the privacy violations will result in negative consequences affecting the individual. An example of consequentialist concerns is the disclosure of sensitive diseases like HIV, schizophrenia. It impacts the reputation of individuals leading to serious implications like anxiety problems and mental illness. Conversely, in deontological concerns [73], individual does not suffer from detrimental negative consequences. However, the individual's data can be accessed by applications even if not used. Though some privacy violations are innocuous but do affect the individuals as they lose control over their data.

Further, the authors in [71] have discussed information privacy threats in terms of organizational threats and systemic threats. The organizational threat [71] occurs when an agent access an individual's (patient) data with malicious intentions. The agent can be an insider or an outsider to the organization. In systemic threats [71], a legal agent residing in the organization can access the individual's (patient) data.

In contrast, Data Protection laws [72], [122] are proposed by different countries to protect the privacy of individuals. For example, Health Insurance Portability and Accountability Act (HIPAA) Act in the USA, IT Act and IT (Amendment) Act in India, Data Protection Directive in the EU. The authors in [72] have discussed de - identification, one of the methods to protect against privacy disclosure. The de - identification [72] removes the information that can help identify the individuals and disclose their privacy. The examples are k - anonymity, l -diversity, t - closeness.

The authors in [74] focus on the individual's (patient) perception in the healthcare domain. In general, individuals (patients) have a dilemma in submitting sensitive information due to the fear of getting it disclosed to their social circles. This limits the individuals (patients) in sharing their medical information. Practical privacy solutions need to be designed such that medical research and innovations can be helpful to society. As a result, privacy - preserving data publishing helps preserve privacy concerns in healthcare applications.

Apart from the applications mentioned above, privacy - preserving data publishing is important where sensitive data is involved.

1.2 Motivation

Data publishing has a multitude of opportunities in the era of the digital revolution. Tons of data gets collected by numerous applications which contain rich, diverse and useful information. Third - party applications can procure the collected data and apply machine learning and data mining techniques to generate insightful observations and predictions. This will facilitate the enhancement of humankind by designing more pro - people socio - economic policies, address real - time health solutions and provide prompt response to catastrophic calamities. Moreover, it will anticipate cutting - edge technological and social innova-

tions and motivate research. Despite innumerable advantages, the published data comes under persistent threat to privacy disclosure due to the presence of sensitive information. This leads to individual(s) resisting in submitting the data accurately and confidently. As a result, privacy - preserved solutions need to be incorporated in the published data.

Various privacy models were proposed in the literature by the researchers against privacy threats. Firstly, k - anonymity [1] model provides a solution to the linking attack. The adversary links the anonymized table with the external table (e.g., the voter's list) based on the common identification information. Thus, it discloses the individual's identity. k - anonymity [1] model protects against identity disclosure. Moving on, l - diversity model addresses the privacy attacks in the k - anonymity model. The adversary uses identification and demographic information to disclose the sensitive information of the individual. l - diversity [2] protects against attribute disclosure. Further, t - closeness [3] model addresses the limitations of l - diversity model. The adversary uses the knowledge about semantic closeness and knowledge of global distribution of sensitive information of the published data to disclose the individual's sensitive information. t - closeness protects against attribute disclosure.

Background knowledge has been playing an important role in the privacy model used in data publishing. In present scenarios, background knowledge has evolved significantly from specific information to social networking profiles; and from published tables to crawling information on the Internet. The background knowledge is commonly described in the sentential form. However, sentential form is not closely bounded as it can be interpreted with different perceptions and different views. Different background knowledge assumptions by the privacy model is as follows:

- An Individual A stays at Zipcode 15030.

- Prostate Cancer does not occur to females.
- High Blood Pressure leads to high risk of Heart Disease.
- Japan has low incidence of Heart Disease.
- Individual A's mother has flu therefore individual A has flu.

The above examples of background knowledge are in sentential form. However, the pertinent question to address is: are these background knowledge similar or different? Before coming to an answer, we first explain the context of the above - defined background knowledge forms.

The first form of background knowledge gives specific information about an individual and its resident zip code. The second form of background knowledge gives information about the occurrence of prostate cancer in a specific gender. The third form gives us information about the inference related to two different diseases. The fourth form gives us information related to the demographics of Japan. The fifth form of knowledge gives information about social connections and inferences based on the property of communicable diseases.

All the above forms of background knowledge are obtained from different sources like social networking sites, research conclusions, facts, observations done on various studies, research conclusions, and many more. In short, all the above forms of knowledge do signify different meaning and context but still comes under the umbrella of background knowledge. As a result, modelling background knowledge will give more insights in designing more stronger, realistic and comprehensive privacy models.

In summary, data privacy has become inevitable in the current data - centric world. Published data is susceptible to privacy attacks orchestrated by the adversary. In particular, the adversary has access to diverse public domain resources

like published tables, external tables (e.g., voter's list) to specific individual information, and many more. With times, the adversarial capabilities have amplified in terms of powerful manipulation capabilities. Selective adversarial knowledge assumption is naive in the current scenario when information is freely available in the public domain. The privacy solution against selective background knowledge will pave the way for other knowledge attacks. Moreover, the privacy disclosure risk will stop individual(s) from submitting data with more conviction. As a result, more stringent and realistic privacy solutions need to be designed against strong and realistic adversarial assumptions.

The above discussion motivates us to study the background knowledge in the privacy - preserving data publishing domain. Further, it encourages us to propose a practical and robust privacy model against comprehensive background knowledge.

1.3 Contribution of the Thesis

In the current era, data has been an important asset. Massive volumes of data get collected from various applications like social networks, location services and many more. The collected data is meticulously mined to study various facets of society. However, at the same time, despite its usefulness, collected data is prone to privacy attacks as it contains sensitive information of individuals. As a result, Privacy - Preserving data publishing has been a challenging research direction.

Our contributions addresses privacy issues in the domain of privacy - preserving data publishing and its application i.e., Social Networks. We summarize our contributions as follows:

- **Privacy - Preserving Data Publishing:** Background knowledge is a potential privacy concern in the domain of privacy - preserving data publishing. A stronger and comprehensive privacy solution will help protect data

against strong adversarial assumptions. Our contributions related to privacy - preserving data publishing are as follows:

- **Modelling Background Knowledge in Privacy - Preserving data Publishing:** We study some of the prominent privacy models in the literature. We have observed that privacy attack due to background knowledge is evitable in the previous privacy models. As a result, we study background knowledge in detail. Background knowledge is an amalgamation of different variants of knowledge. We propose an adversarial model against background knowledge attack. In the adversarial model, we define different types of adversaries based on its capabilities. We analyze existing privacy models against the proposed adversarial model. This study has motivated us to devise a strong privacy model against the adversarial model.

- **Privacy Model against Background Knowledge:** We propose a privacy model $(\theta, [lb, ub]^{+sp}, \alpha)$ Privacy against adversarial background knowledge. We have assumed a strong adversary with access to comprehensive background knowledge. We use the concept of semantic dissimilarity for partitioning data and adding spurious records into the actual data to protect the individual records from privacy disclosure. The objective in adding spurious records to the data is to increase the complexity in terms of guesses and manipulations for the adversary. We have theoretically analyzed the proposed privacy model against a strong adversary. We use two datasets, namely Adult dataset and Census Income dataset from the UCI machine learning repository for experiments of proposed privacy model. The experimental results also show positive results in terms of privacy when compared with existing literature.

- **Privacy - Preserving Data Publishing in Social Networks:** Data Privacy in social networks is an important research direction to address. Due to privacy issues, individuals either do not publish sensitive information or publish it inaccurately. This makes social network data highly inaccurate and inconsistent. Due to the complex structure of social networks, privacy models of relational databases can not be directly incorporated. We address privacy concerns in the field of social networks data publishing.

  - **Rule - Based Anonymization against Inference Attack in Social Networks:** We study the collective inference attack [6], which predicts the unpublished sensitive information using the identification information as well as social relations of neighbours. It used a rough set theory approach to predict the sensitive information. Cai et al. [6] also proposed a data sanitization method against rule - based mining techniques. We have observed weakness [151], [154] in the data sanitization technique [6]. We also propose a rule anonymity model [151] against rule - based mining techniques. It provides a strong privacy guarantee such that the presence of rules should show a negligible impact on the privacy of sensitive information. We have assumed an adversary with capabilities of rule generation irrespective of specific techniques. We have proposed a rule - based anonymization technique that incorporates the principle of rule anonymity. We have theoretically analyzed the proposed technique against a strong adversary. We have used a Facebook dataset from Stanford Large Network Dataset Collection for experiments. The experimental results show a positive impact in terms of privacy against rule - based mining against the existing literature.

  - **De - anonymization against Background Knowledge in Social Networks:** We propose a de - anonymization technique [152] against the adversary's background knowledge. We have assumed that the ad-

versary has comprehensive background knowledge that also considers imprecise and inaccurate semantically similar information . The proposed distance metrics capture the imprecise and inaccurate attribute and structural information into the picture. The proposed de - anonymization technique DeSAN incorporates an aggregate distance - based approach for de - anonymizing users in published social networks against background knowledge. We have used a Facebook dataset from Stanford Large Network Dataset Collection for experiments. The experimental results of the proposed de - anonymization technique show positive results in terms of de - anonymization accuracy. We also propose a privacy preserving technique against comprehensive adversarial knowledge.

1.4 Thesis Outline

The Thesis is organized as follows:

- Chapter 2 emphasizes the literature study and prerequisites required in the thesis. Section 2.2 discusses the prerequisites in terms of attribute type and basic steps of anonymization. Section 2.3 discusses the prominent privacy models like k - anonymity, l - diversity, t - closeness and Differential Privacy. Section 2.4 shows the strengths and limitations of the discussed privacy models. Section 2.5 concludes the chapter.

- Chapter 3 concentrates on background knowledge. Section 3.1 studies Background Knowledge in Privacy - Preserving Data Publishing. In Section 3.2, knowledge sets are discussed. Section 3.3 proposes a Adversarial Model against background knowledge. Section 3.4 analyzes privacy models against background knowledge. Section 3.5 concludes the chapter.

- Chapter 4 proposes a strong privacy model against background knowledge. Section 4.1 introduces the need for a strong privacy model against background knowledge. Section 4.2 shows the implication of Background Knowledge on Published data using an example. Section 4.3 provides a broader perspective on semantic knowledge. Section 4.4 defines the definitions and concepts in building the privacy model. Section 4.5 proposes the privacy model against background knowledge. Section 4.6 presents an algorithm to implement the proposed privacy model. Section 4.7 theoretically analyze the privacy model with strong adversarial assumptions. Section 4.8 evaluates the proposed privacy model by using a real dataset. Section 4.9 concludes the chapter.

- Chapter 5 studies social network, a very prominent application of Privacy - Preserving Data Publishing. Section 5.1 gives a brief overview about social networks. Section 5.2 discusses the transition from Relational Tables to Social Networks in terms of privacy disclosure. Section 5.3 studies privacy in social networks. Section 5.4 motivates in proposing privacy solutions for social networks. Section 5.5 concludes the chapter.

- Chapter 6 discusses inference attack and provides a rule - based anonymization technique against inference attack in social networks. Section 6.1 introduces inference attack due to rule - based mining techniques. Section 6.2 studies thoroughly the existing literature [6] of inference attack due to rule - based mining. It also shows weakness in the existing literature [6]. Section 6.3 models inference attack due to Rule - based Mining. Section 6.4 proposed a rule anonymity model. Section 6.5 presents a Rule - based anonymization technique that abides by the rule anonymity principle. Section 6.6 theoretically analyze the rule - based anonymization. Section 6.7 evaluates the experiments using a real social dataset. Section 6.8 concludes the chapter.

- Chapter 7 proposes a de - anonymization technique against background knowledge in social networks. Section 7.1 introduces the privacy issue due to background knowledge in Social Networks. Section 7.2 studies the ex-

isting literature. Section 7.3 discusses semantic knowledge in social networks. Section 7.4 discusses social network representation and the adversary's background knowledge. Section 7.4.3 defines distance metrics. Section 7.5 proposes a De - anonymization Technique named De - SAN against background knowledge. Section 7.6 presents a privacy - preserving technique against background knowledge. Section 7.6.2 theoretically analyzes the technique against the strong adversarial assumption. Section 7.7 evaluates the de - anonymization technique. Section 7.8 concludes the chapter.

- Chapter 8 concludes the thesis. It also paves the way for future research directions.

CHAPTER 2

Privacy Models

2.1 Introduction

Privacy - Preserving data publishing [32], [38] publishes the data with the help of anonymization techniques. Specifically, the privacy principles in Privacy - Preserving data publishing predominantly protects [3], [32] from identity and attribute disclosure. In privacy - preserving data publishing, the data publisher publishes the data table after applying the privacy principles, rather than analytical and statistical results. The adversary, too, has access to an entire anonymized data table instead of mining results obtained from triggering queries. Consequently, it makes the published data more susceptible as the data publisher can not change the privacy requirement based on the adversary's capabilities. Also, attacks due to background knowledge prominently dominate the literature of privacy - preserving data publishing. As a result, the Privacy - Preserving data publishing domain requires stringent privacy definitions and strong and practical adversarial assumptions to protect published data from privacy attacks.

We discuss the prerequisites for privacy - preserving data publishing in the next section. The prerequisites would facilitate understanding of privacy models in the literature.

2.2 Preliminaries

2.2.1 Types of Attributes

Anonymity [32], [37] is incorporated on the data table to preserve privacy. In general, a data table consists of three types of attributes:

1. Explicit Identifiers: Attributes that clearly identify individuals like SSN, Name and Address.

2. Quasiidentifiers : Attributes whose value taken together can identify an individual like Zip - Code, Birth - Date and Gender. The disclosure of these attributes needs to be done in a controlled manner.

3. Sensitive Identifiers/Attributes : Attributes considered sensitive like Disease and Salary. These attributes need not be disclosed to the third - party premises.

2.2.2 Basic Steps of anonymization

The basic steps needed to anonymize the data table are as follows:

1. Remove the Explicit Identifiers as they uniquely identify the individuals and inhibit individual privacy.

2. Anonymize the quasiidentifiers such that disclosure of the sensitive attributes linked to the particular individuals is restricted.

To implement step (2), anonymization operators [32] are used. Some prominent anonymization operators [32], [37], [38] are generalization, suppression, perturbation, permutation.

2.3 Some Existing Privacy Models

Various privacy models [1], [2], [3], [32] provide the intended quantum of privacy to ease information sharing that calibrate the analysis and decision - making process on published data. k - anonymity [1] protects against identity disclosure. (α, k) - anonymity [13] extends the k - anonymity with α - Deassociation requirement. It protects against attribute disclosure, specifically, homogeneity attack, such that the relative frequency of each sensitive attribute value in a k - anonymous partition is at most α. [13] doesn't consider background knowledge. MultiRelational k - anonymity [68] extends k - anonymity that considers multi relation setting. p - Sensitive k - anonymity [14] extends the k - anonymity principle such that for each k - anonymous partition, each sensitive attribute in a partition must occur at least p times. (k, e) - anonymity [15] provides privacy protection against numerical attributes such that each k anonymous partition contains at least k distinct sensitive attribute values and their range is at least e. The extensions [13], [14], [15] of k - anonymity provides method for attribute protection. (ϵ, m) - anonymity [18] provides the solution to proximity breach attack, specifically addressing numerical sensitive attributes. The attack occurs if the adversary can locate the range of sensitive attribute values of the given individual with high conviction while unable to guess the exact sensitive attribute value. The ratio of numbers of records whose sensitive attribute values satisfies ϵ - neighborhood ($[sensitive\ attribute\ value \pm \epsilon]$) to the size of the partition must be at most $\frac{1}{m}$. In Personalized Privacy [21], the individual manages the privacy of their sensitive attribute values by protecting attribute disclosure. [12] protects against membership disclosure such that such that the probability of presence of individual (record r) in generalized table T' when external table T is present $(P(r \in T'|T))$ is δ, where $\delta = (\delta_{min}, \delta_{max})$. [17] gives a solution to the sequential release of datasets where the datasets belong to the same data table. [17] suggests two privacy notion (X, Y) - anonymity and (X, Y) - Linkability. In (X, Y) - anonymity [17], each value in X is linked to at least k distinct values in Y. In (X, Y) - Linkability [17], the confidence of linking Y from X is at most threshold. Here, X is quasiidentifiers, and Y

is sensitive attributes. The above requirements protect from the disclosure of two datasets due to join operation. m - invariance [20] protects against privacy attack due to the republication of dynamic datasets. A sequence of generalized dataset [20] is m - invariant if every dataset is m - unique (each partition contains at least m records having different sensitive attribute values) and every record present in any/all of the releases have the same signature (sensitive attribute value). l - diversity [2] and t - closeness [3] protects against attribute disclosure. Differential Privacy [22] protects from membership disclosure with strong privacy guarantee. Specifically, we focus on privacy attack due to background knowledge with a strong privacy protection guarantee. We consider the following models for further analysis:

- k - Anonymity
- l - Diversity
- t - closeness
- Differential Privacy

We discuss the above - mentioned privacy models in the coming section.

2.3.1 k - anoymity

A table satisfies k - anonymity if every record in the table is indistinguishable from at least $k-1$ *other record with respect to every set of quasiidentifier attributes; such a table is called a k - anonymous table [1]*

The table is said to be k - anonymous if each record comprising of the quasiidentifiers is identical to at least k - 1 records. For example, Table 2.1 is 4 - anonymous.

In Table 2.1, the Zip - Code and age are the quasiidentifiers, which needs to be anonymized to protect from the linking with external tables like voter's list. In Table 2.1, Disease is a sensitive attribute, which should not be known to the adversary. Here, the quasiidentifier, Zip - Code, is anonymized by replacing the last two

	Non - Sensitive		Sensitive
No	ZipCode	Age	Disease
1	130**	<30	Heart Disease
2	130**	<30	Heart Disease
3	130**	<30	Viral Infection
4	130**	<30	Viral Infection
5	148**	≥40	Cancer
6	148**	≥40	Heart Disease
7	148**	≥40	Viral Infection
8	148**	≥40	Viral Infection
9	130**	3*	Cancer
10	130**	3*	Cancer
11	130**	3*	Cancer
12	130**	3*	Cancer

Table 2.1: An Example of *k* - Anonymous Data Table

digits with a '*' and the age is anonymized by replacing it with a broader range. In Table 2.1, for any given partition, each record is indistinguishable from the three other records based on quasiidentifiers; therefore, it is a 4 - Anonymous partition [1]. The entire data table follows the same property [4]. In other words, the adversary requires to remove at least three records to reveal the individual's health condition. However, the adversary could locate a particular partition where the individual resides but can not link disease value with the individual due to indistinguishability property.

Samarati [68] proposed the basic algorithm of *k* - anonymity for single domain minimal generalization. [65], [66], [67] proposed approximation algorithm. [70] extends *k* - anonymity to multi - relations. With a significantly large value of k, the privacy disclosure is less as the adversary requires to remove more records to narrow down to a single record. In other words, smaller is the value of k; larger is the possibility of the privacy disclosure. But with an exorbitant large k, the number of records in a partition would be larger, which would require a higher level of generalization or suppression. Though it will provide more privacy, it will reduce the utility of the data [69].

Weaknesses in k - Anonymity model:

The k - Anonymity model [2] has 2 attacks as follows:

1. Homogeneity Attack: k - Anonymity [2] can disclose the information due to lack of diversity in the sensitive attributes. For example, Table 2.1 is 4 – anonymous data table; however, if an adversary knows the zip - code and age of the individual and also has the information that record of the individual is present in the published table, i.e. for example if the adversary knows a 31 year old individual living in zip - code 13053. The adversary can easily locate the records 9,10,11,12 in the published data table. Now, in all four records, the disease is cancer; therefore, the adversary without any dilemma would conclude that the individual has cancer. This discloses the privacy of the individual. Homogeneity attack occurred as there was only a single sensitive attribute value (Here, disease), i.e. cancer in the partition; so the adversary can guess the disease (sensitive attribute) with a probability of 1.

2. Background Knowledge Attack: If the adversary personally knows the individual and has the knowledge that record is present in the published table; then k - Anonymity [2] is vulnerable. For example, in Table 2.1, if the adversary knows an individual personally, i.e. 21 year old Japanese who lives in zip – code 13068. Therefore, the adversary can easily locate that the individual's record resides in the first partition comprising record 1,2,3,4. Now, the adversary can guess that the individual either is suffering from heart disease or viral infection, but with a probability of 0.5. However, Japanese people have extremely low chances of heart disease. Therefore, an adversary can conclude that attacker is suffering from a viral infection; discloses the privacy.

2.3.2 l - diversity

l - diversity

Machanavajjhala et al. [2] suggested an l - diversity principle, which provides a solution to the privacy attacks in the k - anonymity. The table T consists of n partitions. *A partition is said to be l - diverse if there are l distinct values for the sensitive attribute.* A table T is l - diverse, if all the partitions in the table are l - diverse. Table 2.2 is 3 - diverse data table. Here, each partition has at least 3 - diverse sensitive attribute values.

	Non - Sensitive		Sensitive
No	ZipCode	Age	Disease
1	1305*	$\leq$40	Heart Disease
2	1305*	$\leq$40	Viral Infection
3	1305*	$\leq$40	Cancer
4	1305*	$\leq$40	Cancer
5	1485*	$>$40	Cancer
6	1485*	$>$40	Heart Disease
7	1485*	$>$40	Viral Infection
8	1485*	$>$40	Viral Infection
9	1306*	$\leq$40	Heart Disease
10	1306*	$\leq$40	Viral Infection
11	1306*	$\leq$40	Cancer
12	1306*	$\leq$40	Cancer

Table 2.2: An Example of l - diverse Data Table

l - diversity [2] principle ensures that the l sensitive attribute values be "well represented". One of the instantiations is Entropy l- diversity. Entropy L diversity [2]: *A table is Entropy l - Diverse [2] if for every partition, entropy of distribution of sensitive attributes is atleast* $\log(l)$. This depicts that the distribution of the sensitive attributes must be uniform in nature and quantitatively atleast $\log(l)$. The entropy of Partition 1 in Table 2.2 is 3. Similarly, the entropy of Partition 2 and 3 in Table 2.2 are 3 as well. As a result, the minimum entropy of (p_1, p_2, p_3) is selected. So, the Table 2.2 is 3 - diverse [2].

Table 2.2 is not vulnerable to the homogeneity attack [2] as discussed in section 2.3.1, as there are 3 diverse sensitive attributes values in each partition. Table 2.2 is not susceptible to the background knowledge attack [2] as discussed in section 2.3.1. Note that the same background knowledge assumption is taken. For example, the adversary knows a 21 - year - old Japanese individual who lives in zip code 13068. It can locate the partition in which the individual's data is present, i.e. partition 3 consisting of record 9, 10, 11,12. There are three diseases {Viral Infection, Heart Disease, and Cancer}. The adversary can eliminate heart disease from the probable list of diseases as Japanese people are less prone to heart disease. But still, the adversary can't guess the Disease between Viral Infection and Cancer.

Weaknesses [3] in l - diversity Model

The l - diversity model [3] has two attacks:

1. Similarity attack: The l - diversity [3] guarantees that the sensitive attribute in each partition is diverse in nature, but does not take semantic meaning into the picture. For example, if a partition consists of 3 diseases gastritis, stomach flu, stomach cancer. Though there are 3 different diseases in the partition, they are semantically close. The three diseases suggest that an individual is suffering from stomach related disease. Here, the privacy of the individual is targeted as the adversary can conclude that the individual is suffering from stomach related diseases. This thwarts the objective of disclosure of sensitive information.

2. Skewness Attack: If the distribution of the sensitive attributes is skewed, then the l - diversity [3] will not protect from attribute disclosure, i.e. new information about individuals are revealed. For example, if in a 2 diverse data, i.e. positive or negative for a deadly disease like HIV, the overall distribution is skewed (For example, from a total of 20 records, 16 are negative

and 4 are positive), which will lead to different interpretations of the partitions. For example, if in partition 1, there are 2 positive and 2 negative, however, it satisfies the 2 diversity constraint but would be interpreted that the 50 % of the population has a chance to be detected positive instead of overall 20% positive cases. If partition 2 has 1 positive and 3 negative, then the interpretation would be different; hence two partitions gives a different level of privacy risks [3].

2.3.3 t - closeness

Li et. al [3] suggested t - Closeness, a privacy solution to protect against l - diversity attacks. *A partition is said to possess t - closeness property if the distance between the distribution of the sensitive attributes in the partition and the distribution of the sensitive attributes in the entire table should not exceed threshold t.* Table 2.3 is an example of t - closeness. The distance between two distribution needs to be quantitatively less than threshold t. Lesser the distance, more identical are the distributions. This would signify that the correlation between the quasiidentifiers and sensitive attributes would be restricted such that the adversary does not get the clearer picture which leads to attribute disclosure. Earth Mover's Distance (EMD) [3] is used to calculate the distance between distributions

2.3.4 Differential Privacy

Differential Privacy [22] provides a potential privacy definition against the individual(s) in statistical data mining scenarios. The presence or absence of a particular individual does not affect the statistical analysis results is the confidence, differential privacy imbibes in the individuals at large. Eventually, it will motivate individuals (users) in submitting its information to databases accomplishing the differential privacy criteria. A more formal explanation of differential privacy [22] is as follows: Let there be two databases $D1$ and $D2$ with a difference of a single database item. A randomized function F satisfies ϵ - differential pri-

	Non - Sensitive		Sensitive
No	ZipCode	Age	Disease
1	1305*	$\leq$40	Heart Disease
2	1305*	$\leq$40	Viral Infection
3	1305*	$\leq$40	Cancer
4	1305*	$\leq$40	Cancer
5	1485*	>40	Cancer
6	1485*	>40	Heart Disease
7	1485*	>40	Viral Infection
8	1485*	>40	Viral Infection
9	1306*	$\leq$40	Heart Disease
10	1306*	$\leq$40	Viral Infection
11	1306*	$\leq$40	Cancer
12	1306*	$\leq$40	Cancer

Table 2.3: An Example of t - closeness Data Table

vacy if $Pr(F(D_1) \in S) \leq exp(\epsilon) * Pr(F(D_2) \in S)$ where $S \subseteq Range(F)$. Indeed, this gives a strong privacy guarantee to an individual against adversary having knowledge of all database rows except the individual. Differential privacy [22], [23] is achieved by adding noise to the statistical query results.

Initially, Dwork [22], [23] proposed Differential privacy for privacy - preserving data analysis scenario [40], but with time it has been extended to privacy - preserving data publishing field [41]. Differential privacy provides a strong privacy guarantee against membership disclosure. But, attribute disclosure and identity disclosure too are essential in privacy - preserving data publishing. As a result, recently, researchers [41], [42] have proposed an amalgamation of differential privacy with k - anonymity. However, privacy attacks due to background knowledge have shown dominance in the evolution of privacy models in privacy - preserving data publishing. In contrast, differential privacy does not discuss the effect of background knowledge in terms of identity and attribute disclosure.

2.4 Strength and Limitations of Privacy Models

In this section, we capture the strength and limitation of the privacy models discussed in section 2.3. Figure 2.1 gives a brief comparison of privacy models.

Privacy Model	Strength	Limitation
k - anonymity	1. Addresses Identity Disclosure. 2. Protects against privacy attacks due to : • Published tables like voter Id	1. Do not address Attribute Disclosure and Membership Disclosure. 2. Does not consider: • Personalized Knowledge • Demographic Knowledge • Correlational Knowledge • Semantic Knowledge • Data Distribution
l - diversity	1. Addresses Attribute Disclosure. 2. Protects against privacy attacks due to: • Personalized Knowledge • Demographic knowledge	1. Does not address Identity Disclosure and Membership Disclosure. 2. Does not consider: • Correlational Knowledge • Semantic Knowledge • Data Distribution
t - closeness	1. Addresses Attribute Disclosure. 2. Protects against privacy attacks due to: • Semantic Knowledge • Data Distribution	1. Do not address Identity Disclosure and Membership Disclosure. 2. Does not consider: • Correlational Knowledge • Demographic Knowledge • Personalized Knowledge
Differential Privacy	1. Protects against Membership Disclosure with strong privacy guarantee. 2. Protects against privacy attacks due to: • Personalized Knowledge	1. Do not address Attribute Disclosure and Identity Disclosure. 2. Does not consider: • Demographic Knowledge • Correlational Knowledge • Semantic Knowledge • Data Distribution

Figure 2.1: Strength and Limitations of Privacy Models

The thesis work is primarily focused on background knowledge attacks in privacy - preserving data publishing; therefore, we keep differential privacy out of the scope of this work.

2.5 Conclusion

This chapter gives an overview of the existing privacy solutions in privacy - preserving data publishing. First, we discuss the prerequisites necessary to under-

stand privacy models. We also discuss some prominent privacy models of privacy - preserving data publishing domain. Further, we capture the strength and weakness of privacy models.

CHAPTER 3

Background Knowledge

3.1 Background Knowledge in Privacy - Preserving Data Publishing

Knowledge in generic term is understanding or awareness about facts, ideas, concepts obtained through rigorous study [4]. However, the context of knowledge in the Privacy - Preserving Data Publishing field is associated with information the adversary has about an individual or group of individuals to disclose privacy. We discuss background knowledge in the context of privacy models and modelling background knowledge as follows:

Background Knowledge in Privacy Models

The privacy models have evolved substantially in the past decades due to the background knowledge attacks in the published data. Background knowledge can range from targeted individual information to generalized information present in the public domain. The adversary using the background knowledge can disclose the privacy of an individual or group of individuals. Initially, the adversary assumed external table like voter's list as background knowledge, which has some quasiidentifiers common with the anonymized table. The k - anonymity [1] model protects against identity disclosure due to the linking of external tables like the voter's list with the anonymized table. It incorporates generalization and suppression as anonymization operators. Nargiz et al. [12] proposed a privacy metric

δ - presence, which protects from membership disclosure. A table is δ - present if the probability of the presence of individual record in the generalized table when the private table is given is between δ_{min} and δ_{max}. The background knowledge assumed is external table like voters list; however, the external table is the superset of the generalized table. Further, Wong et al.[13] proposed a new privacy model named (α, k) - anonymity, which uses k - anonymity and α - deassociation properties together. Here, in a k anonymous partition, the relative frequency of pair of quasiidentifiers and its associated sensitive attribute value should not exceed α. It assumes background knowledge in terms of the external table. [14] proposed a privacy property named p - sensitive k - anonymity, which extends k - anonymity. It protects against attribute disclosure. It assumes background knowledge in terms of external table and level of generalization of quasiidentifier attributes. [15] proposed a privacy model named (k, e) - anonymity, where the published table satisfies k - anonymity property (Here, k - anonymity refers to each partition having at least k distinct sensitive attribute values for the given quasiidentifiers) and range of sensitive attribute values in the k - anonymous partition is at least e. It assumes background knowledge as multiple external tables. Adding to external tables, [16] assumes knowledge of the algorithm to disclose privacy. It suggests a privacy solution in terms of m - confidentiality. Wang et al. [17] proposed (X, Y) privacy, a privacy solution for sequential release of the tables. It assumes background knowledge as the external table, which has common quasiidentifiers with the published anonymized table. [18] proposes a privacy model named $(\epsilon, m)-$anonymity against proximity breach of numeric sensitive attributes. The background knowledge assumed by the adversary is external tables along with the quasiidentifier information of the individual, whose privacy needs to be disclosed. [19] proposed solution of freeform attack, which is depicted in the form $Q \rightarrow s$, where Q and s are attributes of any level. Privacy is disclosed if set attributes Q is associated with a sensitive attribute value s. In [19], the adversary assumed background knowledge as taxonomies of all attributes along with the anonymized table. [20] provides a privacy principle named m - invariance that provides a solution against dynamic re - publication problem that considers

insertion and deletion operations. In [20], the background knowledge assumed is external tables, generalization principle and lifespan of each tuple in the sequential release of tables. [21] assumes background knowledge in terms of external tables and proposes personalized anonymity that provides privacy of sensitive attribute based on the personalized preference of individual(s).

Progressively, l - diversity [2] addresses the weaknesses of k - anonymity, namely homogeneity attack and background knowledge attack. The adversary assumes background knowledge in terms of personalized knowledge and demographic knowledge. Further, t - closeness [3] addresses the weaknesses of l - diversity in terms of similarity attack and skewness attack. The adversary assumed knowledge in terms of semantic knowledge and data distribution. Recently, Dwork proposed differential privacy [22][23], which provides a strong privacy guarantee in terms of membership disclosure. The adversary assumed personalized knowledge.

Modelling of Background Knowledge

Kifer et al.[8] modelled background knowledge in terms of personalized knowledge which includes knowledge about individuals present in the partition of the published table and its entire quasiidentifier information. It obtained the quasiidentifier information from external tables [1]. Moving on, Chen et al. [10] modelled background knowledge in terms of personalized knowledge like individual's and its relations information related to sensitive values. [11] modelled background knowledge in terms of probabilistic data distribution and probabilistic personalized knowledge. [5] modelled knowledge related to negative correlations whereas [18] modelled correlational knowledge. Moving to social networks, [6][24][25][26][27] modelled background knowledge in terms of auxiliary social graphs, structural information, social relations and rule - based inferences. Lastly, the adversary increased its manipulation capabilities [7] by inferring unpublished sensitive attributes by exploiting personalized information supported by facts and

observations available in the public domain. Figure 3.1 summarizes background knowledge in privacy - preserving data publishing.

There are two challenges with respect to background knowledge. The first challenge is to define background knowledge. The definition of background knowledge is open - ended as each privacy model defines its own interpretation of background knowledge. The second challenge is to actually analyze the background knowledge definition of the existing privacy models. This will help in analyzing the effect of background knowledge in PPDP.

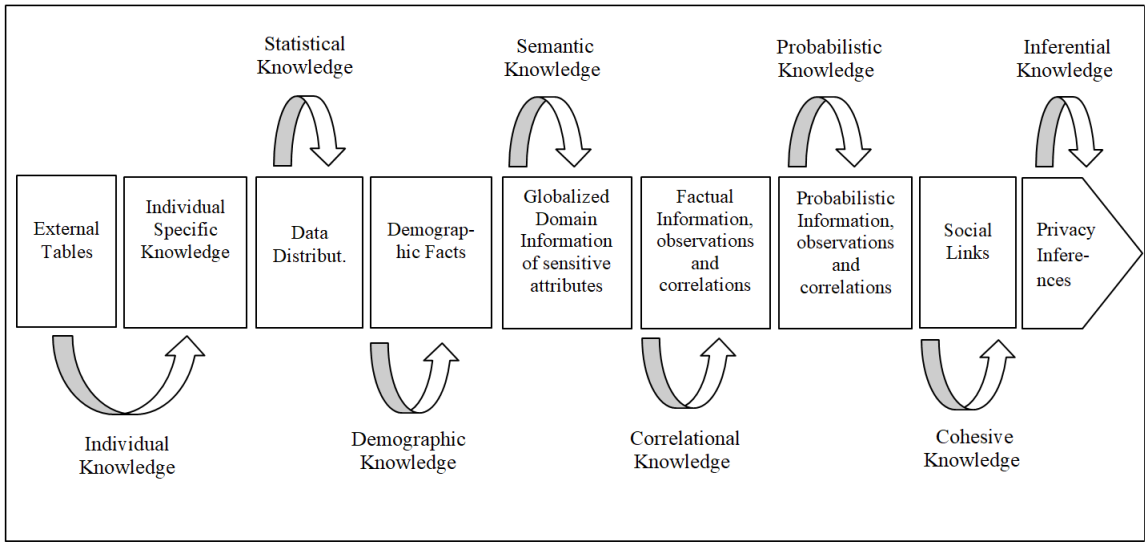

Figure 3.1: Background Knowledge in Privacy - Preserving Data Publishing

3.2 Preliminaries

Background Knowledge can help eliminate the records in the anonymized table. The elimination of the records is possible in two ways [2]:

- *Negative Disclosure.*
- *Positive Disclosure.*

In the *negative disclosure*, the records are eliminated based on the perception of Not going to occur with a higher conviction. For example, Prostate Cancer is not going to occur in Females. In *positive disclosure*, the records are constricted based on the perception of going to occur with a higher conviction. For e.g. Dengue has a high occurrence in New Delhi. The *positive disclosure* and *negative disclosure* approaches when applied with different sets of background knowledge helps in shrinking the probable choices into single convergence.

3.2.1 Notations

Table T is a micro - data table having n records. Schema of table T is defined as $T(q_1, q_2, \ldots, q_j, SA)$, where $\{q_1, q_2, \ldots, q_j\}$ are the set of quasiidentifiers and SA is sensitive attribute which consists of k distinct sensitive attribute values, $SA = \{SA_1, SA_2, \ldots, SA_k\}$. Table T' is an anonymized table that fulfils any given anonymity principle. Schema of table T' is defined as $T'(q'_1, q'_2, \ldots, q'_j, SA)$. Here, $\{q'_1, q'_2, \ldots, q'_j\}$ are the set of anonymized quasiidentifiers and SA is sensitive attribute which consists of k distinct sensitive attribute values, $SA = \{SA_1, SA_2, \ldots, SA_k\}$. D_{QID} and D_{SA} is the domain of quasiidentifiers and sensitive attribute respectively.

3.2.2 Knowledge Sets

In this section, we have used the term published table and external table. Note that the published table and the external table are two different entity. The external table consists of external identifiers as well as quasiidentifiers, whereas, published tables can be anonymized tables (that fulfil any given privacy requirement like k - anonymity etc.) or external tables. In the case of social networks, published tables are auxiliary social network and anonymized social network. Social networks can be represented in tabular form (Refer to chapter 6 and 7). Herewith, we will interchangeably use published tables/anonymized tables for social networks as well. We define the knowledge sets in terms of adversarial capabilities as follows:

Statistical Knowledge: Statistical knowledge is defined as knowledge obtained from the distribution of the data. It would further help the adversary in obtaining inferences and a broader generalized view about the published table. Moreover, rule - based inferences also help the adversary in [6] inferring unpublished sensitive attributes. For example, the people within the age range of 60 and 70 have heart disease with probability 1. This knowledge will help in changing the uniform probabilities of sensitive attribute values for the given quasiidentifier(s) in the published table. Statistical operators and published table help in generating statistical knowledge. Mathematically, it is represented in the form of function as follows:

$$F_{Stat}(op, T^*) \rightarrow STAT_K \tag{3.1}$$

Here, T^* is the published table, and op is the statistical operator. Here, we assume statistical operators to be rule generators, conditional probability, count. $STAT_K$ is the set of all statistical knowledge resulted due to the function F_{Stat} and is represented as $STAT_K = \{stat_{k_1}, \ldots, stat_{k_{stat}}\}$. Note that, statistical knowledge can also be obtained from the published table, where the given individual is not present, but we have assumed that the knowledge of the presence of the individual in the published table is known to the adversary.

Each element of set $STAT_K$ is of the form:

1. $((\bigwedge_{i=1}^{j} q_i : v_i) \Rightarrow SA_k, p)$, where $j \geq 1$, q_i is a quasiidentifier, v_i is the value of a quasiidentifier q_i, SA_k is the k^{th} sensitive attribute value and p is probability of occurrence of the instance of statistical knowledge. The value of p is $0 \leq p \leq 1$. For example, $((age : [60, 70]) \Rightarrow Heart\ Disease, 1)$. (Note that value v_i can be generalized or suppressed [1] too.) It can also be of the form

$Count(SA = SA_k) = n_{SA_k}$ where SA is the sensitive attribute, SA_k is the k^{th} sensitive attribute value and n_{SA_k} is the count value of the sensitive attribute value SA_k.

Individual Knowledge: Individual knowledge is defined as knowledge that is available with the individual itself (adversary in our case). Individual knowledge ranges from specific information related to the individual(s) like identification, sensitive information etc. to more generalized information like external tables. For example, Bob is of Japanese nationality. This is a personalized knowledge as Japanese nationality is known as apriori by the adversary. Individual information along with an anonymized table (where the given individual is present) helps in generating individual knowledge. Mathematically, it is represented in the form of function as follows:

$$F_{In}(IN^I, T^*) \rightarrow IN_K^I \tag{3.2}$$

Here, IN^I is information about individual I present with the adversary and T^* is the published table (here, anonymized table) in which I is present. IN_K^I is the set of all individual knowledge generated by function F_{In} and is represented as $IN_K^I = \{in_{k_1}^I, \ldots, in_{k_{In}}^I\}$.

Each element of set IN_K^I is represented in any of the two forms:

1. $(a\ (r)\ v_a, p)$ where a is attribute that can be quasiidentifier and/or sensitive attribute and/or any attribute that is not present in the anonymized table (but helps in privacy disclosure), v_a is value of attribute a, r is either $=$ or $\neq$ and p $(0 \leq p \leq 1)$ is the probability of occurrence of particular instance of individual knowledge. For example, an instance of individual knowledge is (ZipCode = 15036,1).

2. $\cup_{rec=1}^{x} t(\cup_{i=1}^{m} A_i)_{rec}$ where A is a set of m attributes comprising external identifiers and quasiidentifiers and x is number of records. The above form represents external table. An example of external table is voter's list.

Demographic Knowledge: Demographic knowledge is defined as the knowledge regarding the demographics of the individuals, specifically zip - codes/ locations/ countries/ city/ town/ village present in the published table. In our case, an adversary can use this knowledge to disclose privacy or shrink the probable choice. For example, a Malaria outbreak is not present in New York City at that instance. This information helps the adversary eliminate the record pointing to New York City and can shrink the probable choices. The probability of eliminating the records are 1. Demographic knowledge is generated with the help of information related to demography as well as the published table. Mathematically, it is represented in the form of function as follows:

$$F_{Dem}(PI_{Dem}, T^*) \rightarrow DEM_K \tag{3.3}$$

where PI_{Dem} is demographic information available in public domain, T^* is published table. DEM_K is the set of all demographic knowledge generated by function F_{Dem} and is represented as $DEM_K = \{dem_{k_1}, \ldots, dem_{k_{dem}}\}$.

Each element in set DEM_K is of the form:

1. $(q_i : v_i \ (r) \ SA_k, p)$ where q_i is a location related quasiidentifier, v_i is value of q_i , r is either $=$ or $\neq$, SA_k is the k^{th} sensitive attribute value and p is the probability of occurrence of the instance of demographic knowledge, which is 1 in this knowledge variant. For example, an instance of demographic knowledge is $(ZipCode : 15036 \neq Malaria, 1)$.(Note that value v_i can be generalized or suppressed [1] too.)

Correlational Knowledge: Correlational knowledge is defined as knowledge obtained by relating the quasiidentifier and sensitive attributes simultaneously in the published table. For example, Males do not have ovarian cancer. This correlation is generated when quasiidentifier value Male and sensitive attribute value ovarian cancer are explored together. As males cannot have ovarian cancer so the adversary can eliminate ovarian cancer from the probable choices. The probability of eliminating the records are 1. Correlational information in the public domain and published table help in generating correlational knowledge. Mathematically, it is represented in the form of function as:

$$F_{Corr}(PI_{Corr}, T^*) \rightarrow CORR_K \tag{3.4}$$

Here, PI_{Corr} is correlational information available in the public domain and T^* is a published table. $CORR_K$ is the set of all correlational knowledge generated by function F_{Corr} and is represented as $CORR_K = \{corr_{k_1}, \ldots, corr_{k_{corr}}\}$.

Each element in set $CORR_K$ is represented in the form:

1. $(q_i : v_i \ (r) \ SA_k, p)$ where q_i is a quasiidentifier , v_i is value of q_i , r is either $=$ or $\neq$, SA_k is the k^{th} sensitive attribute value and p is the probability of occurrence of particular instance of correlational knowledge which is 1 in this knowledge variant. For example, an instance of correlational knowledge is $(Gender : M \neq Ovarian\ Cancer, 1)$. (Note that value v_i can be generalized or suppressed [1] too.)

Probabilistic Knowledge: Probabilistic knowledge is defined as knowledge that is probabilistic in nature; that is, one cannot eliminate the records with probability 1. The adversary manipulates records probabilistically based on demographic

and correlational knowledge. For example, India has a high diabetic middle - aged population. This information will result in assigning high probability to the records where diabetes is linked to Indians whereas less probability to other records. Higher is the probability, higher are the chances of disclosure. The core difference between correlational knowledge and demographic knowledge is that the latter focuses on location, whereas the former is generic irrespective of location. Probabilistic knowledge considers both forms. The uncertain information based on demographics and correlations in the public domain and the published table helps generate probabilistic knowledge. Mathematically, it is represented in the form of function as follows:

$$F_{Pr}(PI_{Pr}, T^*) \rightarrow PR_K \tag{3.5}$$

where PI_{Pr} is probabilistic information of correlational and demographic knowledge available in the public domain and T^* is the published table. PR_K is the set of all probabilistic knowledge generated by function F_{Pr} and is represented as $PR_K = \{pr_{k_1}, \ldots, pr_{k_{pr}}\}$.

Each element in set PR_K has the same representation as demographic knowledge and correlational knowledge, except probability p. Here, the value of p is $0 \leq p \leq 1$ instead of 1.

Semantic Knowledge: Semantic knowledge is defined as knowledge related to the semantic similarity of the attribute domain in the published table. This knowledge helps in shrinking the probable choices considerably as it replaces the set of sensitive attribute values with a single broader sensitive attribute value. For example, if the set of the sensitive attribute contains {Gastritis, gastric ulcer, Chronic Gas}, it can be substituted with *stomach disorder*, a broader sensitive attribute

value. Semantic information related to sensitive attribute domain available in the public domain, and the published table helps generate semantic knowledge. Mathematically, it is represented in terms of function as:

$$F_{Sm}(PI_{Sm}, T^*) \rightarrow SM_K \tag{3.6}$$

Here, PI_{Sm} is Semantic information available in the public domain and T^* is published table. SM_K is the set of all semantic knowledge generated by function F_{Sm} and is represented as $SM_K = \{sm_{k_1}, \ldots, sm_{k_{sm}}\}$.

Each element in set SM_K is represented in the form:

1. $(SA_{sm} \approx SA_s)$ where SA_{sm} is a set of sensitive attribute values which are semantically similar and SA_s is a broader sensitive attribute value of set SA_{sm}. For example, an instance of semantic knowledge is $(Gastritis, Gastric\ ulcer, Chronic\ gas \approx Stomach\ disorder)$.

Cohesive Knowledge: Cohesive knowledge is defined as knowledge about social connections obtained from published social networks. The cohesive knowledge consists of structural information of social network like degree information, neighbourhood information, structural properties. Moreover, cohesive knowledge also considers information related to social relations and auxiliary social networks. For example, Alice has three friends i.e., degree information is 3. This knowledge will help the adversary eliminate users who do not have 3 friends and help in disclosing privacy. Information about social connections and the anonymized table (where the individual is present) helps in generating cohesive knowledge. Mathematically, cohesive knowledge is represented in the form of function as follows:

$$F_{Co}(CO^I, T^*) \rightarrow CO_K^I \tag{3.7}$$

Here, CO^I is cohesive information about individual I present with the adversary and T^* is anonymized table in which I is present. Note that social network can be represented in a tabular form (refer to chapter 6 and 7). CO_K^I is the set of all cohesive knowledge generated by function F_{Co} and is represented as $CO_K^I = \{co_{K_1}^I, \ldots, co_{K_{co}}^I\}$.

Each element in set CO_K^I can be represented in any of the two forms:

1. $(s = v, p)$ where s is the structural attribute, v is the value of structural attribute s and p ($0 \leq p \leq 1$) is the probability of occurrence of the particular cohesive knowledge. For example, an instance of cohesive knowledge is (degree = 3, 1).

2. $(Relation : a(r)v, p)$ where a is any attribute, v is the value of attribute a, $Relation$ is Individual I's social relation, r is either $=$ or $\neq$ and p ($0 \leq p \leq 1$) is the probability of occurrence of particular instance of cohesive knowledge. The above representation depicts information about individual's social relation. For example, an instance of cohesive knowledge is (Mother: disease = flu, 1).

3. It can also be represented as social network table(Refer to chapter 6 and 7).

Inferential Knowledge: Inferential knowledge is defined as knowledge used to infer the sensitive attribute of individual using the individual information and known facts available in the public domain. For example, if individual I has diabetes and Diabetes leads to heart - related disease then Individual I has Heart - related disease. Individual information, inferential information in the public domain and anonymized table (where the given individual is present) help generate

inferential knowledge. Mathematically, it is represented in the form of function as follows:

$$F_{Inf}(Inf^I, PI_{Inf}, T^*) \rightarrow INF_K \tag{3.8}$$

Here, Inf^I is the personalized knowledge (cohesive as well as individual knowledge) of individual I that can help in making inference and PI_{Inf} is inferential information available in the public domain and T^* is a published table (here, anonymized table) where I is present. INF_K is the set of all inferential knowledge generated by function F_{Inf} and is represented as $INF_K = \{inf_{k_1}, \ldots, inf_{k_{inf}}\}$.

Each element in set INF_K is of the form:

1. $(((ArB) \bigwedge (BrC)) \Rightarrow\ (ArC),\ p)$ where (ArB) and (BrC) are individual information and inferential information in public domain, respectively, whereas (ArC) is the inferred knowledge, r is a either $=$ or $\neq$ and p $(0 \leq p \leq 1)$ is the probability of particular instance of inferential knowledge. For example, ((I = diabetes), (Diabetes = heart disease) $\implies$ (I = heart disease), 0.8).

Definition 3.1. Background Knowledge: Given an anonymized table T' consisting of a set of quasiidentifiers $\{q'_1, q'_2, \ldots, q'_j\}$ and sensitive attribute SA, Background Knowledge is an assemblage of all defined knowledge sets i.e., statistical, individual, demographic, correlational, probabilistic, inferential, semantic and cohesive knowledge; which helps in disclosing the privacy of an individual or set of individuals. Mathematically, background knowledge is represented as:

$$BK = \{STAT_K \cup IN_K^I \cup DEM_K \cup CORR_K \cup SM_K \cup PR_K \cup INF_K \cup CO_K^I\} \tag{3.9}$$

3.3 Adversarial Model

The main objective of a data publisher is to publish the data table in an anonymized form such that the privacy of the individual(s) is preserved, and at the same time, data doesn't become useless. In similar terms, the adversary's goal is to disclose the individual's privacy using background knowledge . We discuss the mechanism of the adversarial model in the next section.

3.3.1 Mechanism

The adversary has background information about an individual or group of individual(s) subjective to its manipulation capabilities. The adversarial model mechanism has two functions: extracting background knowledge and linking individuals with their respective sensitive attributes. Firstly, the knowledge extractor extracts the background knowledge in terms of knowledge sets (Section 3.2.2) from background information. Secondly, Link will link the individual(s) and its respective sensitive attributes in the anonymized table using background knowledge. If the link leads to privacy disclosure, the adversary becomes successful else unsuccessful.

Figure 3.2 shows the mechanism of the adversarial model. Knowledge Extractor block takes as an input Background Information BI. Background information BI is background knowledge in an unstructured form that the adversary possesses to disclose the individual's privacy and is represented as $BI = \{b_1, \ldots, b_b\}$. Here, we assume the adversary has at least one piece of background information i.e., $b \geq 1$. The output of knowledge extractor block is Background Knowledge BK. The structure of BK is as per section 3.2.2. The Link block takes as an input

the Background Knowledge BK and anonymized table T'. Here, T' consists of n records, where each record consists of j quasiidentifiers and a sensitive attribute. Each record in T' is synonymous with an individual. Note that the anonymized table fulfils any given privacy requirement. Link Block links quasiidentifiers of the individual(s) with its respective sensitive attribute using BK Background Knowledge in an anonymized table T'. The output of the link block is Pr i.e., $P(T'|BK)$. We define the adversarial capabilities in the next section.

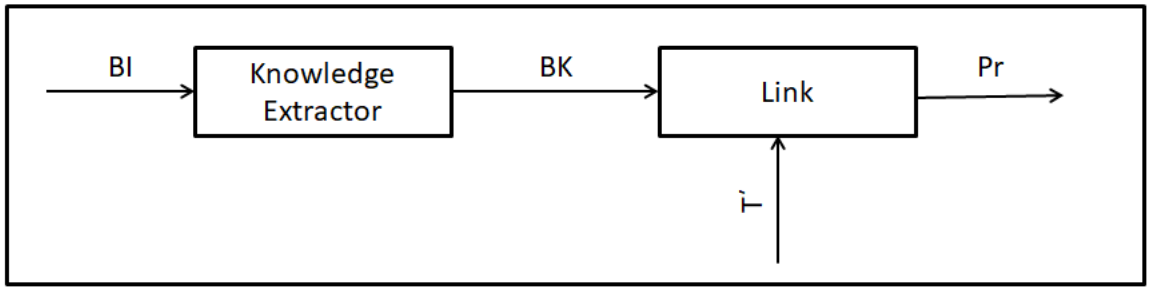

Figure 3.2: A Generic Adversarial Model for Background Knowledge

3.3.2 Adversarial Capabilities

Knowledge sets help adversary in manipulating privacy by linking the sensitive attributes with the quasiidentifiers of the individual(s). Considering a single adversary with access to all the knowledge sets will not be a practical approach as the adversary can have diverse manipulation capabilities. As a result, we divide the knowledge sets to differentiate adversary based on its manipulation capabilities. As a result, we categorize knowledge sets into three broader knowledge sets based on manipulation capabilities. They are as follows:

1. Personalized Knowledge (K_P) : It consists of minimalistic knowledge required in manipulating privacy. Personalized knowledge is a set that consists of Statistical, Individual and cohesive knowledge set.

2. Definite Knowledge (K_D) : It consists of definite (certain) knowledge required in manipulating privacy. Definite knowledge is a set that consists of

Statistical, Individual, Cohesive, Demographic, Correlational and semantic knowledge set.

3. Probabilistic Knowledge (K_{Pr}) : It consists of probabilistic knowledge required in manipulating privacy. Probabilistic knowledge is a set that consists of Statistical, Individual, Cohesive, Demographic, Correlational, Semantic, Probabilistic and Inferential knowledge set.

Certainly, $K_P \subset K_D \subset K_{Pr}$. Furthermore, only K_{Pr} considers probabilistic knowledge. We further define functions that can be accessed by the adversary. They are as follows:

1. $KE(BI) \rightarrow BK$: The function KE takes as an input Background Information BI and outputs Background Knowledge BK. Here, BK can be any one of the K_P, K_D, and K_{Pr} knowledge sets.

2. $LINK(T', BK) \rightarrow Pr$: The function $LINK$ takes as an input the anonymized table T' and Background Knowledge BK (any one of K_P, K_D, and K_{Pr}). The output of the *Link* function is Pr. i.e., $P(T'|BK)$ is a linking probability of the anonymized table T' in the presence of background knowledge BK.

3. $CHECK(Pr) \rightarrow s, f$: The function compares $P(T'|BK)$ with $P(T')$, if the difference is negligible, then privacy is not disclosed f; else, privacy is disclosed s.

The knowledge extractor block and link block in the adversarial model is function KE and function *Link*, respectively. The *Check* function checks whether the adversary is successful in disclosing privacy or not.

We define three types of adversaries based on access to the knowledge sets. Adversary accesses the above - defined functions, but the primary difference exists

in the BK background knowledge set. The adversaries are as follows:

1. *Weak Adversary* has access to the Personalized knowledge set K_P.
2. *Moderate Adversary* has access to the Definite knowledge set K_D.
3. *Strong Adversary* has access to the Probabilistic knowledge set K_{Pr}.

We also define the linking probability $P(T')$ of anonymized table T'. $P(T')$ is the probability of linking the quasiidentifiers of individual(s) with its sensitive attributes in an anonymized table T' that follows any given privacy principle. We now define privacy in an anonymized table T' in the context of background knowledge.

Definition 3.2. Privacy: Given an anonymized table T', an Adversary A and knowledge extractor KE, T' preserves privacy against adversary A if the difference of linking probability of the anonymized Table T' in presence of BK and without BK is negligible.

$$|P^A_{KE}(T'|BK) - P^A_{KE}(T')| \leq \Delta \tag{3.10}$$

Here, Δ is negligible. Note that anonymized table T' fulfils any given privacy principle. The interpretation of the above definition is that the presence of background knowledge does not affect the privacy of individual(s) in the anonymized table T'. This acts as a privacy guarantee against background knowledge in data publishing scenarios.

Privacy Models protects against privacy attacks due to Background Knowledge in the Privacy - Preserving Data Publishing Domain. The privacy model that can provide a strong privacy guarantee against background knowledge is competent to protect the privacy of the individual(s) against adversarial capabilities in the published data. We define a strong privacy model against background knowledge as follows:

Definition 3.3. Strong Privacy Model: Given a Privacy Model P, an anonymized table T'_P generated using P, an Adversary A and knowledge extractor KE, P is a strong privacy model if the anonymized table T'_P fulfils equation 3.10.

Here, the Privacy model P follows any given privacy principle and the Anonymized Table T'_P abides by that privacy principle. The adversary A has access to the adversarial capabilities as per Section 3.3. A privacy model is said to be strong if the anonymized table generated using the privacy model fulfils equation 3.10. On the other side, a privacy model is not said to be strong if the anonymized table generated using the privacy model does not fulfil equation 3.10. The privacy model that is not strong is prone to privacy attacks due to background knowledge. Here, Background knowledge considers all the knowledge variants discussed in Section 3.2.

3.3.3 Important Observations

We establish some important observations between the adversaries defined in section 3.3.2. The observations will implicitly help analyse the impact of the broader knowledge sets (i.e. K_P, K_D and K_{Pr}) on the anonymized table T'. Here, we assume that each knowledge set can eliminate at least a single record in the anonymized table T'. As a result, in general, the number of records in T' to be removed by the adversary is at least 8.

Theorem 3.1. *Given an anonymized table T' that fulfils a privacy principle Y, if T' preserves privacy against Moderate Adversary then it implicitly preserves privacy against Weak Adversary.*

Proof. Let T' be an anonymized table fulfilling privacy principle Y. Here, $T' = \{t_1, \ldots, t_n\}$ and Y is any existing privacy principle. Let A_w be a weak adversary and A_m be a moderate adversary. The Background Knowledge BK, for a moderate and a weak adversary, is K_D and K_P, respectively. Here, n is the total records

present in T', whereas k is the records removed from T' using background knowledge. Note that K_P and K_D are not probabilistic. The linking probability of any record t_i when K_P can link a record/individual (t_i) to a sensitive attribute is 1 else the linking probability is $\frac{1}{n-(k-1)}$. The linking probability of T' when K_P is present is as follows:

$$
\begin{aligned}
P_{KE}^{A_w}(T'|K_P) &= P_{KE}^{A_w}(t_1, t_2, \ldots, t_n|K_P) \\
&= P_{KE}^{A_w}(t_1|K_P) \cdot P_{KE}^{A_w}(t_2|K_P), \ldots, P_{KE}^{A_w}(t_n|K_P) \\
&= \prod_{t_i \in K_P} (1) \cdot \prod_{t_i \notin K_P} \left(\frac{1}{n-(k-1)} \right) \\
&= 1^k \cdot \left(\frac{1}{n-(k-1)} \right)^{n-(k-1)} \\
&= 1 \cdot \left(\frac{1}{n-(k-1)} \right)^{n-(k-1)} \\
&= \left(\frac{1}{n-(k-1)} \right)^{n-(k-1)}.
\end{aligned}
$$

The linking probability of any record t_i when K_D can link a record/individual (t_i) to a sensitive attribute is 1 else the linking probability is $\frac{1}{n-(k-1)}$. The linking probability of T' when K_D is present is as follows:

$$
\begin{aligned}
P_{KE}^{A_m}(T'|K_D) &= P_{KE}^{A_m}(t_1, t_2, \ldots, t_n|K_D) \\
&= P_{KE}^{A_m}(t_1|K_D) \cdot P_{KE}^{A_m}(t_2|K_D), \ldots, P_{KE}^{A_m}(t_n|K_D) \\
&= \prod_{t_i \in K_D} (1) \cdot \prod_{t_i \notin K_D} \left(\frac{1}{n-(k-1)} \right) \\
&= 1^k \cdot \left(\frac{1}{n-(k-1)} \right)^{n-(k-1)} \\
&= 1 \cdot \left(\frac{1}{n-(k-1)} \right)^{n-(k-1)} \\
&= \left(\frac{1}{n-(k-1)} \right)^{n-(k-1)}
\end{aligned}
$$

Here, $n(K_P) = 3$ and $n(K_D) = 6$. Therefore, the minimum number of records removed is 3 and 6 due to K_P and K_D, respectively i.e. $\left(\frac{1}{n-2}\right)^{n-2} < \left(\frac{1}{n-5}\right)^{n-5}$. As a result, $P_{KE}^{A_w}(T'|K_P) < P_{KE}^{A_m}(T'|K_D)$. Also, $K_P \subset K_D$. Therefore, for any T', if privacy is preserved against a moderate adversary implies that privacy is preserved against a weak adversary. ■

Theorem 3.2. *Given an anonymized table T' that fulfils a privacy principle Y, if T' preserves privacy against Strong Adversary then it implicitly preserves privacy against Moderate Adversary.*

Proof. Let T' be an anonymized table fulfilling a privacy principle Y. Here, $T' = \{t_1, \ldots, t_n\}$ and Y is any existing privacy principle. Let A_s be a strong adversary and A_m be a moderate adversary. The Background Knowledge BK, for a moderate and a strong adversary, is K_D and K_{Pr}, respectively. Here, n is the total records present in T', whereas k is the records removed from T' using background knowledge. Note that K_{Pr} is probabilistic whereas K_D is not. The linking probability of any record t_i when K_{Pr} can link a record/individual (t_i) to a sensitive attribute is 1 else the linking probability is $\frac{1}{n-(k-1)}$. The linking probability of T' when K_{Pr} is

present is as follows:

$$\begin{aligned} P_{KE}^{A_s}(T'|K_{Pr}) &= P_{KE}^{A_s}(t_1, t_2, \ldots, t_n|K_{Pr}) \\ &= P_{KE}^{A_s}(t_1|K_{Pr}) \cdot P_{KE}^{A_s}(t_2|K_{Pr}), \ldots, P_{KE}^{A_s}(t_n|K_{Pr}) \\ &= \prod_{t_i \in K_{Pr}} (p) \cdot \prod_{t_i \notin K_{Pr}} \left(\frac{1}{n-(k-1)} \right) \\ &= p^k \cdot \left(\frac{1}{n-(k-1)} \right)^{n-(k-1)} \end{aligned}$$

Here, $n(K_{Pr}) = 8$ and $n(K_D) = 6$. Therefore, the minimum number of records removed is 8 and 6 due to K_{Pr} and K_D, respectively i.e. $\left(\frac{1}{n-5}\right)^{n-5} < \left(\frac{1}{n-7}\right)^{n-7}$. As a result, $P_{KE}^{A_m}(T'|K_{Pr}) < P_{KE}^{A_s}(T'|K_D)$. Also, $K_D \subset K_{Pr}$. Therefore, for any T', if privacy is preserved against a strong adversary implies that privacy is preserved against a moderate adversary. ■

More are the knowledge sets; more the adversary has the power to manipulate the privacy. Based on theorem 1 and theorem 2, we observe the following:
Strong Adversary $\Rightarrow$ *Moderate Adversary* $\Rightarrow$ *Weak Adversary*

3.4 Weakness in Privacy Model against Background Knowledge

This section analyses the privacy models named k - anonymity, l - diversity and t - closeness against the proposed adversarial model.

Theorem 3.3. *The linking probability of a k - anonymous table T' when BK is present is non - negligible against a moderate adversary.*

Proof. Let T' be a k - anonymous [1] table. Here, $T' = \{t_1, \ldots, t_n\}$ and each record is synonymous to an individual. T' consists of j quasiidentifiers and a sensitive attribute. A table T' is k - anonymous if each and every partition in the table T' is k - anonymous. We assume that linking the external table with the anonymized table can shrink k - anonymous table to a k - anonymous partition. Therefore, the linking probability of a partition when background knowledge is not present is $(\frac{1}{k})^k$. The moderate adversary accesses the functions as defined in section 3.3.2 as follows:

1. $KE(BI) \rightarrow K_D$: This step extracts knowledge as per section 3.2.2 from the Background Information (BI). As, knowledge can be obtained from external table and anonymized table T' apart from individual knowledge. Therefore, $BK \neq \phi$.

2. $LINK(T', K_D) \rightarrow PR$: The input to the $LINK$ function is anonymized table T' and background knowledge K_D. Here, r is the number of records removed from k - anonymous partition due to BK. This step calculates the linking probability when the background knowledge BK is present.

$$
\begin{aligned}
P_{KE}^{A_m}(T'|K_D) &= P_{KE}^{A_m}(t_1, t_2, \ldots, t_k|K_D) \\
&= P_{KE}^{A_m}(t_1|K_D) \cdot P_{KE}^{A_m}(t_2|K_D), \ldots, P_{KE}^{A_m}(t_k|K_D) \\
&= \prod_{t_i \in K_D} (1) \cdot \prod_{t_i \notin K_D} \left(\frac{1}{k-r}\right) \\
&= 1^r \cdot \left(\frac{1}{k-r}\right)^{k-r} \\
&= 1 \cdot \left(\frac{1}{k-r}\right)^{k-r} \\
&= \left(\frac{1}{k-r}\right)^{k-r}
\end{aligned}
$$

3. $CHECK(PR) \rightarrow s$: This step compares the linking probability as follows:

$$\begin{aligned}
\left| P_{KE}^{A_m}(T'|K_D) - P_{KE}^{A_m}(T') \right| &= \left| \left(\frac{1}{(k-r)} \right)^{k-r} - \left(\frac{1}{k} \right)^{k} \right| \\
&= \left| \left(\frac{1^{k-r}}{(k-r)^{k-r}} \right) - \left(\frac{1^{k}}{k^{k}} \right) \right| \\
&= \left| \left(\frac{1}{(k-r)^{k-r}} \right) - \left(\frac{1}{k^{k}} \right) \right| \\
&= \left| \frac{(k-r)^{r}}{(k-r)^{k}} - \frac{1}{k^{k}} \right| \\
&= \left| \frac{(k-r)^{r}}{(k-r)^{k}} \right| \left(\because \frac{1}{k^{k}} << \frac{(k-r)^{r}}{(k-r)^{k}} \right)
\end{aligned}$$

A function $f(k)$ is non - negligible [28][29] if $\exists c \in N$ such that $\forall k_0 \in N$, there is a $k \geq k_0$ such that $f(k) \geq k^{-c}$. Here, $f(k) = \frac{(k-r)^r}{(k-r)^k}$. We simplify $f(k)$ in terms of k as follows:

$$\begin{aligned}
f(k) &= \left(\frac{(k-r)^{r}}{(k-r)^{k}} \right) \\
&= \left(\frac{(k-r)^{r}}{\left((k-r)^{log_{(k-r)k}} \right)^{\frac{k}{log_{(k-r)^k}}}} \right) \\
&= \left(\frac{(k-r)^{r}}{k^{\frac{k}{log_{(k-r)k}}}} \right)
\end{aligned}$$

There exists $c = \frac{k}{log_{(k-r)}k}$ such that $k \geq 3$ ($k_0 = 3$) and $1 \leq r \leq k-1$, $\frac{(k-r)^r}{k^{\frac{k}{log_{(k-r)^k}}}} \geq k^{-\frac{k}{log_{(k-r)^k}}}$. As a result, $f(k) \geq k^{-c}$. Therefore, k - anonymity doesn't preserve

privacy against moderate adversary as $|P_{KE}^{A_m}(T'|K_D) - P_{KE}^{A_m}(T')|$ is non-negligible.
■

Theorem 3.4. *The linking probability of l - diverse table T' when BK is present is non - negligible against a moderate adversary.*

Proof. Let T' be a l - diverse table[2]. Here, $T' = \{t_1, \ldots, t_n\}$ and each record is synonymous to an individual. T' consists of j quasiidentifiers and a sensitive attribute. A table T' is l - diverse if each and every partition in table T' is l - diverse. We assume that basic identification knowledge can shrink the l - diverse table into a l - diverse partition. The linking probability of partition when background knowledge is not present is $(\frac{1}{l})^l$. The moderate adversary accesses the functions as defined in section 3.3.2 as follows:

1. $KE(BI) \rightarrow K_D$: This step extracts knowledge as per section 3.2.2 from Background Information (BI). As knowledge can still be obtained apart from the basic identification information and anonymized table T'. Therefore $BK \neq \phi$.

2. $LINK(T', K_D) \rightarrow PR$: The input to the $LINK$ function is anonymized table T' and background knowledge K_D. Here, r is the number of sensitive attribute values removed from l - diverse partition due to BK. This step calculates the linking probability when the background knowledge BK is present.

$$
\begin{aligned}
P_{KE}^{A_m}(T'|K_D) &= P_{KE}^{A_m}(t_1, t_2, \ldots, t_l|K_D) \\
&= P_{KE}^{A_m}(t_1|K_D) \cdot P_{KE}^{A_m}(t_2|K_D), \ldots, P_{KE}^{A_m}(t_l|K_D) \\
&= \prod_{t_i \in K_D} (1) \cdot \prod_{t_i \notin K_D} \left(\frac{1}{l-r}\right) \\
&= 1^r \cdot \left(\frac{1}{l-r}\right)^{l-r} \\
&= 1 \cdot \left(\frac{1}{l-r}\right)^{l-r} \\
&= \left(\frac{1}{l-r}\right)^{l-r}
\end{aligned}
$$

3. $CHECK(PR) \rightarrow s$: This step compares the linking probability as follows:

$$
\begin{aligned}
\left| P_{KE}^{A_m}(T'|K_D) - P_{KE}^{A_m}(T') \right| &= \left| \left(\frac{1}{l-r}\right)^{l-r} - \left(\frac{1}{l}\right)^{l} \right| \\
&= \left| \left(\frac{1^{l-r}}{(l-r)^{l-r}}\right) - \left(\frac{1^l}{l^l}\right) \right| \\
&= \left| \left(\frac{1}{(l-r)^{l-r}}\right) - \left(\frac{1}{l^l}\right) \right| \\
&= \left| \frac{(l-r)^r}{(l-r)^l} - \frac{1}{l^l} \right| \\
&= \left| \frac{(l-r)^r}{(l-r)^l} \right| \left(\because \frac{1}{l^l} << \frac{(l-r)^r}{(l-r)^l} \right)
\end{aligned}
$$

A function $f(l)$ is non - negligible [28][29] if $\exists c \in N$ such that $\forall l_0 \in N$, there is a $l \geq l_0$ such that $f(l) \geq l^{-c}$. Here, $f(l) = \frac{(l-r)^r}{(l-r)^l}$. We simplify $f(l)$ in terms of l as follows:

$$
\begin{aligned}
f(l) &= \left(\frac{(l-r)^r}{(l-r)^l}\right) \\
&= \left(\frac{(l-r)^r}{\left((l-r)^{log_{(l-r)}l}\right)^{\frac{l}{log_{(l-r)}l}}}\right) \\
&= \left(\frac{(l-r)^r}{l^{\frac{l}{log_{(l-r)}l}}}\right)
\end{aligned}
$$

There exists $c = \frac{l}{log_{(l-r)}l}$ such that $l \geq 3$ ($l_0 = 3$) and $1 \leq r \leq l-1$, $\frac{(l-r)^r}{l^{\frac{l}{log_{(l-r)}l}}} \geq l^{-\frac{l}{log_{(l-r)}l}}$. As a result, $f(l) \geq l^{-c}$. Therefore, l - diversity doesn't preserve privacy against moderate adversary as $|P_{KE}^{A_m}(T'|K_D) - P_{KE}^{A_m}(T')|$ is non - negligible. ■

Theorem 3.5. *The linking probability of t - closeness anonymized table T' when BK is present is non - negligible against a moderate adversary.*

Proof. t - closeness [3] addresses attribute disclosure, but doesn't address identity disclosure as a result can not be a stand - alone property applied on data table. So, k-anonymity and t - closeness [3] can be applied together for anonymization of data table. As a result, proof is same as k - anonymity. ■

3.5 Conclusion

Background Knowledge plays a vital role in the evolution of privacy models. Adversary using background knowledge can disclose the privacy of the individual(s). In this chapter, we propose an adversarial model for Background knowledge. We analyze the privacy models like k - anonymity, l - diversity and t - closeness against the proposed adversarial model. We believe that this study will

help in modelling a strong privacy model against background knowledge that preserves the privacy of individuals and data while publishing data in public.

CHAPTER 4

Privacy Model against Background Knowledge

Privacy attacks due to background knowledge prominently dominate the privacy - preserving data publishing domain. The adversary uses the background knowledge to compromise the privacy of the individual(s). There arises a need to protect individual's data with stringent privacy solutions against background knowledge. Inevitably, background knowledge is a serious threat to data privacy due to its diversity and ease of availability.

We have observed two interesting takeaways related to background knowledge in Chapter 3. They are as follows:

Observation 1: Each privacy model addresses background knowledge partly.

Argument: Each privacy model has its own background knowledge assumptions. The background knowledge assumptions are not comprehensive as one privacy model targets a set of background knowledge whereas becomes victim to other set of background knowledge.

Observation 2: Each privacy model assumes adversary to have limited capabilities.

Argument: In the current scenario, knowledge is freely available in the public domain. Moreover, adversary has access to advanced manipulation techniques for

privacy disclosure purpose. As a result, assuming limited adversarial capabilities becomes a naive assumption.

The above observations show the need for a strong privacy model that considers comprehensive background knowledge against a strong adversary. In this chapter, we show the implications of the background knowledge in terms of privacy disclosure on privacy - preserved data. We provide a broader perspective to semantic knowledge and prove that it has wider implications concerning privacy disclosure. We propose a privacy model that not only provides a solution towards semantic knowledge but also preserve the privacy of data against the comprehensive background knowledge. We analyze the proposed privacy model and show its practicality with the experimental results.

4.1 Introduction

In the modern digital age, data has become an essential commodity for competitive advantages in doing business. Tremendous volumes of data generated from various sources in every moment, which not only makes useful knowledge available to our society, but also allows organizations to infer useful results for business analytics, health symptoms prediction, weather forecasting, location - based services and so on. Moreover, the observations made from these data help in moulding government policies, improving the user experience, incorporating new market dimensions, exploring research trends and handling a health crisis. However, the online collected data, in recent times, pose a crucial concern is data privacy. There are many instances [134], [137], [138] of data breaches that show significant security issue to tackle with while availing various services from these collected data.

Although some techniques [32], [135], [136] such as data obfuscation, data suppression is useful to preserve data privacy, the published data found vulnerable against the comprehensive background knowledge, where the adversary can

compromise the sensitive attributes of the individual present in the published table. The background knowledge can be obtained from basic identification information, external table, demographic information, published observations and research conclusions and information from social connections (As discussed in Chapter 3). To overcome the privacy disclosure threat, different privacy models have been used in applications that address [32] privacy issues against background knowledge. For example, k - anonymity [1] model protects against the external table (e.g. voters list), l - diversity model [2] protects against personal identification knowledge and partly from demographic information, t - closeness [3] protects against the knowledge of global distribution of the sensitive attributes.

In the next section, we discuss the implication of background knowledge on the published privacy - preserved data using an example.

4.2 Implication of Background Knowledge on Published Data

It is assumed that the background knowledge is comprehensive and vast in nature due to the free availability of the primary identification information of individuals on the Internet. We further assume that the background knowledge can accommodate different variants of knowledge, namely, statistical knowledge, individual knowledge, demographic knowledge, correlational knowledge, probabilistic knowledge, inferential knowledge, semantic knowledge and social connections. We apply different types of knowledge on the privacy preserved published data table, and then, we analyze privacy in terms of diversity of sensitive attributes.

Scenario: Suppose a hospital publishes Table 4.1 for research purpose. The adversary has somehow got access to Table 4.1. The adversary knows the presence of individual "Y" in the Table 4.1. The goal of the adversary is to compromise the privacy of the individual "Y" by knowing its sensitive attribute value.

Prerequisites: The Table T' satisfies the privacy principles of k - anonymity, l - diversity and t - closeness. Table T' protects against attacks like linking attack, homogeneity attack, background knowledge attack (as specified in l - diversity [2]), similarity attack and skewness attack. Table T' consists of 12 records. The initial assumption of adversary related to the sensitive attribute values is equivalent to the number of records. As a result, the probability of each sensitive attribute value is $\frac{1}{12}$ as each sensitive attribute value has uniformly linked to an individual "Y".

ZipCode	Age	Gender	Disease
1503*	4*	*	Stomach Cancer
1503*	4*	*	Breast Cancer
1503*	4*	*	Lymphoma
1503*	4*	*	Melanoma
1503*	4*	*	Heart Disease
1503*	4*	*	Ebola
1503*	4*	*	Prostate Cancer
1503*	4*	*	Diarrhoea
1504*	7*	*	Dementia
1504*	7*	*	Kidney Stone
1504*	7*	*	Heart Disease
1504*	7*	*	Pneumonia

Table 4.1: An Example of Anonymized Table T'

Analysis: The anonymized table can reveal statistical information about data distribution. Statistical knowledge though cannot eliminate the records but do change the probabilities. For example, the adversary applies *statistical knowledge* onto the table T'. Suppose that the statistical knowledge is Count(Disease = "Heart Disease") = 2. Therefore, the probability of heart disease is $\frac{2}{12}$. The probability of sensitive attribute value "Heart Disease" has been increased from 0.0833 to 0.1667. The probability of the remaining sensitive attribute values remains unchanged. Therefore, the statistical knowledge, without any specific individual knowledge, has been able to increase the probability. Person specific knowledge with the adversary can also lead to privacy disclosure. For example, the adversary applies *individual knowledge* (adversary specific knowledge) on the published

table T'. Suppose, the individual knowledge is "Individual Y, a Gender = Female lives in Zipcode = 15032 and country = India". Typically, this will eliminate the partition 3. Therefore, the probability of the sensitive attribute values after applying individual knowledge is $\frac{1}{8}$. Therefore, the increase in the probability of sensitive attribute values is from 0.0833 to 0.125. Individual knowledge has narrowed down the sensitive attribute values from 12 to 8.

The adversary can exploit the demographic information (specifically based on location) about the individual, which can result in privacy disclosure. For example, the adversary applies *demographic knowledge* on the published table T'. Suppose, demographic knowledge is "Ebola not present in India". This knowledge will help in eliminating the sensitive attribute value "Ebola" from partition 2. The probability of sensitive attribute values after applying demographic knowledge has changed from 0.125 to 0.1428. Demographic knowledge removes a sensitive attribute value and narrows down the number from 8 to 7. The relationship between quasiidentifier values and sensitive attribute values can be exploited by the adversary leading to privacy disclosure. For example, the adversary now applies *correlational knowledge* onto the published table T'. Suppose, the correlational knowledge is "Females do not have Prostate Cancer". This knowledge will help in eliminating the sensitive attribute value "Prostate Cancer" from partition 2. The probability of sensitive attribute values after applying correlational knowledge has changed from 0.1428 to 0.1667. Correlational knowledge removes a sensitive attribute value and brings down the diversity of sensitive attribute values from 7 to 6.

Semantically related sensitive attribute values can lead to privacy disclosure. For example, the adversary now applies semantic knowledge onto the published table T'. Suppose, the *semantic knowledge* is "Stomach Cancer, Breast Cancer, Lymphoma and Melanoma are semantically equivalent to Cancer". This knowledge will help in summarizing the sensitive attribute values in partition 1 to a single sensitive attribute value i.e. Cancer. The probability of the sensitive attribute

values after applying semantic knowledge has changed from 0.1667 to 0.3333. Semantic knowledge has summarized the sensitive attribute values and narrowed the diversity from 6 to 3.

The demographic and correlational knowledge can be probabilistic in nature. Indeed, this can change the probabilities of sensitive attribute values and lead to privacy disclosure. For example, the adversary now applies *probabilistic knowledge* onto the published table T'. Suppose, the probabilistic knowledge is "Heart Disease has a high probability(0.45) of occurrence in India". This knowledge will not eliminate the records but will change the scenario of uniform probability distribution amongst sensitive attribute values. The probability of sensitive attribute values, namely Heart Disease, Cancer and Diarrhoea, after applying probabilistic knowledge is 0.45, 0.275 and 0.275, respectively. This knowledge has not eliminated any records but definitely differentiated the chances of linking sensitive attribute value with individual "Y" in terms of probability.

Inferences between the domain of sensitive attribute values (present or not present in an anonymized table) can lead to privacy disclosure. For example, the adversary now applies *inferential knowledge* onto the published table T'. Suppose that the inferential knowledge is "Hypertension leads to a heart disease with high probability (0.55)". This inference can be helpful if an adversary has specific individual information. Further, the adversary has the *individual knowledge* "Individual Y suffers from hypertension". Now, the inference of hypertension and heart disease can lead to privacy disclosure. The inference has increased the probability of heart disease. The probability of sensitive attribute values namely Heart Disease, Cancer and Diarrhoea after applying inferential knowledge is 0.6667, 0.1667 and 0.1667, respectively. This knowledge has definitely increased the chances of heart disease being the sensitive attribute value considerably. Social connections namely information about the family, social links like knowledge about the common groups, friends etc. can lead to privacy disclosure. For example, the adversary now applies *cohesive knowledge* onto the published table T'. Suppose, the

cohesive knowledge is "Individual Y's father suffers from Heart Disease". This knowledge will not help eliminate the sensitive attribute values but when combined with any previous knowledge sets has a potential for privacy disclosure. If *inferential knowledge* is applied onto cohesive knowledge and is depicted as "If Father has heart disease then children have a high probability(0.6) to get heart disease (by heredity)". It will help heart disease to link to the individual "Y" successfully as the probability of heart disease is $0.85712 \approx 1$.

The background knowledge leads to privacy disclosure as the adversary has discovered individual Y's sensitive attribute value (i.e. heart disease). The probability of heart disease has shown a significant increase which has lead to privacy disclosure. The privacy attack due to background knowledge occurred not due to less diversity of the sensitive attribute values but due to more availability of background knowledge. Moreover, we discuss that semantic knowledge has a broader interpretation than in [3] in the next section.

4.3 Semantic Knowledge: A Broader Perspective

Semantic knowledge explores the semantic similarity amongst the published data. Semantic similarity pinpoints to ontological proximity between two terms/ concepts/ words/ sentences. In this work, we focus on sensitive attributes. Here, we assume ontological proximity to be the semantic distance between two sensitive attributes in a table. For example, if we consider the disease domain, then prostate cancer and breast cancer both are similar as both are related to diseases of cellular proliferation [31]. If these set of diseases arrive in a single partition, then with the semantic domain knowledge, the adversary can learn that individual has cancer. Though the diseases are syntactically dissimilar but have a vague difference in terms of semantic meaning. As a result, less is the proximity; more related are the sensitive attribute values. In other words, more related are the sensitive attribute values in a partition; higher is the chances of privacy disclosure. Moreover, nowadays, for trend analysis, the data analysts look for specific keywords in the

search/posts to generate the ranking of the trends. The posts can be partitioned by using any of the popular privacy models. If the partition consists of syntactically dissimilar but semantically similar posts, then with semantic knowledge, the adversary can know the contents of the individual's post, which will obstruct the privacy. Furthermore, sensitive attribute values need not be a single word; there arises a need to protect the data from background knowledge as well as semantic knowledge. Semantic knowledge with domain information can shrink the size of the equivalence class considerably. Let us illustrate this with an example.

ZipCode	Age	Gender	Disease
1503*	2*	*	Prostate Cancer
1503*	2*	*	Breast Cancer
1503*	2*	*	Lymphoma
1503*	2*	*	Melanoma

Table 4.2: 4 - anonymous, 4 - diverse partition

Table 4.2 shows an equivalence class that is 4 - anonymous and 4 - diverse. As the data has more records of cancer patients so as per the statistical distance, the following equivalence class/partition is generated. However, if we know the domain of the sensitive attribute (here it is a disease), then without any apprehensions, we can conclude that the individual has cancer. Nevertheless, the simple semantic classification will not lead to privacy disclosure. Here, if we consider simple classification, then all the four diseases are not similar, as they occur in different parts of the human body. But if we consider semantically, then all the four diseases are similar as they show similar behaviour of abnormal cell growth. As a result, semantic knowledge has the potential to disclose privacy without actually eliminating the three sensitive attribute values. Therefore, semantic knowledge has broader implications in terms of privacy disclosure.

As a result, a more realistic approach for privacy model is required that can not only differentiate the sensitive attribute values based on their semantic meanings but also consider other variants of knowledge. In this chapter, we present a privacy model that preserves the privacy of data against background knowledge.

4.4 Preliminaries

Let table $T = \{r_1, r_2, \ldots, r_n\}$ consists of n distinct records. Each record r_i is distinct and points to an individual. Table T consists of t quasiidentifiers and a sensitive attribute SA. The schema is $T(q_1, q_2, \ldots, q_t, SA)$. Sensitive attribute SA is a set consisting of j distinct sensitive attribute values represented as $SA = \{SA_1, SA_2, \ldots, SA_j\}$. The record r_i^{th} of table T would be of the form $\{r_i[q_1], r_i[q_2], \ldots, r_i[q_t], r_i[SA_i]\}$ (SA_i is a particular sensitive attribute value). Domain of quasiidentifiers and sensitive attributes are represented as D_{QI} and D_{SA}, respectively.

4.4.1 Intrinsic Notions

The basic definitions and concepts that are required in formalizing the privacy model is as follows:

Definition 4.1. Partition: For a table T, a partition p_i where $p_i \subseteq T$, consists of r ($r < n$) records where each record r_x^p is distinct and associated to an individual. It is represented as $p_i = \bigcup_{x=1}^{r} r_x^p$. Also, $T = \bigcup_{i=1}^{k} p_i$.

We specifically focus on finding a semantic dissimilar partition. We would first define semantic distance, and then, it will be followed by semantic dissimilar partition. To calculate the semantic distance, we use ontology [31] created by domain of the sensitive attributes.

Definition 4.2. Semantic Distance: For any given $SA_i, SA_j \in SA$ and domain ontology O_{SA} of SA, the semantic distance between two sensitive attributes SA_i and SA_j is defined as

$$Sem_D \underset{i \neq j}{(} SA_i, SA_j) = E_{O_{SA}}(SA_i, SA_j) = n \tag{4.1}$$

Here, $E_{O_{SA}}$ is the edges on the ontology O_{SA}, SA_i and SA_j are two sensitive attribute values of SA, n is the number of edges. The equation signifies the number of edges between two sensitive attributes in the ontology O_{SA}. Semantic Distance is commutative in nature, that is, $Sem_D(SA_i, SA_j) = Sem_D(SA_j, SA_i)$ gives the same output.

Definition 4.3. Semantically Dissimilar Partition: For any given partition p where $p \in T$, the partition p consists of j_p distinct sensitive attribute values such that $SA^p = \{SA_1^p, \ldots, SA_{j_p}^p\}$, then p is said to be semantically dissimilar partition if it satisfies the below condition

$$\underbrace{Sem_D(SA_i^p, SA_j^p)}_{\substack{\forall i,j \in p \\ i \in \{1,\ldots,j_p-1\} \\ j \in \{2,\ldots,j_p\} \\ i \neq j}} \geq \theta \tag{4.2}$$

The above condition implies that semantic distance between all the sensitive attributes in a partition p must be at least θ. For example, consider a partition p which consists of 3 diseases {CAD, Gastritis, Viral Infection}, which are sensitive attribute values. The semantic threshold $\theta \geq 6$ is given. Then, based on the ontology [31], the distance between the pairs is calculated as i) CAD - Gastritis = 7 ii) CAD - Viral Infection = 6 iii) Gastritis - Viral Infection = 7. The above pairs fulfil the semantic threshold criteria. Therefore, the partition p is semantically dissimilar partition.

Definition 4.4. $[lb, ub]$ **bound partition:** For any given partition p where $p \in T$, a partition is said to be $[lb, ub]$ bound if it satisfies the below condition

$$lb \leq j_p \leq ub \tag{4.3}$$

Here, j_p is the number of distinct sensitive attribute values in a partition p, lb is the lower bound of distinct number of sensitive attribute values where $lb \geq 3$ and ub is the upper bound of distinct number of sensitive attribute values in a partition.

We calculate the upper bound ub for our model as follows: For a given table T, take a sample S where $S \subset T$ and create y semantically dissimilar partitions i.e. $S = \bigcup_{i=1}^{y} p_i$ where $(y < k)$, upper bound ub is calculated as follows:

$$ub = \frac{\sum_{i=1}^{y} (n(SA_i^p))}{y} \tag{4.4}$$

The above equation calculates an average of the partitions created by taking a sample from the microdata table T. The sample S consists of j distinct sensitive attribute values. This definition helps in creating an upper bound for the partition not more than ub number of sensitive attribute values should reside in a partition. Instead of l - diversity [2], we use this mechanism as l - diversity becomes rigid at times, as the partition requires at least l diverse attribute values. As we use semantic dissimilar partitions, we can not always place exactly l semantically dissimilar sensitive attribute values into a partition. Furthermore, we decide the upper bound dynamically after analyzing the dataset as each dataset has different distributions.

Definition 4.5. α **clustered partition:** For a given $[lb, ub]$ semantically dissimilar partition p with j_p distinct sensitive attribute values in table T, the partition p is said α clustered partition if it satisfies the following two conditions

1. $n(SA_i^p)_{i=1,\ldots,j_p} \leq \alpha$
2. $E(p) \geq log(j_p)$

We calculate α dynamically as follows: For a given Sample S of table T and sensitive attributes $SA = SA_i^S = \{SA_1^S, \ldots, SA_j^S\}$, α is calculated as:

$$\alpha = \frac{\sum_{i=1}^{j} n(SA_i^S)}{j} \tag{4.5}$$

The above equation calculates the upper bound of α which signifies the number of redundant sensitive attribute values accommodated in a single partition. For example, if the partition p = {CAD, Gastritis, Viral Infection} and α = 5. Then, at most 5 redundant records consisting of CAD as a sensitive attribute value is accommodated. The α has been selected dynamically as redundancy varies for different data sets. Lower bound of α is not selected due to the negative impact in the usefulness of data as data distributions are not always uniform. At the same time, our goal is that the adversary could not get more information than intended and hence the entropy of a particular partition p after adding at most α redundant records should be at least $log(j_p)$. For example, in the above partition p, the entropy should be at least $log(3)$ as there are 3 sensitive attribute values.

Definition 4.6. sp **- concealed partition:** The partition p' is said to be sp - concealed partition if it satisfies the following two conditions:

1. Each partition p' consists of spurious records $r_{sp}^{p'}$ such that $p' = \{(r_1^{p'}, r_2^{p'}, \ldots, r_r^{p'}), r_{sp}^{p'}\}$.
2. $sp_{SA}^{p'} \notin BK_A^{p'}$.

Here, a set of spurious records $r_{sp}^{p'} = \{r_{sp_1}^{p'}, r_{sp_2}^{p'}, \ldots, r_{sp_u}^{p'}\}$ where $u \geq 1$ are added into a partition p'. The structure of the i^{th} spurious record $r_{sp_i}^{p'}$ in partition p' is $(r_{sp_i}^{p'}[q_1], r_{sp_i}^{p'}[q_2], \ldots, r_{sp_i}^{p'}[q_t], r_{sp_i}^{p'}[sp_{SA_i}])$ (Here sp_{SA_i} is a particular i^{th} spurious sensitive attribute value). $sp_{SA}^{p'}$ is a set of distinct spurious sensitive attribute values for a particular partition p'. $BK_A^{p'}$ is the set of probable sensitive attribute values that are likely to be selected by the adversary obtained using background knowledge (Refer definition 3.1 for background knowledge). In other words, the

partition should not contain the spurious sensitive attribute values present in the background knowledge.

4.5 $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private: Privacy Model

Privacy disclosure is successfully linking the sensitive attribute to an individual with high conviction (probability). In literature [3], [33] privacy disclosure is related to three types of disclosure, i.e., *membership disclosure*, *identity disclosure* and *attribute disclosure*. In *membership disclosure*, the adversary learns that an individual is present in the published dataset. In *identity disclosure*, the adversary can link a record in the dataset to an individual. In *attribute disclosure*, the adversary can get more information about the attributes rather than the individual. Analyzing the three types of disclosure, in the current scenario, *membership disclosure* is the least significant. The reason being the basic information of the individuals available in the public domain. While *identity disclosure* and *attribute disclosure* are the two disclosures that the privacy models target. *Identity disclosure* may lead to *attribute disclosure* in case of lack of diversity in sensitive attribute. k - anonymity provides a solution to identity disclosure but doesn't consider attribute disclosure. Identity disclosure will not link the individual to sensitive attributes when sensitive attributes and quasiidentifiers are not correlated (bifurcated into two tables). Moreover, getting identity information like quasiidentifiers, i.e., age, zipcode, gender etc. is comparatively easy due to availability in the public domain. In current times, *attribute disclosure too can lead to identity disclosure* . We explain using the following scenario: Consider an adversary which successfully narrows down to a partition by using the basic information. If a partition consists of distinct sensitive attribute values, the adversary can eliminate the sensitive attribute values with the enormous background knowledge and lead to identity disclosure [153]. Therefore, attribute disclosure is a substantial threat. The privacy models like l - diversity, t - closeness have provided solutions to the attribute disclosure but doesn't consider identity disclosure. We now explain the basis of the proposed privacy model.

Privacy of dataset is inversely proportional to the number of records eliminated from the published table. More the number of records eliminated less is the privacy preserved. The adversary initially observes the table T' consisting of j sensitive attribute values as the probability $\frac{1}{j}$. With the quasiidentifiers being known to the adversary, the number of records gets drastically shrunk to a partition, consisting of a set of records with j_p sensitive attribute values. A partition can disclose privacy by following any of the approaches:

1. By eliminating the sensitive attribute values gradually with the help of background knowledge.
2. By summarizing the sensitive attribute values to a broader context.
3. By analyzing the uneven data distribution of sensitive attribute values.

The above approaches have been discussed in the literature independently or partly, but not collectively. Therefore, there is a need to come up with a privacy model that takes care of all these scenarios and protects against privacy disclosure (both attribute and identity disclosure) of the individual(s). We define the proposed $(\theta, [lb, ub]^{+sp}, \alpha)$ - *Private* as follows:

Definition 4.7. $(\theta, [lb, ub]^{+sp}, \alpha)$ **- Private:** A table T' is $(\theta, [lb, ub]^{+sp}, \alpha)$ Private if each partition p'_i in Table T', where $T' = \bigcup_{i=1}^{k} p'_i$ accomplishes the following conditions C:

1. Each partition p'_i is a semantically dissimilar partition.
2. Each partition p'_i is a $[lb, ub]$ bound and an α clustered partition.
3. Each partition p'_i is a sp - concealed partition.

4. In each partition p'_i, the probability of linking of quasiidentifiers of each record $r_i^p \in p'_i$ with each j_p sensitive attribute values is between $\left[\frac{1}{lb+sp_{SA}^{p'_i}}, \frac{1}{ub+sp_{SA}^{p'_i}}\right]$.

where $sp_{SA}^{p'_i}$ is the set of spurious sensitive attribute values in partition p'_i. If p_i satisfies condition C, then p_i is $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private. Table T' is $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private if and only if $\forall p_i \in T', \{p_1, \ldots, p_k\}$ satisfies condition C.

The proposed privacy model not only protects from variants of knowledge but also from data distribution. Moreover, it also ensures that the probability of linking the individual with the sensitive attribute in a partition is between $\left[\frac{1}{lb+sp_{SA}^{p'_i}}, \frac{1}{ub+sp_{SA}^{p'_i}}\right]$. As a result, the proposed privacy model is more realistic and robust due to its background knowledge assumption.

4.6 The Algorithm

This section provides the details of $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private model.

The algorithm is divided into 4 phases - 1) Parameter Selection; 2) Partitioning; 3) Spurious Record Generation; and 4) Anatomization. The Parameter Selection phase selects tuning parameters θ, α and $[lb, ub]$ and provides flexibility to select from the dataset dynamically. The Partitioning phase depends on the selected tuning parameters. The Spurious Record Generation phase adds spurious records to protect against background knowledge BK. The Anatomization phase bifurcates the given table into two tables (i) quasiidentifier table (QI_T), and (ii) sensitive attribute table (SA_T). The phases are explained as follows.

4.6.1 Phase 1: Parameter Selection

This phase selects the tuning parameters from the given microdata Table (dataset) dynamically. Rather than giving a fixed value without actually analyzing the data can, at times, deal with improper partitioning.

1. θ : It is a semantic threshold. This parameter helps in creating the partitions that are semantically different; in other words, the sensitive attributes in a partition are not semantically similar. We explain it with an example. If partition p consists of sensitive attribute values {Viral Infection, Flu, Influenza} where all these diseases point to respiratory ailments. The adversary can summarize the sensitive values leading to privacy disclosure.

2. $[lb, ub]$: The partitions must have diversity in terms of the number of sensitive attribute values. Therefore, we are using upper bound and lower bound on the sensitive attributes in a partition. Here, we have presumed $lb = 3$, as for lesser the number of sensitive attributes more are the chances of privacy disclosure. ub is the upper bound on the number of sensitive attributes in a partition.

3. α: α leads to uniformity in the number of redundant sensitive attribute values. α acts as an upper bound for redundancy.

Algorithm 1 Phase 1 - Parameter Selection

INPUT: A Micro data table T, Ontology O.

OUTPUT: Tuning parameters θ, ub and α.

1: Take Sample $S \subset T$.

2: **for each** record r_i^S in S **do**

3: Calculate SA^S and $n(SA^S)$.

4: **end for**

5: **for each** i in (SA_i^S) **do**

6: Calculate $Sem_D(SA_i^S, SA_{i'}^S)$ such that $i \neq i'$ & $i = 1, \ldots, j-1$,$i' = 2, \ldots, j$.

7: **end for**State Calculate θ

$$\theta = \frac{\sum_{i=1,i'=2,i\neq i'}^{j-1,j} Sem_D(SA_i^S, SA_{i'}^S)}{\sum_{j=2}^{n}(j-1)}$$

8: **for each** distinct sensitive attribute values SA^S in S **do**

9: Create partitions $\{p_1, \ldots, p_y\}$

10: **end for**

11: **for each** partition p_i in S **do**

$$ub = \frac{\sum_{i=1}^{y} n(SA_i^p)}{y}$$

12: **end for**

13: **for each** distinct sensitive attribute values SA^S in Sample S **do**

$$\alpha = \frac{\sum_{i=1}^{j} n(SA_i^S)}{j}$$

14: **end for**

The algorithm for parameter selection selects a sample S such that $S \subset T$. Based on the sample, the SA^S and $n(SA^S)$ are calculated where SA^S is a set of distinct sensitive attribute values and $n(SA^S)$ is the number of distinct sensitive attribute values in a sample S. In this case, $SA^S = SA = \{SA_1, \ldots, SA_j\}$. The ontology [31] is created based on the domain of the sensitive attributes. Semantic Distance

is calculated between each pair of the sensitive attribute values. The average of these pairs is then selected as θ. The main rationale for taking the average is to take the central value of the distribution, which will help in better partitioning and less usage of dummy records. The algorithm can be partially used or omitted if the data publisher wants as it can explicitly assign any or all values to the tuning parameters. Lines 1 - 4 takes the sample and calculate the distinct set of sensitive attribute values from the sample. Lines 5 - 9 finds the semantic distance between corresponding pairs of sensitive attribute values and calculate θ. Lines 10 - 16 creates the partitions of a sample S as per the partitioning algorithm (Phase - 2) and calculates α and ub.

4.6.2 Phase 2: Partitioning

The partitioning phase creates partitions that are semantically dissimilar, and the number of sensitive attribute values in a partition is confined between lower bound and upper bound. The redundancy of the sensitive attribute values are kept in - check such that it is not more than α as to protect from uneven distribution of different sensitive attribute values in a partition. We have incorporated another checkpoint wherein the redundancy should not lead to loss of information; entropy of the partition should not be less than $log(j_p)$, where $j_p = n(SA^p)$ is the number of distinct sensitive attribute values in a particular partition p. A $Flag$ is used to convey the open and closed partitions. Here, open partitions are the partitions that have not exceeded the upper bound of the sensitive attribute values and have room to accommodate distinct semantically dissimilar sensitive attribute values, whereas closed partitions have exceeded the upper bound. Spurious records are added into the partitions. The maximum number of records in a partition is $(\alpha * ub) + (\alpha * sp_{SA})$ where sp_{SA} is the number of distinct number of spurious sensitive attributes.

In Algorithm 2, line 1 creates the initial partition p_1. The Lines 2 - 22 adds records of table T into partitions. Each partition has to fulfil a set of conditions, that is,

each record is only added to a partition if a sensitive attribute of the record is semantically dissimilar with the sensitive attributes present in the partition SA^p. Also, the number of distinct sensitive attributes should be between lb and ub. The redundancy of a particular partition is checked with the help of α_i^{SA}, the number of redundant sensitive attribute i in partition p_j. Once the partitions are generated, Lines 23 - 28 adds the spurious records.

Algorithm 2 Phase 2: Partitioning

INPUT: A Micro data table T and tuning parameters θ, α, lb and ub.
OUTPUT: sp - concealed partitions.

1: Create Initial partition p_1
2: **for each** record r in T **do**
3: **for each** j in partition p_j **do**
4: **if** $flag = open$ **then**
5: **if** $(n(SA^j) = 0)$ **then**
6: Add $[SA_1]$ record in partition p_j.
7: **else**
8: **for each** i in SA^p **do**
9: **if** $(Sem_D(SA_i^p, SA^j) \geq \theta$ && $n(SA^j) \in [lb, ub])$ **then**
10: **while** $((\alpha_i^{SA} < \alpha)$ && $(E(p_j) \geq log(n(SA^j))$ **do**
11: **if** $(r[SA_i] \notin SA^j)$ **then**
12: Add $r[SA_i]$ record in partition p_j
13: **else**
14: Add record $[SA_i]$ in partition p_j
15: **end if**
16: **end while**
17: Create a new partition p_{j+1}.
18: **end if**
19: **end for**
20: **end if**
21: **end if**
22: **end for**
23: **end for**
24: **for each** j in partition p_j **do**
25: **if** $(n(SA^p) = ub || lb \leq n(SA^j) \leq ub)$ && $(\alpha_i^{SA} \leq \alpha))$ **then**
26: flag = closed
27: Add $Spurious\_Records(p_j)$
28: **end if**
29: **end for**

4.6.3 Phase 3: Spurious Record Generation

Algorithm 3 Spurious Record Generation

```
1:  procedure SPURIOUS_RECORDS(p_j)
2:      for each each partition p_j do
3:          sp^p_SA = Add k% of n(SA^p)
4:          for each each i in sp^p_SA_i do
5:              sp^p_QI_i = Random(QI) in p_j
6:              sp^p_i = (sp_QI_i, sp_SA_i)
7:          end for
8:      end for
9:      return sp^p_i
10: end procedure
```

The function for adding spurious record works as follows: In each partition, k% of the distinct number of sensitive attribute values present in a partition p_j are added as spurious sensitive attribute values in p_j. The spurious quasiidentifiers sp_{QI} is randomly selected from the partition p_j. sp_{SA} is selected as per definition 4.8 (as per lines 2 - 9).

Before going into detail with the mechanism, we explain the important parameters that are instrumental in the mechanism as follows:

1. *Social Connections*(SC) are a set of sensitive attribute values that are most probable to be known to the adversary based on the social links that can be established in a partition p. It is a cohesive knowledge.

2. *Public Information*(PI) is a set of sensitive attribute values that are most probable to be known to the adversary based on the public domain knowledge of correlations between quasiidentifiers and sensitive attributes, probabilistic and inferential knowledge, knowledge based on demographics that can be established in a partition p.

The mechanism to add spurious sensitive attribute values are as follows: The spurious sensitive attribute values in a partition p is selected such that it does not contain any sensitive attribute values that are present in the probable set of social connections (SC), public information (PI), semantically and syntactically, similar attribute values in a partition p. The publisher adds spurious sensitive attribute values into the partition p. We define the spurious sensitive attribute value as follows:

Definition 4.8. Spurious Sensitive Attribute Value: For any given spurious sensitive attribute value in a partition p, $sp^p_{SA_i} \cap \{SC^p,\ PI^p,\ SA^p\} = \phi$ and $Sem_D(sp^p_{SA_i}, SA^p) \geq \theta$.

Here, $sp^p_{SA_i}$ is a spurious sensitive attribute value in a partition p, SC_p, PI_p and SA_p are the set of probable social connection sensitive attributes in partition p, set of probable public information sensitive attributes in partition p and the set of sensitive attribute values already present in the partition p. k is useful quantification that helps in providing privacy. However, k needs to be selected cautiously as this would make the data useless too. The data publisher can select based on the privacy required for a data set.

4.6.4 Phase 4: Anatomization[30]

Anatomization is a privacy technique to publish the sensitive data [30]. It uses l - diversity to partition the data where the probability of each sensitive attribute value should not be less than $\frac{1}{l}$. After creating sp - concealed partitions, it bifurcates the data into two different tables, quasiidentifier table and sensitive attribute table. The quasiidentifier table consists of quasiidentifiers and group id. The sensitive attribute table consists of the group id that is corresponding to the quasiidentifier table, sensitive attributes and the count of the sensitive attributes for a partition. The technique came into existence to overcome the shortcomings of generalization in terms of utility. The utility is affected by generalization and suppression as it fails in capturing the distribution of the data in quasiidentifier

partition. The anatomization protects the utility by bifurcating the tables rather than generalizing. We do not use the l - diversity principle in our proposed privacy model. Also, as we are adding spurious records, the utility is definitely going to get affected. To lessen the consequence, we use anatomization instead of generalization and suppression.

4.6.5 Working Example

ZipCode	Age	Gender	Disease
15036	20	M	CAD
15071	32	F	Breast Cancer
15036	24	F	Viral Infection
15038	27	M	Prostate Cancer
15075	37	F	Gastritis
15078	31	M	Viral Infection
16032	41	F	Gastric ulcer
16035	45	M	Kidney Failure
16031	48	F	Flu

Table 4.3: Inpatient Micro - data Table

An overview of the working of our proposed privacy model is explained with an example as follows: Table 4.3 consists of the Patient Microdata. We apply the algorithm on Table 4.3. Based on the phase 1, we get $\alpha = 1$, $ub = 3$, and $\theta \geq 7$. Now, in phase 2, we generate partitions based on the tuning parameters as follows: p_1 = {CAD, Viral Infection, Prostate Cancer}, p_2 = {Breast Cancer, Gastritis, Viral Infection }, p_3 = {Gastric Ulcer, Kidney Failure, Flu }. The partitions p_1, p_2 and p_3 are semantically dissimilar, $[lb, ub]$ bound and α clustered. Following that, the spurious records are added to the partitions (as per phase 3). After phase 4, we get table 4.4 and table 4.5 as an output.

ZipCode	Age	Gender	Group Id
15036	20	M	1
15036	24	F	1
15038	27	M	1
15038	24	F	1
15071	32	F	2
15075	37	F	2
15078	31	M	2
15075	32	F	2
16032	41	F	3
16035	45	M	3
16031	48	F	3
16032	45	F	3

Table 4.4: Quasiidentifier Table (QI_T)

Group Id	Disease	Count
1	CAD	1
1	Viral Infection	1
1	Prostate Cancer	1
1	Arthritis	1
2	Breast Cancer	1
2	Gastritis	1
2	Viral Infection	1
2	Nerve Compression Syndrome	1
3	Gastric Ulcer	1
3	Kidney Failure	1
3	Flu	1
3	Depression	1

Table 4.5: Sensitive Attribute Table (SA_T)

4.7 Analysis of $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private: Privacy Model

We analyze the proposed privacy model against the adversarial model defined in Chapter 3.

Theorem 4.1. *The linking probability of a $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private Table T' when BK is present is negligible against a strong adversary.*

Proof. The number of records present in the $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private table T' :

$$T' = k * ((\alpha * ub) + (\alpha * sp_{SA}))$$

Here, k is the number of partitions in the table. α is the number of redundant records of a particular sensitive attribute value. ub is the upper bound of distinct number of sensitive attribute value that can reside in a partition. sp_{SA} is the distinct number of spurious sensitive attribute values in a partition. All the partitions are semantically dissimilar. We have assumed uniform partitions. Table T' has been bifurcated into two tables quasiidentifier Table QI_T and sensitive attribute Table SA_T. The schema of quasiidentifier Table (QI_T) and sensitive attribute Table (SA_T) are $QI_T(QI', G_{id})$ and $SA_T(G_{id}, SA', Count(SA'))$ respectively. In schema of QI_T, QI' is the set of quasiidentifiers along with spurious quasiidentifiers and G_{id} is the group id of a particular partition. Similarly, in schema of SA_T, G_{id} is the group id of a particular partition corresponding to the QI_T table's partition, the corresponding sensitive attribute values SA' and count of the sensitive attribute value as $Count(SA')$. The set SA' consists of sensitive attribute values along with spurious sensitive attribute values. For simplicity, we omit the column of count in SA_T. This will abide by the condition 4 of definition 4.7. We assume a uniform partition and, as a result, do not consider the parameter α for further calculations. Here, we assume that the strong adversary A_s is able to shrink to a particular partition. Therefore, the linking probability of a partition when background knowledge is not present is $\left(\frac{1}{ub+sp_{SA}}\right)^{ub+sp_{SA}}$.

1. $KE(BI) \rightarrow K_{Pr}$: This step extracts knowledge as per section 3.2.2 from the Background Information (BI). As, knowledge can be obtained from anonymized table T' and public domain apart from individual knowledge. Therefore, $BK \neq \phi$.

2. $LINK(T', K_{Pr}) \rightarrow PR$: This step takes as input the anonymized table T' and background knowledge K_{Pr}. The output is $P_{KE}^{A_s}(T'|K_{Pr}) = \left(\frac{1}{ub+sp_{SA}}\right)^{ub+sp_{SA}} + \Delta$. Here, Δ is negligible. We have added spurious sensitive attribute values as per Definition 4.8. But, Δ is added due to individual knowledge.

3. $CHECK(PR) \rightarrow f$: The $|P_{KE}^{A_s}(T'|K_{Pr}) - P_{KE}^{A_s}(T')|$ is negligible as the increase in probability is negligible due to spurious sensitive attribute values.

As a result, the strong adversary A_s is unsuccessful in linking the individual to a sensitive attribute value. The privacy attack due to background knowledge is not successful on T'. Therefore, $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private does preserve privacy against strong adversary as $|P_{KE}^{A_s}(T'|K_{Pr}) - P_{KE}^{A_s}(T')|$ is negligible. ■

4.7.1 Rationale

In our thesis, Theorem 3.3, Theorem 3.4, and Theorem 3.5 analyses the existing privacy models (k-anonymity, l-diversity and t-closeness, respectively) against the adversarial model (As discussed in Chapter 3). The proposed privacy model is analysed (Theorem 4.1) against the same adversarial model (As discussed in Chapter 3) in our thesis. The analysis shows that the proposed privacy model abides by the definition of a strong privacy model (As discussed in Chapter 3), whereas the existing models do not. Here, strong pinpoints to the adversarial capabilities. The proposed strong privacy model protects against background knowledge (As defined in Chapter 3) that is comprehensive and realistic. The previous models assumed selective background knowledge which is prone to privacy attacks. Moreover, the proposed privacy model protects from attribute and identity disclosure in terms of privacy disclosure. Due to its background

knowledge assumption that an adversary can avail, the proposed privacy model is stronger notion.

4.8 Experiments and Results

4.8.1 Datasets

We use the following setup for experiments of the proposed privacy model. We use two datasets for our experiments, namely Adult Data set and Census Income data set from UCI machine learning repository [34]. Adult data set consists of a total of 30162 records after excluding the null values. Census Income (KDD) dataset consists of a total of 196130 records after excluding null values.Table 4.6 and 4.7 gives us a brief description of the datasets used in the experiment in terms of the attributes, its type as well as the domain of attributes.

Attribute	Type	Distinct values
Age	Numeric	72
Workclass	Categorical	8
Education	Categorical	16
Race	Categorical	5
Sex	Categorical	2
Native Country	Categorical	41
Occupation	Categorical	14

Table 4.6: Adult Data Set

4.8.2 Prerequisites

The experiments are performed on 3.20GHz Intel Core i5 machine with 4GB RAM. The proposed algorithms are implemented in Java. We have used ARX data anonymization tool [35] for k anonymity and l - diversity privacy models. To study the influence of semantic similarity on sensitive attribute values, we create different variants from Adult and census income dataset, respectively. ADT1, ADT2, ADT3, ADT4 and ADT5 are datasets from Adult dataset each with a different sensitive attribute. CEN1, CEN2, CEN3, CEN4 and CEN5 are datasets from

Attribute	Type	Distinct values
Age	Numeric	91
Class of Worker	Categorical	9
Education	Categorical	17
Major Industry Code	Categorical	24
Race	Categorical	5
Sex	Categorical	2
Reason of Employment	Categorical	6
Full or Part time employment	Categorical	8
Tax filer stat	Categorical	6
Detailed household summary in household	Categorical	8
Country of birth self	Categorical	42
Major Occupation Code	Categorical	15

Table 4.7: Census Income Data Set

Census dataset each with a different sensitive attribute. For example, in ADT1, the quasiidentifiers are Age, Workclass, Education, Race, Sex and Native Country. Occupation is a sensitive attribute. We summarize Datasets in Table 4.8.

Data Set	Sensitive Attribute
ADT1	Occupation
ADT2	Race
ADT3	Workclass
ADT4	Education
ADT5	Native Country
CEN1	Major Occupation Code
CEN2	Race
CEN3	Class of Worker
CEN5	Education
CEN4	Country of birth self

Table 4.8: Summarization of Data Sets

4.8.3 Experimental Results

We evaluate $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private model based on efficiency and privacy. The parameters θ and ub are deciding factors for partitioning. The semantic similarity focuses on how closely the sensitive attributes need to be related semantically, which helps in deciding the upper bound of the partition. We have experimented

on 10 datasets to know the impact of θ and ub on each other. The results are shown in Table 4.9.

Data Set	Distinct SA	θ_l	ub_l	θ_u	ub_u
ADT1	14	5	3	6	3
ADT2	5	2	5	3	2
ADT3	8	3	3	4	2
ADT4	16	5	3	6	2
ADT5	41	3	4	4	4
CEN1	15	5	4	6	3
CEN2	5	2	5	3	2
CEN3	9	3	3	4	2
CEN4	17	5	3	6	3
CEN5	42	4	3	4	3

Table 4.9: Interdependence of θ and ub

θ_l and θ_u are the lower and the upper values of the parameter θ. Similarly, ub_l and ub_u are the lower and the upper values of the parameter ub with respect to θ_l and θ_u respectively. Generally, θ_u is preferred as compared to θ_l as it would give more semantically distinct partition. In other words, more privacy from semantic knowledge. The case where ub is less than 3, θ_l value is selected (As depicted in ADT2 in Table 4.9). The two crucial factors that help in tuning the parameters are 1) Semantic closeness amongst the sensitive attribute values 2) The number of sensitive attribute values. If sensitive attribute values are not semantically correlated, then θ would be on a greater side. Similarly, if sensitive attribute values are semantically correlated then the value of θ would be less. In table 4.9, this has been depicted in dataset ADT1 and ADT5. In general, the range of θ is $Min(Sem_D^{SA}) \leq \theta \leq Max(Sem_D^{SA})$, where $Min(Sem_D^{SA})$ and $Max(Sem_D^{SA})$ are minimum and maximum semantic distance amongst the sensitive attribute values of sensitive attribute SA in a dataset (The semantic distance is obtained from an existing ontology). More is the semantic distance; less is the value of ub and vice verse. In Table 4.9, ADT1 has $\theta = 6$ and $ub = 3$ whereas ADT5 has $\theta = 4$ and $ub = 4$. The range of ub is $lb \leq \theta \leq SA_j$, where lb is 3 and SA_j is the number of distinct sensitive attributes of a dataset.

α is a redundancy parameter which is selected with the help of sample S of the published dataset. We compare the sampling methods for selecting more precise parameter α. We will only consider ADT1 and CEN1 for further evaluations as the distribution is similar in all ADT's and CEN's. We experiment the datasets using two sampling methods that is random sampling method and stratified sampling method. Both methods are random in nature. Figure 4.1 and Figure 4.2 shows that stratified sampling generates a constant parameter α. Stratified sampling works best with uniformly distributed data. The random sampling method is more precise with the original (original value is computed for the complete dataset rather than sample S) α value as compared to stratified sampling. The original value of α is 2154 and 13075 of ADT1 and CEN1 dataset, respectively. Therefore, the random sampling method is selected.

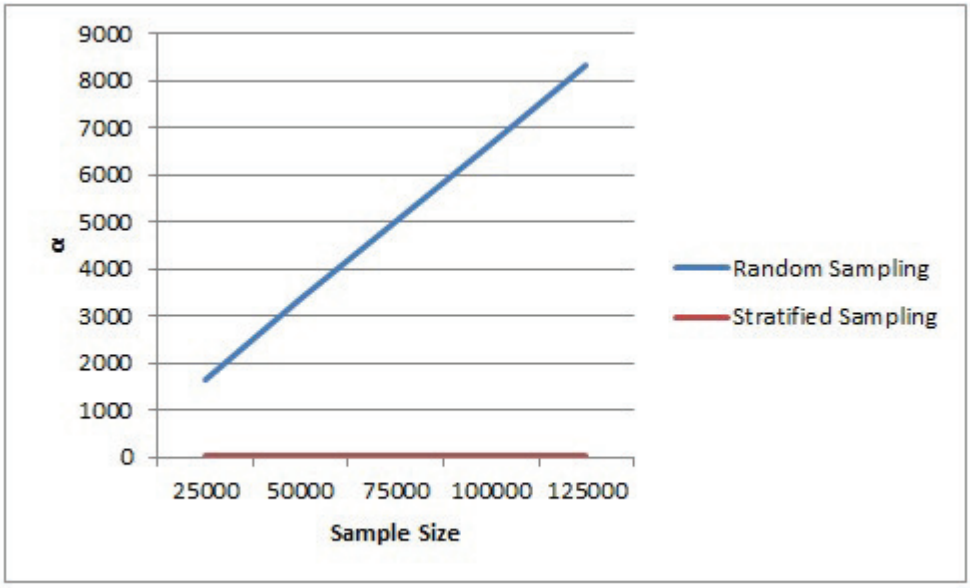

Figure 4.1: Impact of Sampling methods and α in ADT1 Dataset

We experiment the impact of parameters θ and ub on partitioning phase. We take a sample of ADT1, ADT3 and ADT4 dataset and experiment different combinations of θ and ub on the datasets. Lower value of θ will lead to less privacy as semantically close sensitive attribute values has lower θ. (As shown in Table 4.9). In Figure 4.3 in ADT1, higher value of θ will require more number of partitions

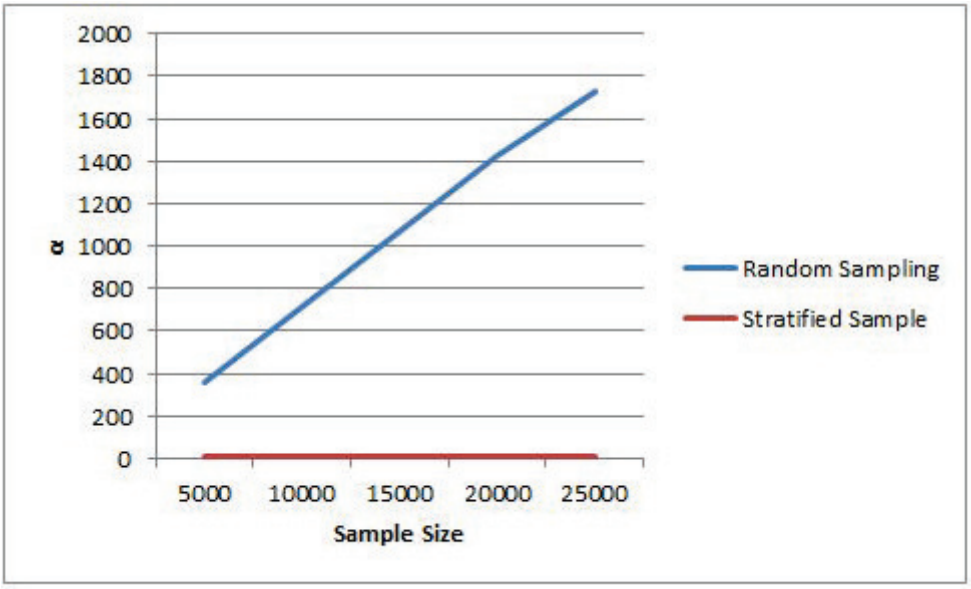

Figure 4.2: Impact of Sampling methods and α in CEN1 Dataset

to accommodate the semantically dissimilar sensitive attribute values. In Figure 4.4 in ADT4, higher value of θ will result in increase in the number of partitions. However, in Figure 4.5, in ADT3, as the domain of sensitive attributes are narrower, the impact of θ is less compared to ADT1 and ADT4. If the value of θ decreases; the partition consists of more semantically correlated attributes which will decrease the privacy. On the other hand, if value of θ increases; the number of sensitive attributes decreases in a partition, which will affect the privacy as there will be less number of guesses. As a result, the best value of θ should be the average. For example, In ADT1, the value of θ is 5 or 6. In a nutshell, an efficient partitioning depends on the three factors: number of sensitive attribute values, semantic closeness amongst the attribute values and distribution of a dataset. We now evaluate our proposed model based on privacy parameters.

We consider ADT1 dataset for evaluating privacy. We compare the privacy of the proposed model with l - diversity and k - anonymity model. We have chosen θ = 6, and it is corresponding ub value as 3. To show the uniformity in comparison, we have chosen l and k as 3. The reason for choosing θ = 6 is that lower θ would lead to semantically similar partitions whereas higher θ would lead to lower ub,

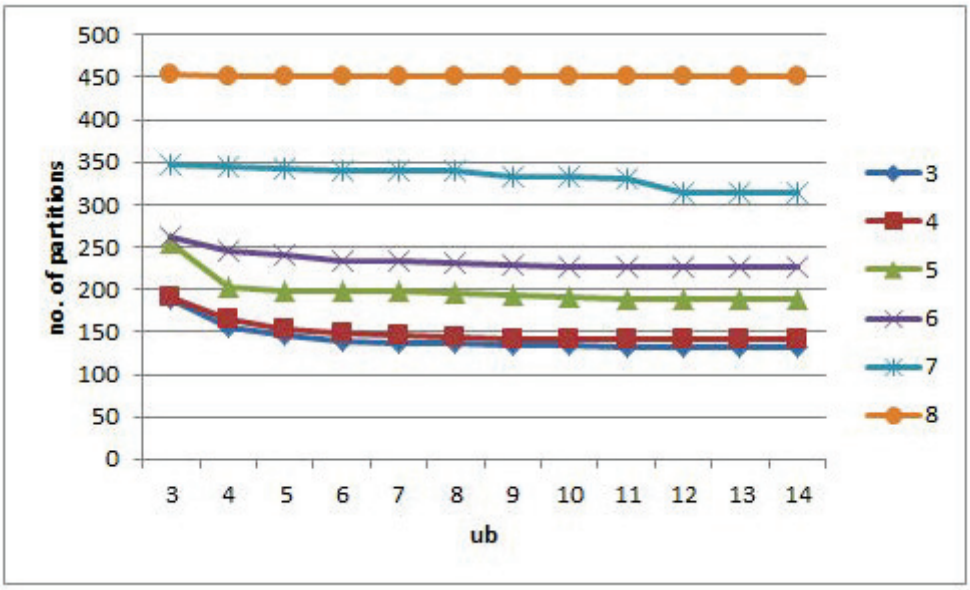

Figure 4.3: Impact of θ and ub on partitioning in ADT1

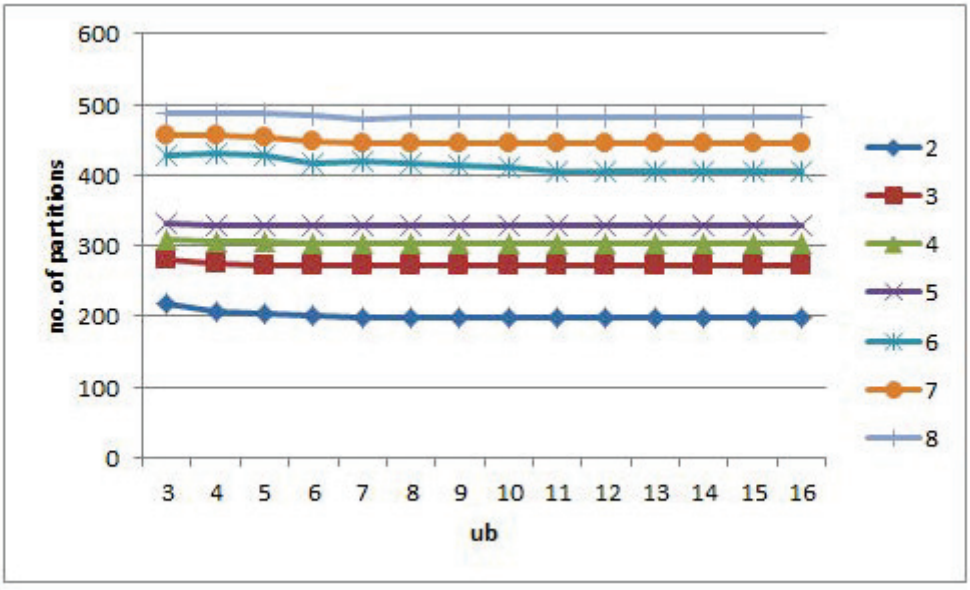

Figure 4.4: Impact of θ and ub on partitioning in ADT4

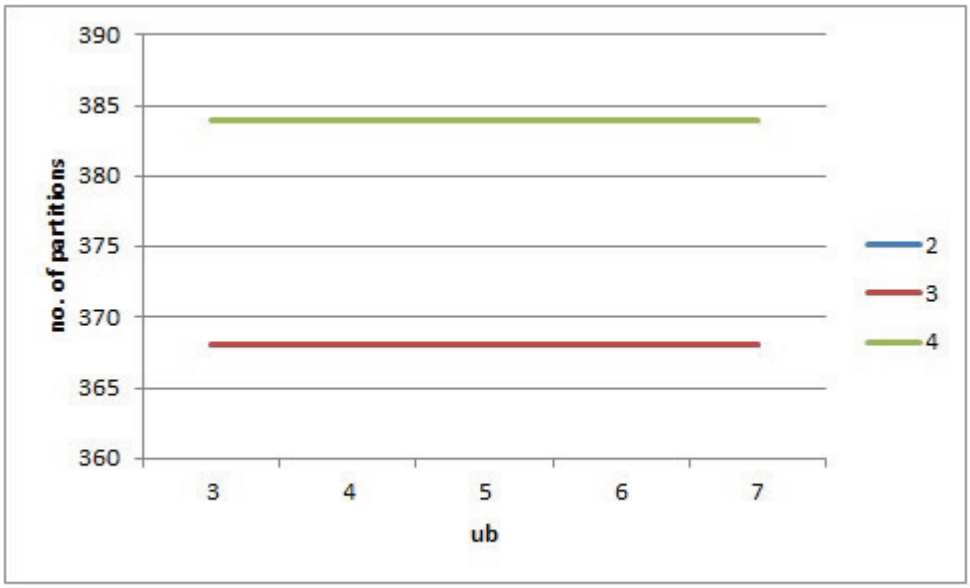

Figure 4.5: Impact of θ and ub on partitioning in ADT3

which would decrease privacy as there is lesser number of guesses to eliminate the sensitive attribute values.

Privacy Model	Entropy
Proposed privacy model $(\theta = 6, ub = 3)$	3.19
k - anonymity, l - diversity $(k = 3, l = 3)$	2.373
l - diversity $(l = 3)$	2.3734

Table 4.10: Entropy Comparison

Privacy: We measure privacy in terms of entropy and conditional entropy. The main objective of the proposed model is that in the presence of the background knowledge, the adversary should not compromise the privacy. Therefore, we measure the conditional entropy as it measures the uncertainty of sensitive attribute when background knowledge is given. More is the uncertainty more is privacy. We explain the above claim with the argument as follows:

Our proposed privacy model adds spurious sensitive attributes to preserve pri-

vacy. The entropy of the proposed privacy model is greater than the k - anonymity and l - diversity model. Suppose, j_p is the number of distinct sensitive attribute values in a partition p in k - anonymity and l - diversity model and $j_p + sp^p_{SA}$ is the number of distinct sensitive attribute values in a partition p. sp^p_{SA} is the spurious sensitive attribute values added in the proposed privacy model. $log(j_p + sp^p_{SA}) > log(j_p)$. As a result, entropy is higher in our proposed privacy model. Moreover, spurious sensitive attribute value increases the conditional entropy in the proposed privacy model. Our experimental results support the same notion.

Table 4.10 shows the entropy comparison on the ADT1 dataset. Our proposed model shows 34.4 % increase in entropy, which depicts more privacy. Figure 4.6 shows the results where our proposed model performs better as compared to the l - diversity and k - anonymity model when background knowledge is present. In terms of conditional entropy, our proposed model shows a significant increase in uncertainty in compromising the sensitive attribute (e.g. occupation).

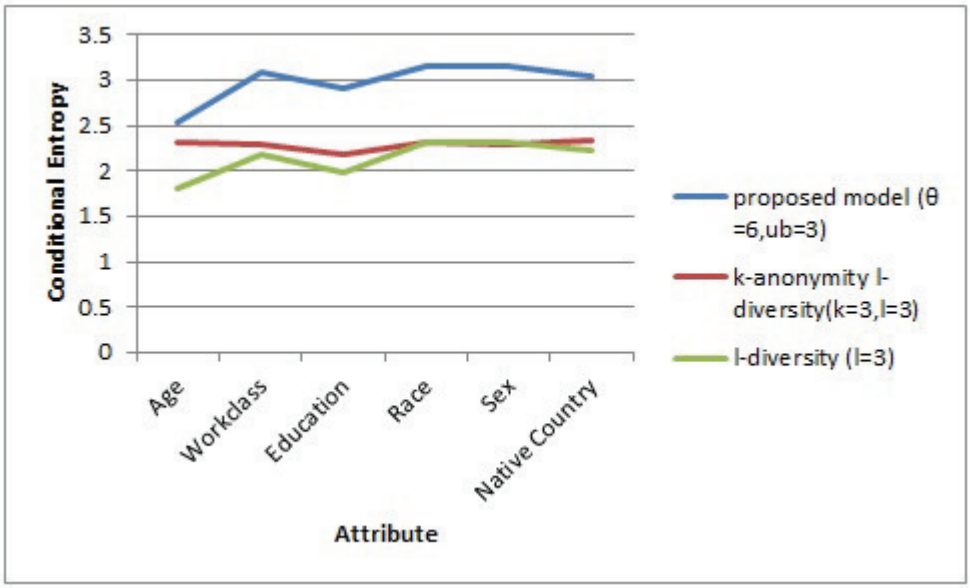

Figure 4.6: Privacy against Background Knowledge

4.9 Conclusion

Data Privacy has become a potential concern in the data - centric world. In this paper, we have formally defined a novel privacy model $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private which protects against background knowledge. Moreover, we have proved the wider implications of semantic knowledge. We theoretically analyze $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private against the strong adversary. We evaluate the proposed model experimentally and demonstrate that $(\theta, [lb, ub]^{+sp}, \alpha)$ - Private model preserves privacy against background knowledge.

The proposed work focuses on privacy solution. The proposed work can also be applied to different applications like location - based services and social networks as it considers a comprehensive background knowledge assumption.

CHAPTER 5

Privacy - Preserving Data Publishing in Social Networks

Social Networks is a prominent application of privacy - preserving data publishing domain. With times, social networks have seen massive development in terms of wider reach and acceptance across the globe. Data privacy is a pertinent research problem in social networks in the current data - centric era. Privacy attacks due to background knowledge are inevitable in social networks as well. This chapter in our thesis links our study of background knowledge in the relational table to social networks. In this chapter, we discuss how social networks are different from relational tables concerning privacy disclosure. We also study various privacy disclosure approaches in the existing literature of social networks data publishing followed by motivation to consider the research problems addressed in chapter 6 and chapter 7.

5.1 Introduction

The current decade has seen Social Networks flourish rapidly across the globe. Presently, the number of social network users [139], [140] also has seen a heavy surge on the Internet. The data in [139] projects around 4.33 billion social users at the starting of the year 2021 and is further going to increase progressively. As a result, millions of social networks data generated by social network users is invaluable to understand the dynamics of society. However, the third - party applications can access the collected social networks data to mine insightful knowledge

and observations. Indeed, this will give constructive inputs in moulding social - economic government policies, understanding the business trends and leverage research prospects. Recently, the COVID - 19 pandemic has proven the impact of social networks data [141], [142] on contact tracing and identifying superspreaders to cease the virus transmissions. This makes the social networks data an asset and a significant contributor to the betterment of society.

Typically, the social network's data consist of age, location, gender, political inclination, occupation, education, hospital check - ins, social relations, social communications, likes, dislikes and hobby groups. There are several instances of data breach [143], [144], [146] acting as a critical threat to the social networks users. The social network's data consist of users' sensitive information [145], and its disclosure can lead to devastating consequences affecting the users' reputation and trust. Some examples of sensitive information [145] in social networks data are location, political inclination, gender, salary, and many more. The sensitive information in social network data makes the data susceptible to privacy attacks. On the other hand, information like identification information, social profiles, social connections, previously published data are substantially available in the public domain. The adversary can easily access this information and utilize it for disclosing the user's sensitive information. Furthermore, the adversary has widened its resources by accessing diverse data and applying powerful data manipulation techniques. As a result, Data Privacy in Social Networks is a potential research challenge considering its popularity and magnitude across the population.

5.2 Transition from Relational Tables to Social Networks

The relational table is represented in a tabular form that contains identification information (known as quasiidentifiers) and sensitive information. The social networks, in general, is depicted as a graphical structure where nodes represent users and edges represent the relations. In relational tables, the record (each record

points to a single user/individual) act as an independent entity that doesn't get affected by the other records. In contrast, a user in social networks act as a dependent entity that get affected by the other users. Here, in social networks, the user's social relations changes with any structural change. As a result, data privacy in social networks is more complex and challenging as compared to relational tables.

Relational tables have rich and widespread literature on data privacy solutions to protect against privacy disclosure due to the adversary's background knowledge. Some prominent data privacy solutions are k - anonymity [1], l - diversity [2], t - closeness [3]. In social networks, the user's social structure possess useful information and can be exploited by the adversary to disclose the user's privacy. As a result, the privacy solutions for relational tables can not be directly incorporated into social networks, considering their structural differences. The social networks literature has several data privacy solutions to protect user's privacy against background knowledge possessed by the adversary. Various data privacy solutions have been designed to protect the user's privacy against the adversary's background knowledge in social networks. Various privacy solutions [25], [63], [155] in social networks are on the line of k - anonymity and l - diversity. The privacy solutions consider different background knowledge assumption. For example, k - degree anonymity [59] assumes background knowledge as degree information of some users. k - candidate anonymity [26] assumes background knowledge in terms of vertex refinement queries, subgraph queries and hub fingerprint queries. k - anonymity and l - diversity approaches in social networks [39] , [56] protects against neighborhood attacks (neighborhood information of some users). k - isomorphism [58] protects against neighbourhood attack graph (It consists of neighbourhood information as well as the node information of a user whose privacy needs to be disclosed). k - automorphism [60] protects against structural attacks that comprise degree information, sub - graph information, single - hop neighbourhood information and hub information. k - degree l - diversity [61] protects against degree information that can help disclose sensitive information. k^2

- degree anonymity [57] protects against friendship attack (It consists of degree information of users with friendship relation). A graph - based privacy preservation framework [64] assumes background knowledge as probabilistic information about some users as well as knowledge related to correlations between any information. Probabilistic Indistinguishability [156] considers weighted single - hop neighbour information and it's degree information. (k, l) - anonymity [157] considers the length of shortest path between the user (whose privacy needs to be disclosed) and other users. $(k, \Gamma_{G,l})$ - adjacency anonymity [158] assumes neighbour information about some users. Certainly, background knowledge plays an important role in the data privacy solutions for social networks similar to relational tables.

The scope of the adversary broadens in terms of accessibility to knowledge with the emergence of social networks. For social networks, the adversary can have knowledge [25] ranging from information about the number of friends of the user, neighbourhood information, information about the relation between users, identification information, measures like betweenness, centrality, and many more. Recently, the adversary has expanded its capabilities in terms of knowledge [6], [7] related to correlations between information and rule - based mining techniques in social networks. The magnitude of knowledge has widened when considering social networks. Besides, social relations play a critical role in disclosing privacy as it adds more value to the existing set of knowledge. Certainly, Background Knowledge (adversary's knowledge) has significantly evolved with the progression of social networks.

In addressing the privacy concerns, a need arises to analyze the background knowledge possessed by the adversary. Modelling the background knowledge in social networks is challenging but at the same time crucial due to its complexity and widespread acceptance. In the next section, we discuss the existing privacy disclosure approaches in the social networks literature.

5.3 Analyzing Privacy in Social Networks

Privacy in social networks is disclosed when the adversary can identify user/s and disclose their sensitive information. The adversary can possess information ranging from auxiliary graphs (other than the published graphs) to specific user (individual) information like the number of friends, the information about neighbours, and many more. The privacy of users can be disclosed in social networks using the information possessed by the adversary. The possibility of disclosure leaves users in great distress when submitting their data to social networks. Either the users submit inaccurate data or restrain to submit data. With the increased manipulation capabilities of the adversary, it can still disclose users' unpublished sensitive information.

Specifically, there are two types of approaches used for privacy disclosure in the existing literature of social networks. A brief overview of the approaches are as follows:

- **Privacy Disclosure due to De-anonymization:** De - anonymization [24], [27], [48], [49], [63] in social networks has evolved considerably in the last decade. Narayanan et al. [24] de - anonymizes the social networks by identifying seeds (common users) and then propagate users (other than seeds) based on structural properties like eccentricity, edge directionality, degree information. Nilizadeh et al. [48] de - anonymizes social networks using community structure. Pedarsani et al. [47] de - anonymizes social networks based on only the structural information. Ji et al. [49] de - anonymizes social networks using structural and profile (attribute) information. Li et al. [27] considers heterogeneous social networks and maps users based on structure and profile information. Qian et al. [7] de - anonymizes social networks using knowledge graph. Typically, Privacy disclosure occurs when an adversary identifies users in the published social network data using background knowledge [24]. This in turn will disclose the sensitive information associ-

ated with the users. User's sensitive information is disclosed if published on social networks. We discuss this approach in detail in chapter 7.

- **Privacy Disclosure due to Inference:** Inference attack [6], [7], [55], [131], [132], [133], [147], [148], [149], [150] has gained momentum with the rise in machine learning and data mining techniques. Inference attack in social networks considers inferring location, inferring user's information (attribute), inferring social relations, and many more. Crandall et al. [150] discuss inferring of the social relations between users when they have multiple instances of their presence at approximately similar geographical location and time. Dong et al. [55] infer the user's demographics, in particular, gender and age, based on their communications patterns in mobile social networks. Mislove et al. [149] discuss inferring user's information (attributes) based on their communities in the social networks as users sharing similar information (attributes) will be able to form denser communities. He et al. [132] discuss inferring user's information (attributes) based on their social relations in social networks. Ryu et al. [147] discusses inferring a user's sensitive information (attribute) based on the publicly available information (attribute) associated with its social relations and the impact of this disclosure on its social relations. Gong et al. [133] infer user's information (attribute) based on user's social structure and user's behaviour information. Gong et al. [133] modelled attribute inference attacks using the social - behaviour - attribute network model. Zhong et al. [148] address demographic inference from location information (check - ins) available on the social networks. It specifically targets human mobility to infer the demographic information of the user. Nie et al. [132] predict occupation by collecting all occupation information available on various online social platforms using the user attribute learning model. Jurgens [131] infers user's location based on the sparse location information available with its social relations in social networks. Cai et al. [6] modelled inference attack using attribute and relation - based classifier. Cai et al. [6] infers user's sensitive information using rule - based mining. Qian et al. [7] modelled the inference attack using a knowledge graph, with

the assumptions that the adversary has more specific information about an individual whose privacy is to be disclosed. Specifically, it has focused on correlational knowledge in terms of inference. In general, Privacy disclosure occurs if an adversary infers the user's unpublished/unknown sensitive information. The adversary uses different data mining techniques as well as background knowledge to infer the sensitive information. It is applicable when sensitive information is not published/ not known in the social network data. We discuss this approach in detail in chapter 6.

Both the above approaches are capable enough to disclose the sensitive information. Nevertheless, de - anonymization identifies users but will not disclose sensitive information if unpublished/unknown. As a result, inference attack is more severe and powerful in terms of privacy disclosure in social networks data.

5.4 Motivation

Social Networks Data Publishing is a flourishing research domain at the juncture of technological advancement. Privacy is a matter of concern while publishing social data due to users' sensitive information. Anonymization techniques [1], [2], [25], [63] are widely used for preserving the privacy of user(s) data while publishing or sharing it on the digital platform. Nevertheless, there are several instances, where the social networks data got de - anonymized, and privacy got disclosed [6], [7], [24], [27], [47], [49], [63], [150]. Due to the de - anonymization threat, many users cautiously publish their sensitive information, either resist or submit less accurate sensitive information. Despite that, the adversary can still infer the unpublished sensitive information from the published dataset using the background knowledge. The adversary can accumulate background knowledge spanning from auxiliary graph to specific user (individual) information like identification information, social relations, and many more. Lately, the adversary has become more powerful by applying various deep learning and data mining techniques to infer user's unpublished sensitive information. Modelling extensive

and exhaustive background knowledge will pave the way for stronger and comprehensive privacy solutions to protect user's social data. As a result, background knowledge in social networks is a challenging and important research issue to address.

This has motivated us to study background knowledge in social networks. De - anonymization and inference attacks are the two privacy disclosure approaches that need to be analyzed. In our thesis, we address the two research problems in social networks as follows:

- We model inference attack due to rule - based mining techniques in social networks. We also propose a new privacy model against inference attack due to rule - based mining techniques. (Chapter 6)
- We devise a de - anonymization technique against comprehensive adversarial background knowledge in social networks. It will help in designing more robust privacy models against comprehensive background knowledge in social networks. (Chapter 7)

5.5 Conclusion

Social Networks is a prominent application of the privacy - preserving data publishing field. Social Networks expand the scope of the adversary's background knowledge and helps compromise the privacy of its user(s). In this chapter, we discuss the privacy disclosure approaches due to background knowledge in social networks. We also discuss the motivation for considering two research problems addressed in chapter 6 and chapter 7.

CHAPTER 6

Rule - based Anonymization against Inference Attack in Social Networks

Inference attack due to rule - based mining technique has garnered attention in social networks data publishing which motivates the adversary to infer the unpublished sensitive information using rule - based mining techniques. This will indeed disclose the privacy of social network users who have restrained from publishing their sensitive information. A strong privacy solution will encourage social network users to submit information more confidently and at the same time preserve privacy. In this chapter, we address the following: First, we have rigorously analyzed the existing data sanitization technique [6] against inference attack using rule - based mining techniques. We have found weakness [151], [154] in [6]. Second, we have modelled inference attack due to the rule - based mining technique in terms of the association of a sensitive attribute with an identification attribute. Third, we propose a privacy model - Rule Anonymity [151] against inference attack due to rule - based mining technique. Rule anonymity can provide a strong privacy guarantee such that the presence of rules should show negligible impact on the privacy of sensitive information. Fourth, the proposed model uses a rule - based anonymization technique that follows the principle of rule anonymity. The proposed privacy model is analyzed against strong adversarial capability. Fifth, the proposed model experiments on a social dataset and the experimental results show the proposed model is practical in preserving privacy.

6.1 Introduction

Social networks data have captivated a lot of market decisions in recent times. Humongous volumes of data get collected by various networking applications. The collected data is rigorously analyzed and mined to understand the dynamics of users in society in terms of concerns and their probable solutions. This indeed facilitates in designing policies, business solutions and empowering socio - economic growth. Despite several advantages of social network data, privacy has been an important concern. Third - party applications exploit published social network data as it contains sensitive user information. The disclosure of user's sensitive information will result in privacy disclosure. Therefore, privacy and utility are two loggerheads of social network data publishing and is a challenging research concern to address. De - anonymization in social networks [25] is dominating the current research direction in social networks. Various techniques [6], [7], [24], [26], [27], [132], [133] are used to model background knowledge of adversary. What makes de - anonymization in social networks a challenging research direction is the complexity of information that social networks possess. Specifically, the adversary has profile information as well as the social relations of users in social networks. This has drastically increased the adversary's capabilities. Additionally, user information related to location, identification, communication, behaviour across diverse platforms is captured by the adversary extrapolating the situation. As a result, social data users are apprehensive about publishing information. Either they do not publish sensitive information or publish inaccurate sensitive information. However, the adversary can still predict the sensitive information by various existing techniques [6], [7], [55], [131], [132], [133], [147], [148], [149], [150] like rule - based mining, link prediction techniques, structural properties, social - behaviour - attribute model, knowledge graph, based on location. This is known as an inference attack in social networks. The competence to disclose user's privacy makes inference attack a potential research problem in social networks. We discuss some related works for inference attack in social networks as follows.

He et al. [132] talks about inferring user's sensitive information using their social relations. The intuition of the work is that socially well - connected people have high chances of sharing similar information. In other words, a user's unpublished/unknown sensitive information (attributes) is inferred by considering the information (attributes) of their social relations in social networks. Specifically, the social relations considered in this work is 'friends'. [132] has modelled the privacy inference using Bayesian networks. The two cases discussed for privacy inference are single - hop inference and multiple hop inference. The single - hop inference considers the information of immediate friends of the user for privacy inference. Whereas multiple hop inference considers the information of extended friends of the user for privacy inference if the information of immediate friends is unknown. [132] incorporates the Bayes Decision rule for predicting the user's unpublished/unknown sensitive information. Mislove et al. [149] target the problem of inferring the user's unpublished /unknown information by using the published information of other users (sparse in number) in social networks along with the social network graph. The privacy inference suggested in [149] is based on two observations. Firstly, friends share more common information (attributes), and secondly, friends with common information (attributes) can lead to a dense community structure. [149] models the privacy inference using the community structure that is created by closely linked users sharing certain common published information in social networks. Dong et al. [55] address the inference of user's demographics (gender and age) based on their mobile communication pattern (behaviours). Specifically, the user's communication pattern [55] shows observations like homophily on gender and age, cross - generation communication and demographics dynamics. The authors [55] model the demographic prediction using a double dependent variable factor graph model, also known as the *WhoAmI* approach. For demographic prediction, it [55] considers the correlation of age/gender with features as well as interrelations between gender and age for prediction. In particular, features [55] considered are individual feature (structural properties of a user), friend feature (demographic information about

immediate friends of a user) and circle feature (demographic information about the user's triads). In other words, [55] models $P(A, B|M, I)$ to predict the user's demographics (gender and age). Here, A and B are the gender and age of the users (whose age and gender are not published), M is the mobile network, and I is the information (features) of users present in M. Crandall et al. [150] infer social ties of users based on their Spatio - temporal coincidences. In this work [150], the location and time information is obtained from social networks and is assumed to be available in the public purview. Specifically, the parameters [150] important in quantifying the spatial and temporal coincidences between two users are cell size, temporal range and the number of co - occurrences. The work [150] claims that sparse information about location and time can infer social ties with high conviction leading to privacy disclosure. The proposed probabilistic model [150] (named Spatio - temporal Co - occurrences) models the social ties with spatial and temporal information. It [150] incorporates Bayes Law for prediction. Jurgens [131] infers user's location from the information obtained from their social network. The information considered in [131] is the social relations of users (ego network of users) and some location information (GPS) of users in social networks. The author in [131] proposed a method named spatial label propagation to infer the location of the users in social networks. It can infer a user's location from sparse location information available with its social relations in social networks. In general, the nearest (in terms of distance) social relation of a user has a high conviction of sharing the same location. But, what if the social relations of a user have no location information associated with it? [131] incorporates a mechanism that considers the scenario when the social relations of the user do not have the associated location information helpful for inferring location. Zhong et al. [148] infer demographic information (like gender, age, education, and many more) of users with the help of its location check - ins in social networks. In particular, the user's location check - in [148] contains information (features) like temporality, spatiality and location knowledge. Temporality refers to the time pattern of the user in terms of the day, weekdays, weekend. For example [148], users with interest in food have restaurant check - ins at breakfast, lunch and dinner time. Spatiality refers to the

location pattern of the user in terms of its movement. For example, a tourist (user) has check - ins for monuments and travel spots. Location knowledge [148] is the user's point of interest (POI) for visiting the particular location check - in. For example [148], students go to college for acquiring knowledge. Specifically, it adds more information to the given location check - in. The authors in [148] proposed a framework named Location to Profile (L2P) to infer user's demographic information in social networks based on their location check - ins. The framework captures the user's location check - in features like temporality, spatiality and location knowledge. Gong et al. [133] infers user's sensitive information (attribute) using their social network structure and behavioural information. [133] models the attribute inference attack using a social - behaviour - attribute network model that incorporates the user's social structure, user's behaviour information and user information (attributes) altogether. The social network [133] is transformed into a social - behaviour - attribute network by adding nodes that represent information (attribute) and behaviour and add edges that connect behaviour and information (attribute) to its user. The attribute inference attack [133] predicts the unpublished sensitive information (attributes) of the targeted users using the publicly available knowledge in terms of social structures, user profiles and user's behavioural information. Nie et al. [132] infer user's attribute information (here, occupation is considered) using social media analytics. The user [132] can have accounts across various social media platforms. Diverse information is submitted on different social media platforms based on its objective. This information can be collected from heterogeneous social media platforms and can disclose more user information than intended. Moreover, different occupation [132] can be related based on their user's social media communication. [132] addresses the above problems by proposing a model named graph - constrained multi - source multi - task learning model for inferring occupation. Ryu et al. [147] discuss inference of a user's sensitive information based on its neighbour's public information. Secondly, whether publishing the given user's sensitive information affects the privacy of other users in its neighbourhood. Qian et al. [7] infers user's unpublished sensitive information in social networks with the help of the adversary's knowledge. [7] models

the adversary's knowledge using knowledge graphs. Typically, Knowledge [7] is represented in the form (s, r, o) where s is the subject, r is the relation, and o is an object. The adversary considers knowledge [7] like structural properties, social relations, facts, user - specific information, information related to demographics and correlational knowledge. In particular, correlational knowledge [7] deals with the correlation between two different pieces of knowledge and is instrumental in privacy inference. Specifically, the variants of correlations discussed in [7] are mutual correlations, Inclusion and soft correlations. In [7], de - anonymization is followed by the privacy inference in social networks using a knowledge graph (generated by the adversary). Cai et al. [6] modelled inference attack using attribute and relation - based classifier. Initially, an attribute - based classifier generates rough set theory based decision rules to infer unpublished sensitive information (attributes). Subsequently, attribute and relation - based classifier [6] are applied together to infer the unpublished sensitive information (attributes), where relation based classifier is based on the intuition that more common information (attributes) the neighbours share; more are the chances of sharing the same sensitive information.

We specifically focus on inference attack using rule - based mining techniques due to their ability to predict unpublished sensitive information. In the current scenario, when rule - based mining techniques are instrumental in mining information, the same can be applied to disclose the privacy of social network users. Indeed, this motivates us to study inference attacks due to rule - based mining techniques in social networks.

6.2 Inference Attack using Rule - based Mining Techniques in Social Networks

Rule - Based Mining [53] has been an important technique in predictive analysis. The same technique can be incorporated in social network data to predict

unpublished sensitive information. Cai et al. [6] is the first in the literature to have modelled an inference attack using rule - based mining technique. It [6] has used a rough set approach to generate rules which in turn will help predict the unpublished sensitive information. Cai et al. [6] has also proposed a data sanitization technique as a solution against inference attack using a rule - based mining technique. The technique [6] uses perturbation and removal of information as its manipulation methods to achieve privacy against inference attack. However, there are two issues with the data sanitization technique [6]. First, the data sanitization technique does not remove the association between sensitive information and identification information in case of perturbation. Second, the data sanitization technique does not provide a strong privacy guarantee against inference attack due to rule - based mining. We analyze the existing literature in the next section.

6.2.1 Literature Review

Adversarial knowledge has always been the focal point of the evolution of the privacy - preserving data publishing field. Initially, modelling adversarial knowledge [9], [10], [11] was extensively studied for relational tables. Specifically, the adversarial knowledge modelled were personal information [8], factual information [5], correlational knowledge [9], probabilistic knowledge [10], [11]. In the current decade, the focus has been shifted from relational tables to social networks. Basically, the social network consists of information in terms of profile and structure information. Social network literature has several instances of modelling of adversarial knowledge [24], [27], [49], [63], [127], [128], [129]. However, in recent times, inference attack [6], [7], [52], [130], [131], [133] in social network been much discussed research challenge as it predicts the unpublished sensitive information. Cai et al. [6] modelled inference attack due to rule - based mining and suggested data sanitization technique as a solution. It is a challenging attack as it can even occur without the knowledge of any personal information. We thoroughly review the data sanitization technique [6] in the next section.

Review of Cai et al.'s [6] Data sanitization Technique

Cai et al. [6] proposed a collective inference attack that infers unpublished sensitive information by mining information in the social circles. The information used for mining is user - profiles and social relationships. The basic intuition for collective inference attack is that there are high chances of sharing sensitive information in social relations if they share more non - sensitive information. To model collective inference attack, rough set approach [51] is used to generate rules. They proposed a data sanitization technique to protect against inference attack.

Data Sanitization Technique

Cai et al. [6] proposed a data sanitization technique named collective method against inference attack. It is based on manipulating the Privacy Dependent Attributes (PDA) and Utility Dependent Attributes (UDA). The PDA and UDA are calculated using Rough Set Theory. The PDA are the set of attributes that help in generating rules to predict unpublished sensitive attributes. Similarly, UDA is the attributes that help in maintaining the usefulness of the social data. The steps of the collective method are as follows:

- If PDA and UDA are disjoint sets, then remove the PDA.
- If PDA and UDA are not disjoint sets then,
  - Calculate the Core set, which is the intersection of privacy and utility attributes.
  - Remove attributes present in the PDA set while not in the Core Set.
  - Perturb the attributes in the Core Set.

The Perturbation substitutes a more generalized value for categorical attributes. Perturbation maps numerical attribute with the ratio of the difference of the numeric attribute value and minimum value in the dataset for a particular attribute category to the range. The range is the ratio of the difference of the maximum and minimum value of the dataset to the level of generalization. Link anonymization technique is used when the application is not specified. It removes the most indistinguishable link of a user such that the variance of its probability of possible sensitive attributes does not exceed a threshold Δ'.

Weakness in [6]'s data sanitization technique

There is a weakness [151], [154] in the data sanitization technique [6], as it does not remove the association between sensitive and identification attributes in case of perturbation. We explain it with the following example.

Consider G is a social network graph which is captured in Table T [6]. The table T consists of only user id and attributes. It excludes the relationships in terms of generating rules. Table T consists of n records which are of two types namely the training records (records whose sensitive attributes are published) and testing records (records whose sensitive attributes are not published). The set of training records are represented as N_{tr} and the set of testing records are represented as N_t. An example is shown in Table 6.1 [6] where n is 6. The identification attribute is Music and Movies whereas, the sensitive attribute is Political view. Here, $N_{tr} = \{A, B, C, D\}$ whereas, $N_t = \{E, F\}$.

Id	Music	Movie	Political View
A	Taylor Swift	God's Not Dead	Conservative
B	Taylor Swift	God's Not Dead	Conservative
C	George Strait	Son of God	Liberal
D	George Strait	Son of God	Liberal
E	Taylor Swift	God's Not Dead	?
F	George Strait	Son of God	?

Table 6.1: A Social Data Table T

Here, privacy dependent and utility dependent attributes are Music and Movies (PDA and UDA are the same). As a result, Music and Movies will be perturbed. The perturbed values for Taylor Swift, George Strait, God's not dead, and Son of God are American singer - songwriter, American country music singer, drama / family and drama / history, respectively. Here, the rules can be generated as the association between the sensitive attributes and identification attributes are not removed by perturbation. Also, the knowledge of the domain of the attributes will help in mapping the perturbed values to the original attribute values as attributes are not removed. As a result, the set of rules are generated on N_{tr} are as follows R = {(George Strait, Son of God $\implies$ Liberal), (Taylor Swift, God's Not Dead $\implies$ Conservative)}. The predicted testing records is N_t as shown in Table 6.2. As a result, privacy is not preserved in the data sanitization method [6] as rules can still be generated.

Id	Music	Movie	Political View
E	Taylor Swift	God's Not Dead	Conservative
F	George Strait	Son of God	Liberal

Table 6.2: Predicted Testing Records in Social Data Table T

It is noted that a new privacy model [151] against inference attack is imminent and we believe that rule - based anonymization technique can play an important role to defend against this attack.

6.3 Modelling Inference Attack due to Rule - based Mining

6.3.1 Preliminaries

Definition 6.1. Social Network Graph: A social network graph is defined as $G(U, A, E)$, where U is a set of users, A is a set of attributes, and E is a set of edges.

Here, $U = \{U^s \cup U^{ns}\}$ where U^s is a set of users whose sensitive information is published and U^{ns} is a set of users whose sensitive information is not published. Attributes are constituted of identification attributes and sensitive attributes. As a result, $A = \{A^{id} \cup A^S\}$ where A^{id} is a set of identification attributes, and A^S is a set of sensitive attributes. Edges are bifurcated into User - User edge and User - Attribute edge. $E = \{E^{U-U} \cup E^{U-A}\}$ where E^{U-U} is a set of User - User edge which connects two users and E^{U-A} is a User - Attribute edge which connects the user to its attribute. The main objective is to predict unpublished sensitive attribute in the social graph. The technique used for prediction is rule - based mining. To achieve this, we need to capture the attribute contents in a tabular form. This has indeed motivated us to capture social network graph into a social data table. We define the social data table as follows:

Definition 6.2. Social Data Table: The Social data table is depicted as $T(ID, A, R)$ where ID is set of users, $A = QI \cup S$ is set of identification attributes QI and sensitive attributes S and R is a set of user - user edges.

The social data table captures the social network graph. We show that the conversion is lossless using a mapping function $\Gamma(G) = T$. The mapping of social network graph G to social data table T as follows: $\Gamma(U) = ID$ (Maps Users), $\Gamma(A) = A$ (Maps Attributes), $\Gamma(E^{U-U}) = R$(Maps User - User edges) and $\Gamma(E^{U-A}) = T(ID, A)$ (Maps User - Attribute edges). Similarly, T can be captured into G.

6.3.2 Rule Formulation

We formulate rules in terms of identification attributes and sensitive attributes as rules are the backbone of rule - based mining technique. A rule consists of two components: antecedent and consequent. Antecedent consists of if part while consequent consists of *then* part. The basic structure of a rule is $antecedent \implies consequent$. Before defining a rule, we define the foundations that are required for rule formulation. We capture the social data table into transactional data as $T(ID, A)$, where ID is a user identifier, and A is the attribute set. A is an at-

tribute set consisting of identification attributes QI and sensitive attributes S. Let $A = \{q_1, \ldots, q_k, s_1, \ldots, s_d\}$ where $k, d \geq 1$. Let $T \subseteq A$ be a set of n transactions $T = \{T_1, \ldots, T_n\}$. Each transaction T_i is associated with an identifier ID and y where $y \leq (k + d)$ is set of attributes. The main objective is to generate rules where QI set of identification attributes is antecedent, and S set of sensitive attributes are consequent. We define the structure of a rule as follows:

Definition 6.3. Structure of Rule: For sets QI and S, for any attribute $q_i \in QI$ and $s_j \in S$, a transaction T_i is a rule of the form $q_i \implies s_j$ if $q_i \subset A$, $s_j \subset A$ and $q_i \cap s_j = \phi$.

The social network data is incomplete and inaccurate in nature. As a result, rule generated from social network data needs to be quantified based on relation between q_i and s_j. We define metrics [52], [53] for rules as follows:

Definition 6.4. Support: For sets $QI \subset A$ and $S \subset A$, for any attribute $q_i \in QI$ and $s_j \in S$, Support sup is the frequency of the occurrence of rule $q_i \implies s_j$ in T.

$$sup(q_i \implies s_j) = n(q_i \cup s_j) \tag{6.1}$$

Support sup gives the frequency of the rule in table T. However, it does not provide any information about the strength of the rule as it does not quantify a strong relation between q_i and s_j. Due to incomplete social data, frequency alone can not be a metric to find the relation between q_i and s_j. Less accurate rules can lead to false predictions. Therefore, one needs to have high conviction in the rule such that the rule can be used to predict the unpublished sensitive attributes. Therefore, we define confidence as follows:

Definition 6.5. Confidence: For sets $QI \subset A$ and $S \subset A$, for any attribute $q_i \in QI$ and $s_j \in S$, Confidence $conf$ is the ratio of the frequency of a rule $q_i \implies s_j$ in T to the frequency of q_i in T.

$$conf(q_i \implies s_j) = P(s_j|q_i) = \frac{n(s_j \cup q_i)}{n(q_i)} \tag{6.2}$$

Confidence is the conditional probability of consequent s_j when antecedent q_i is given. Confidence gives a strong association in terms of q_i and s_j. We define the rule as follows:

Definition 6.6. Rule: For any attribute $q_i \in QI$ and $s_j \in S$, $q_i \implies s_j$ is a rule if conditions (1) and (2) are fulfilled:

1. $sup(q_i \implies s_j) \geq t_{sup}$
2. $conf(q_i \implies s_j) \geq t_{conf}$

Here, t_{sup} and t_{conf} are the threshold for support and confidence, respectively. The value t_{sup} can be $2 \leq t_{sup} \leq n$(n is the number of transactions in the table T). The value t_{conf} can be $0 \leq t_{conf} \leq 1$.

6.3.3 Modelling Adversary

We devise an adversarial model [151] for inference attack using rule based mining.

Adversary

In social networks, *data publisher*'s task is to publish anonymized social network such that the statistical information can be obtained, but the user (individual) privacy is preserved. On the other hand, *adversary*'s goal is to predict unpublished sensitive attributes in the anonymized social network. *Adversary* has access to published anonymized social network graph, user's basic information and rule

- based mining techniques. The adversary has the capability to generate rules based on a published social network graph.

Adversarial Model

We describe the capabilities accessible by the adversary to execute inference attack due to rule - based mining.

Mechanism: The adversary uses n anonymized records of the social graph into set of training records N_{tr} and set of testing records N_t. Here, $N_{tr} = \{n_{tr_1}, \ldots, n_{tr_{n_{tr}}}\}$ and $N_t = \{n_{t_1}, \ldots, n_{t_{n_t}}\}$. The sensitive information is published in n_{tr} training records while it is not published in n_t testing records. The set of training records N_{tr} act as an input to the Rule generator and output is a set of rules R. Any mining technique can be used for generation of the rules. The set of testing records N_t and Rules R act as an input to the predictor block. The output of the predictor block is n'_t. Here, n'_t is $P(N_t|R)$. It signifies the probability of the testing records when R is present. Figure 6.1 shows the diagram of mechanism.

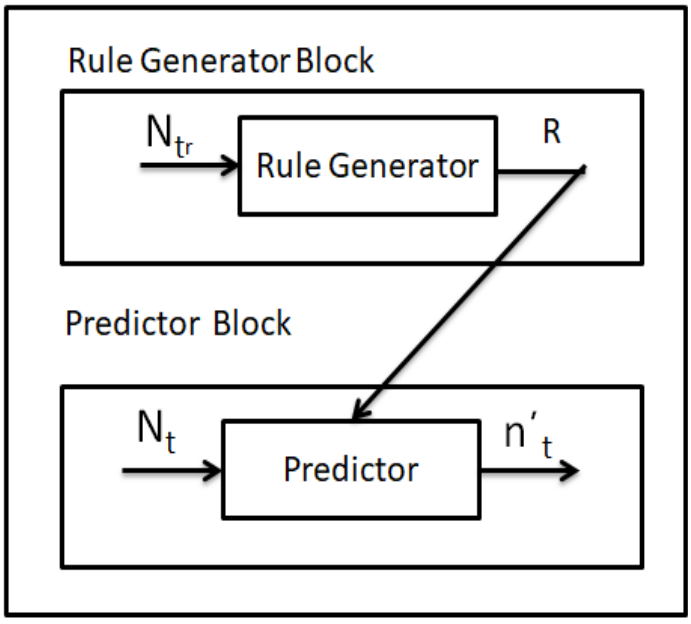

Figure 6.1: A Generic Rule - generator and Prediction model

Capability of Adversary Adv_R**:** The adversary Adv_R has access to n anonymized

transactions with a attributes. Here, $a = \{QI \cup S\}$, where QI is a set of identification attributes and S is a set of sensitive attributes. Identification attribute set is $q = \{q_1, \ldots, q_k\}$ and set of sensitive attribute is $s = \{s_1, \ldots, s_d\}$ where $k, d \geq 1$. The adversary Adv_R has access to the mechanism as described in Figure 6.1. It primarily consists of Rule Generator Block and Predictor Block. *Rule Generator block* takes as an input N_{tr} which consists of both identification attributes and sensitive attributes. It generates a set of rules $R = \{R_1, \ldots, R_{ru}\}$ where $ru \geq 1$. Here, we assume that the adversary has at least one rule corresponding to the testing records. The testing records N_t consists of identification attributes while sensitive attributes are unknown. *Predictor block* predicts sensitive attributes in n_t testing records using R. We define two blocks in the form of function as follows:

- $RG(N_{tr}) \longrightarrow R$: The function RG takes a set of training records N_{tr} as input, and the set of rules R is the output.
- $Pred(N_t, R) \longrightarrow n'_t$: The function $Pred$ takes a set of testing records N_t and a set of rules R as input. The output is probabilistic.

Apart from two functions, the adversary also has access to given functions. $\Gamma(G) \rightarrow T$: The function Γ captures the social network graph G to the social data table T. $Check(n'_t) \rightarrow s, f$: This function checks the probability $P(N_t|R)$ and compares it with $P(N_t)$, if the difference is negligible then privacy is preserved f and if the difference is non - negligible then privacy is disclosed s.

6.4 Rule Anonymity

We define privacy in terms of rule - based mining known as Rule Anonymity [151]:

Definition 6.7. Rule Anonymity: Given the social graph G and set of rules R, a predictor p and set of s distinct sensitive attribute values $s \in S$, the graph G is

rule anonymous if the difference of prediction probability of sensitive attribute S in the presence of R and without R is negligible.

$$|P_p^S(G|R) - P_p^S(G)| \leq \Delta \tag{6.3}$$

Here, Δ is negligible. A social network graph G is rule anonymous if it fulfils the condition in equation 6.3. In other words, for any given anonymized social network graph G, rules R have negligible impact on the sensitive attribute S, then G is rule anonymous. Rule Anonymity acts as a privacy guarantee in the social network against inference attack due to rule - based mining technique. If any given privacy - preserving technique fulfils equation 6.3, then rules have a negligible impact on the prediction of sensitive attribute due to rule - based mining technique. We now analyze data sanitization technique [6] with respect to rule - anonymity definition.

Theorem 6.1. *The prediction accuracy of a sensitive attribute in [6]'s data sanitization in social graph G is non - negligible against* Adv_R .

Proof. Let graph G be a data sanitized social graph. G consists of total of n_{tr} training records and n_t testing records. G consists of s distinct sensitive attribute values in the sensitive attribute S. The prediction accuracy of sensitive attribute S in n_t testing records when rules are not present is $P_p^S(G) = (\frac{1}{s})^{n_t}$. The Adv_R accesses the functions described in section 6.3.3 and applies it on G as follows:

- $\Gamma(G) \rightarrow T$: Step 1 captures the social graph G into the social data table T.
- $RG(N_{tr}) \rightarrow R$: Step 2 generates rules in N_{tr} and output is R. G does not give any privacy guarantee in terms of rule generation. It removes the most dependent privacy attributes and perturbs the privacy attributes that affect utility. Also, perturbation only replaces the old value with a perturbed one, but it doesn't eliminate the association between the attributes. Here, training

records don't contain any anonymization operations to remove the association, as a result, rules are generated. Therefore, $R \neq \phi$.

- $Pred(N_t, R) \rightarrow n'_t$: Step 3 calculates the prediction accuracy of sensitive attribute S in N_t when rules R is present. The probability when any testing record n_{t_i} is present in R is 1, otherwise the probability is $\frac{1}{s}$. The calculations are as follows:

$$
\begin{aligned}
P_p^S(G|R) &= P(n_{t_1}, n_{t_2}, \ldots, n_{t_{n_t}}|R) \\
&= P(n_{t_1}|R) \cdot P(n_{t_2}|R) \ldots P(n_{t_{n_t}}|R) \\
&= \prod_{n_{t_i} \in R} (1) \cdot \prod_{n_{t_i} \notin R} \left(\frac{1}{s}\right) \\
&= 1^r \cdot \left(\frac{1}{s}\right)^{n_t - r} \\
&= 1 \cdot \left(\frac{1}{s}\right)^{n_t - r} \\
&= \left(\frac{1}{s}\right)^{n_t - r}
\end{aligned}
$$

- $Check(n'_t) \rightarrow s$: Step 4 compares the prediction accuracy as follows:

$$
\begin{aligned}
\left| P_p^S(G|R) - P_p^S(G) \right| &= \left| \left(\frac{1}{s}\right)^{n_t - r} - \left(\frac{1}{s}\right)^{n_t} \right| \\
&= \left| \frac{\left(\frac{1}{s}\right)^{n_t}}{\left(\frac{1}{s}\right)^{r}} - \left(\frac{1}{s}\right)^{n_t} \right| \\
&= \left| \left(\frac{1}{s}\right)^{n_t} \left[\frac{1}{(\frac{1}{s})^r} - 1 \right] \right| \\
&= \left| \left(\frac{1}{s}\right)^{n_t} \left[\frac{1}{(\frac{1^r}{s^r})} - 1 \right] \right| \\
&= \left| \left(\frac{1}{s}\right)^{n_t} \left[\frac{1}{(\frac{1}{s^r})} - 1 \right] \right| \\
&= \left| \left(\frac{1}{s}\right)^{n_t} \left[s^r - 1 \right] \right| \\
&= \left| \left(\frac{1}{s^{n_t}}\right) \left[s^r - 1 \right] \right| \\
&= \left| \frac{s^r}{s^{n_t}} - \frac{1}{s^{n_t}} \right| \\
&= \left| \frac{s^r}{s^{n_t}} \right| \left(\because \frac{1}{s^{n_t}} << \frac{s^r}{s^{n_t}} \right)
\end{aligned}
$$

A function $f(n_t)$ is non - negligible [28], [29], if $\exists c \in N$ such that $\forall n_{t_0} \in N$, there is a $n_t \geq n_{t_0}$ such that $f(n_t) \geq n_t^{-c}$ Here, $f(n_t) = \frac{s^r}{s^{n_t}}$. We simplify $f(n_t)$ in terms of n_t as follows:

$$
\begin{aligned}
f(n_t) &= \frac{s^r}{s^{n_t}} \\
&= \frac{s^r}{\left(s^{log_s n_t}\right)^{\frac{n_t}{log_s n_t}}} \\
&= \frac{s^r}{n_t^{\frac{n_t}{log_s n_t}}}
\end{aligned}
$$

There exists $c = \frac{n_t}{log_s n_t}$ such that $n_t \geq 2$ ($n_{t_0} = 2$), $s \geq 2$ and $1 \leq r \leq$

$n_t - 1$, $\frac{s^r}{\frac{n_t}{n_t^{log_s n_t}}} \geq {n_t}^{-\frac{n_t}{log_s n_t}}$. As a result, $f(n_t) \geq {n_t}^{-c}$. Therefore, $f(n_t)$ is non - negligible. Hence, data sanitization technique is not rule anonymous as $|P_p^S(G|R) - P_p^S(G)|$ is non - negligible. Therefore, privacy is disclosed.

■

In the next section, we provide a new anonymization technique against inference attack using rule - based mining capturing rule anonymity principle.

6.5 Rule - based Anonymization

We propose a rule - based anonymization technique [151] against inference attack. The basic idea is to generate rules with less conviction so that unpublished sensitive attributes can not be predicted. Basically, the rule consists of two parameters, namely, confidence (t_{conf}) and support (t_{sup}). If rules generated have less conviction, then rules can't be useful for prediction. In other words, rules generated with low confidence and support parameters will be ineffective for predicting unpublished sensitive attributes. We achieve the above objective by adding spurious rules. Rule - based anonymization technique is divided into two phases: Phase 1 is the *Rule Generation* and, Phase 2 is the *Anonymization technique* against rule - based mining.

6.5.1 Phase 1: Rule Generation

Rule Generation is a procedure for generating rules using any rule mining techniques. The rules generated should abide by definition 6.6. The generated rules should have at least t_{sup} support and have at least t_{conf} confidence. First of all, it generates rules (R_I) by any method in the existing literature. After the rules are generated, it checks the support and confidence parameters in the rule R_I. Only those rules in R_I that abide by the thresholds are selected in R. The parameters of RULE_GENERATION Procedure is In which can take as input either social Network Graph G or Social data table T (any one of the two inputs can be used as

input) and thresholds of support (t_{sup}) and confidence (t_{conf}). The procedure returns a set of Rules R. Line 2 captures Social Network Graph G to Social Data Table T. Line 3 generates Rules R_I (intermediate rules) that follows any rule - based mining method RG. Lines 3 - 10 checks whether rules R_I have atleast t_{sup} support and t_{conf} confidence. If the condition is true, then add to the final Rule set R. Line 11 returns the set of rules R.

Algorithm 4 Rule_Generation Procedure

```
1:  procedure RULE_GENERATION(In, t_sup, t_conf)
2:      Capture Social Network Graph G to Social Network Data Table T.
3:      R_I = RG(ID,A).
4:      for each i in R_I do
5:          if sup(R_I_i) ≥ t_sup then
6:              if conf(R_I_i) ≥ t_conf then
7:                  Add R ← R_I_i
8:              end if
9:          end if
10:     end for
11:     return R
12: end procedure
```

6.5.2 Phase II: Anonymization Technique

The anonymization technique anonymizes the social graph G to protect against rule - based mining. First of all, rules are generated using the procedure RULE_GENERATION, which takes input as G or T and threshold t_{sup} and t_{conf}. Next, k% spurious rules are added in T. We determine the value of k as follows: Let, R_s be a set of rules which are generated after adding the spurious rules R' in social data table T'. It is depicted as $R_s \leftarrow$ RULE_GENERATION(T',t_{sup},t_{conf}). k is selected in such a way that anonymized social network graph G' is rule anonymous. Let δ be the number of spurious rules added to the social graph such that the below condition is satisfied.

$$|P_p^S(G'|R_s) - P_p^S(G')| \leq \Delta \tag{6.4}$$

Here, Δ is negligible. Therefore, $k\% = \delta$ as adding δ spurious rules guarantees rule anonymity.

The structure of the spurious rule is synonymous to a user in G'. The spurious user has a user id and attributes. A user has two relation edges, i.e. user to attribute relation and user to user relation. The user to attribute relation is the spurious rules as it consists of attributes. The problem is confined to rule - based mining considering attribute information. However, user to user relation of spurious users is randomly selected from a set of user - user edges (E^{U-U}).

Algorithm 5 Anonymization Technique

INPUT: Social Network Graph G, k, t_{sup}, t_{conf}.
OUTPUT: Anonymized Social Network Graph G'.

1: Capture Social Network Graph G to Social Network Data Table T.
2: R = RULE_GENERATION(T,t_{sup}, t_{conf}).
3: R' = k% spurious rules.
4: Add $T' \leftarrow R'$.
5: Recapture Anonymized Data Table T' to Anonymized Social Network Graph G'.

Line 1 captures the social network graph G to the social network table T. Line 2 calls the *RULE_GENERATION* procedure to generate rules. Line 3 extracts k% spurious rules. Spurious rules are added in such a way that rule anonymity is achieved. Line 4 adds R' to T'. Line 5 captures T' to G'.

6.5.3 Working Example

We explain the algorithm using an example. We take Table 6.1 consisting of n records. Here, $N_{tr} = \{A, B, C, D\}$ whereas $N_t = \{E, F\}$. Suppose $t_{sup} = s$ and $t_{conf} = t$. Suppose, there are r rules generated with respect to the given thresh-

olds. As a result, rule anonymity is not achieved. We add n spurious rules such that rule anonymity is achieved. Here, $n = \{n_1, n_2\}$ spurious rules are added such that rule anonymity is achieved for a given $t_{conf} = t$ and $t_{sup} = s$ as shown in Table 6.3. As a result, unpublished sensitive attribute in N_t is not predicted. Therefore, rule - based anonymization prevents inference attack due to rule - based mining.

Id	Music	Movie	Political View
A	Taylor Swift	God's Not Dead	Conservative
B	Taylor Swift	God's Not Dead	Conservative
C	George Strait	Son of God	Liberal
D	George Strait	Son of God	Liberal
E	Taylor Swift	God's Not Dead	?
F	George Strait	Son of God	?
G	n_1^1	n_1^2	n_1^3
H	n_2^1	n_2^2	n_2^3

Table 6.3: A Rule - Anonymized Social Data Table T'

6.6 Analysis of Rule - Based Anonymization

The proposed rule - based anonymization technique is analysed against an adversary Adv_R.

Theorem 6.2. *The prediction accuracy of sensitive attribute in Rule - based anonymized social Graph G' is negligible against Adv_R.*

Proof. Let Graph G' be a rule - based anonymized Graph. G' consists of total of $n_{tr'}$ training records and $n_{t'}$ testing records. G' consists of s distinct sensitive attribute values in sensitive attribute S. G' consists of k% i.e. δ spurious rules. The prediction accuracy of sensitive attribute S in $n_{t'}$ testing records when rules are not present is $P_p^S(G') = (\frac{1}{s})^{(n'_t+\delta)}$. Now, the Adv_R accesses the functions described in section 6.3.3 and applies it on G' as follows:

- $\Gamma(G') \rightarrow T'$: It captures the anonymized social graph G' to the anonymized social data table T'.

- $RG(N_{tr'}) \rightarrow R_s$: Step 2 generates rules using set of training records $N_{tr'}$ and output is R_s. Here, δ spurious rules are added in such a way that Rule Anonymity is guaranteed in a published social graph G'.

- $Pred(N_{t'}, R_s) \rightarrow n'_{t'}$ and $Check(n'_{t'}) \rightarrow f$: Step 3 predicts the sensitive attributes in testing records when rules R_s are present. But, as G' is rule anonymized; as a result, the prediction accuracy is negligible. Therefore, privacy is preserved.

■

6.7 Experiments and Results

We evaluate the rule - based anonymization technique on real social dataset to validate our claim of rule anonymity.

6.7.1 Dataset

We have used Facebook dataset [50] to evaluate our rule - based privacy anonymization technique. The Facebook dataset [50] consists of the attributes as well as social relations. We have taken 48 attributes as identification attributes where each attribute is an attribute value that belongs to a broader attribute category. Broadly, the attributes are education type, location, work with, language and gender. Here, gender is a sensitive attribute. Farahbaksh et al. [54] have cited that gender can be inferred with the help of location and employment information so, we have selected the above set of information. Here, each attribute value is $1/0$ that represents the presence/absence of the particular attribute in the user's attribute set. Table 6.4 gives the statistical information of the Facebook dataset. We have implemented rule - based anonymization in Python.

Information	Facebook
Number of Nodes	1045
Number of Edges	53498
Number of Attributes	48
Number of Sensitive Attributes	2
Average degree	51.19

Table 6.4: Statistical Information: Facebook Social Network DataSet [50]

6.7.2 Prerequisites

We use Rule Generator *RG* as association rule mining. We consider different combinations of confidence threshold and support threshold for rule generation. Our main goal is to achieve rule anonymity. We study the effect of confidence and support on the value of δ to attain rule anonymity. We have considered different confidence thresholds t_{conf} as 0.6, 0.75 and 1.0. Different support thresholds for the above - specified confidence thresholds are also being used. It helps in analyzing the impact of confidence and support on δ. We have not taken lower confidence and support threshold, as it would generate false rules.

6.7.3 Experimental Results

Figure 6.2, Figure 6.3 and Figure 6.4 shows the relation between δ, t_{sup} and t_{conf} when rule anonymity is being achieved. More spurious rules are required when confidence and support threshold are low to achieve rule anonymity. Figure 6.2 shows the relation between δ and t_{sup} for confidence threshold $t_{conf} = 0.60$. Lower is the support threshold, more is the value of δ to achieve rule anonymity. Figure 6.3 and Figure 6.4 also depicts the relation between δ and t_{sup} for confidence threshold $t_{conf} = 0.75$ and $t_{conf} = 1.0$ respectively. Here, δ decreases with the increase in support thresholds. The initial decrease in δ for $t_{conf} = 1$ is sudden as compared to $t_{conf} = 0.75$ and $t_{conf} = 0.60$. This depicts that for higher confidence threshold, δ shows sudden decrease when support is initially increased. However, all the three graphs show similar behaviour in terms of δ when maximum support thresholds are taken, i.e. 6 , 80 and 80 for confidence thresholds 1.0, 0.75 and 0.60 respectively. In summary, higher confidence and support thresholds will

achieve rule anonymity with lower δ. As a result, rule anonymity will be achieved with fewer spurious rules, which will indirectly help in preserving utility. Figure 6.5 provides a better visualization by diagrammatically blending Figure 6.2, Figure 6.3, and Figure 6.4.

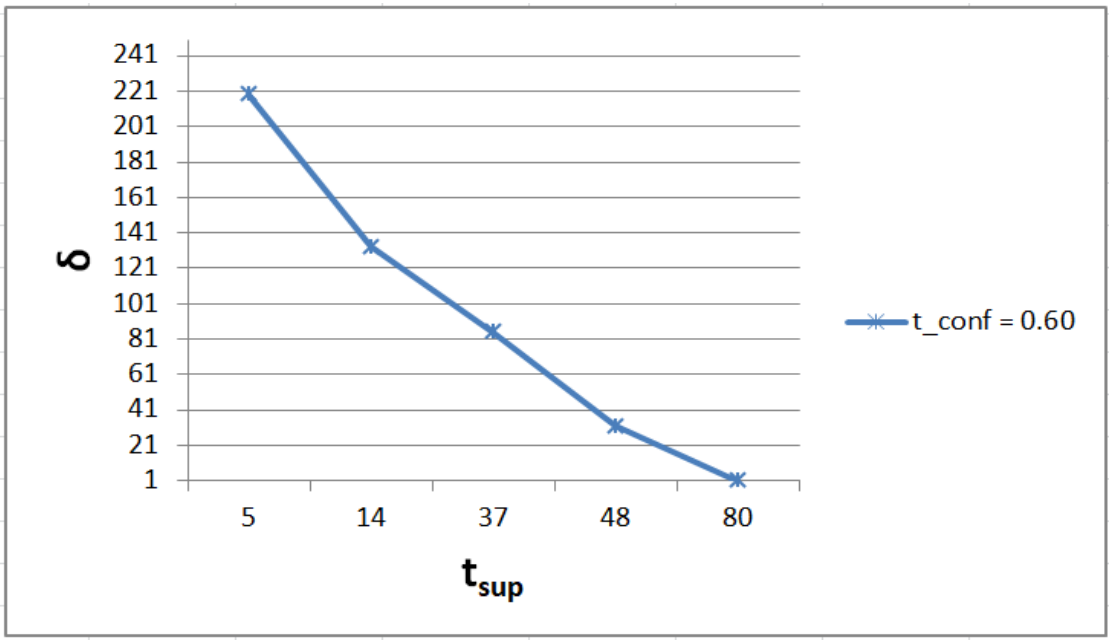

Figure 6.2: Effect Of δ on social dataset with $t_{conf} = 0.60$

Privacy Inference

We use social dataset [50] as in Table 6.4 to evaluate privacy inference. We compare our proposed technique with data sanitization [6] and naive anonymization (user names replaced with pseudonyms) in terms of privacy inference. Privacy inference occurs when the unpublished sensitive attribute is predicted. In other words, rules are generated for the given thresholds to predict unpublished sensitive attributes. As a result, we measure privacy inference in terms of the number of rules generated for a given threshold. Privacy decreases if the rules are successfully generated as the attacker can predict the sensitive attributes. Less is the number of rules generated, less are the chances of privacy inference in terms of prediction of sensitive attributes. We compare data sanitization technique [6] and naive anonymization (user names are replaced with pseudonyms)

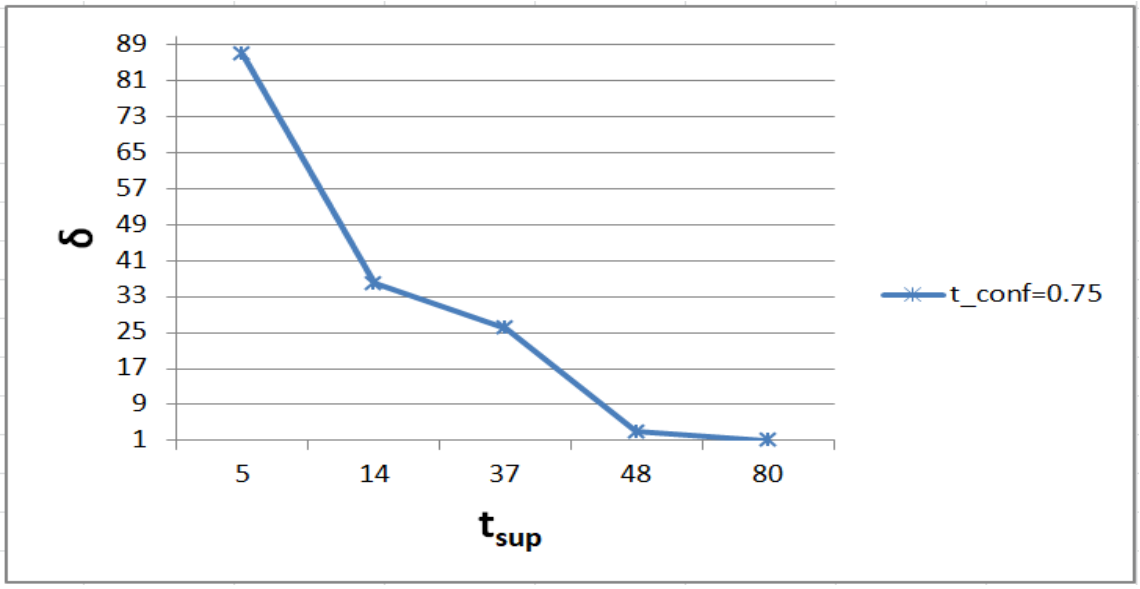

Figure 6.3: Effect Of δ on social dataset with $t_{conf} = 0.75$

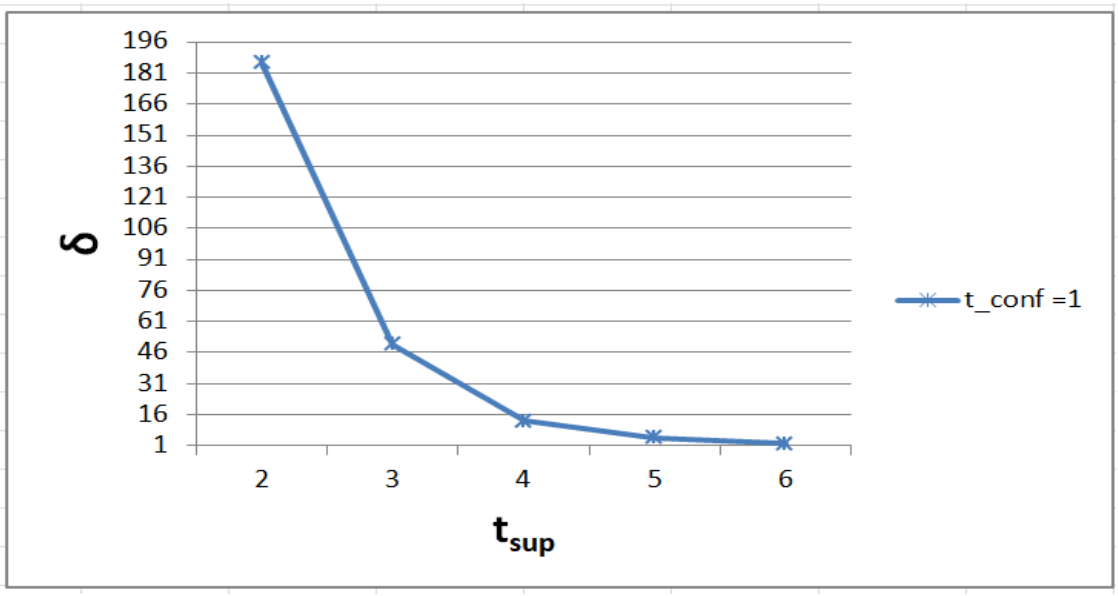

Figure 6.4: Effect Of δ on social dataset with $t_{conf} = 1.0$

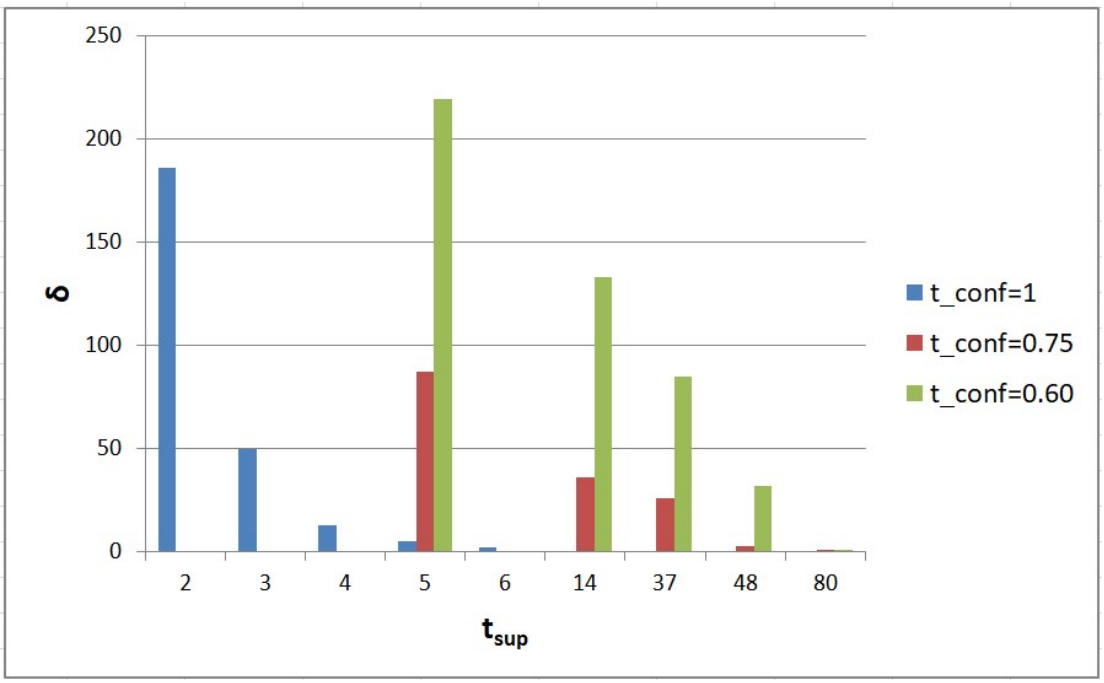

Figure 6.5: Effect Of δ on social dataset with t_{conf} and t_{sup}

with our existing proposed rule - based anonymization technique in terms of rules generation. Table 6.5 summarizes the privacy inference comparison. The data sanitization technique performs better as compared to naive anonymization in lower confidence thresholds. However, the data sanitization technique and naive anonymization show similar behaviour in presence of higher thresholds. Our proposed rule - based anonymization technique achieves rule anonymity where the rules generated are negligible (in our experiments, it is zero). As a result, the proposed rule based anonymization technique preserves privacy against rule - based mining as compared to the existing literature.

Technique	t_{conf}	t_{sup}	Rules
Data Sanitization	0.6	5	29
Data Sanitization	0.75	48	2
Data Sanitization	1.0	5	3
Naive Anonymization	0.6	5	95
Naive Anonymization	0.75	48	2
Naive Anonymization	1.0	5	5

Table 6.5: Privacy Inference

6.8 Conclusion

The Rule - based mining technique has been an effective method in predicting unpublished sensitive attributes in social networks. Inference attack is a serious privacy concern, as it can predict unpublished sensitive attributes using rule - based mining technique in social networks. We have devised an adversarial model against rule - based mining techniques. We have defined a strong privacy guarantee - Rule anonymity - that considers knowledge of rule based mining technique. We have discussed in detail about the data sanitization approaches in existing literature, which have been evaluated using the proposed adversarial model. Rule - based anonymization technique has been proposed that captures the rule anonymity principle. The proposed model is experimented using a social network dataset, and the experimental results have shown that the proposed model outperforms related approaches in terms of preserving privacy. The research work discussed in this chapter can be expanded further with the privacy versus utility aspect, which is left as the future scope of the proposed work.

CHAPTER 7

De - anonymization against Background Knowledge in Social Networks

Social networks data has multi - faceted opportunities in solving problems but suffers from privacy issues due to background knowledge. Moreover, the adversary has become more competitive and comprehensive in terms of its background knowledge capabilities. Modelling background knowledge in social networks can help understand the strength of an adversary and envisage stronger privacy protection techniques. In this chapter, we propose a de - anonymization technique against the adversary's background knowledge. We assume that the adversary has knowledge that is imprecise and semantically similar. The assumption of published social network considers users, attributes and structural information to be incomplete and inaccurate. The aggregate distance metrics suggested in the proposed technique considers imprecise attribute and structural information in to picture. In the proposed de - anonymization technique [152], termed as DeSAN, we consider the published graph to be anonymized such that users, relations and attribute information can be removed. The DeSAN uses distance for mapping between users. To the best of our knowledge, our proposed technique DeSAN is the first to consider semantic adversarial knowledge that collectively incorporates inaccurate and imprecise attribute and structure information while de - anonymizing. Less is the distance; more are the chances to be de - anonymized, which helps design more robust and comprehensive privacy models. We also discuss the proposed privacy - preserving technique against background knowledge. Experimental results of the proposed DeSAN have shown useful and encouraging

results in terms of de - anonymization accuracy.

7.1 Introduction

Humongous social network data gets collected by various networking applications in everyday life, and in particular, data consists of user's identification information, relations, interests and interactions. Network applications outsource data to third party applications, which provides insight about demographics, current trends, business decisions and can even track a pandemic [45]. As a result, social network data has become a significant contributor to society's social and economic growth. Contrary to the umpteen advantages, social network data is prone to privacy threat as it contains the user's sensitive information. Adversary compromises the privacy of social network users by using knowledge in terms of structural and identification information. Moreover, in the current instance, access to knowledge is much easier as it is freely available in the public domain. Therefore, modelling knowledge in terms of the adversary is essential and challenging research direction in the field of social network data publishing.

In recent times, de - anonymization [63] in social networks has gained immense attention in the research community. Generally, de - anonymization in social networks map the anonymized graph with the adversary's background knowledge graph in order to identify the user. Subsequently, user identification leads to attribute disclosure. Adversary's background knowledge graph can range from a social graph, other than an anonymized graph, to more specific user information. As a result, de - anonymization helps in analyzing and modelling the knowledge an adversary can possess.

7.2 De - Anonymization in Social Networks

Various de - anonymization [7], [46], [24], [47], [49], [6], [27], [63] techniques are presented and discussed extensively in the literature. In particular, seed - based and seed - free de - anonymization approaches are broadly explored. The seed - based de - anonymization [24], [63], [125] approach is initiated by identifying seeds (some users) followed by propagating the structure - based identification mechanism from the seeds to the social graph users. On the other side, the seed - free de - anonymization [123], [124] approach identifies users in the social graph based on structural information like degree. Apart from the above approaches, there are de - anonymization approaches [63] based on knowledge graphs [7], heterogeneous social networks [27], rule - based mining [6].

Typically, de - anonymization identifies users in published social network using the adversary's background knowledge. Adversary's background knowledge can range from published auxiliary social graphs to more specific individual information. As a result, de - anonymization techniques use structure and/or identification information to narrow down users in the published anonymized graphs. Different similarity measures [7], [27], [126], [26] have been proposed in the literature to attain the above objective. More is the similarity between the published graph and the adversary's background knowledge; more is the conviction to identify users. We discuss some existing literature in the next section.

7.2.1 Literature Review

Backstrom et al. [46] modelled two attacks, namely, active and passive. The active attack implants the dummy users and connects them with target users. Once the anonymized graph is published, the adversary looks for the generated subgraph into the anonymized social network. Once found, the adversary can manipulate the privacy of target users. It is assumed that the adversary has the knowledge

of a set of target users, whose privacy needs to be compromised. In the passive attack, the adversary searches itself in the published anonymized network by using the basic identification information. Once found, it manipulates the privacy of its neighbours. Here, knowledge is the adversary's identification information. Narayanan et al. [24] modelled a de - anonymization attack by mapping anonymized social graph with the adversary's background knowledge. The adversary's background knowledge in [24] is aggregate auxiliary information and individual auxiliary information. The aggregate auxiliary information is a social graph other than the anonymized graph, and auxiliary individual information is information about specific target users. In [47], the authors proposed a theoretical framework for de - anonymization. Many de - anonymization techniques [48], [3], [49] have been proposed in the literature that incorporates different methods for re - identification of users. Some well - known approaches are community based approach [48], using different similarity measures for graph matching [27] [24] [26].

With time, the adversary's manipulation capabilities have also been broadened in terms of knowledge collection and assumption. Li et al. [27] models a de - anonymization attack, where the adversary can aggregate user information from various heterogeneous social networks. Specifically, the de - anonymization attack uses structure and profile to match user on different social network platforms, which helps in collecting information about users that might be present on one social network while absent on the other social network. Subsequently, Qian et al. [7] moved a step further by modelling the inference attack in the social network. The work has proposed a two - step attack which consists of de - anonymization followed by privacy inference. The attack predicts the unpublished sensitive information of the user using the adversary's background knowledge. The work in [7] considers adversary's background knowledge as structural information, personal information, statistical information and factual information. They have considered correlations between different information.

In a nutshell, the assumption of the adversary's background knowledge has evolved in social network de - anonymization. Initially, the assumption about adversary's background knowledge was auxiliary social graph [46], [24]. Gradually, the adversary's background knowledge assumption in social network de - anonymization expanded significantly in terms of different knowledge [7], [6], [3] like individual information, correlation - based knowledge, structural knowledge, probabilistic knowledge.

However, knowledge related to semantic proximity [152] has not been discussed in the literature. We discuss semantic knowledge in the following section.

7.3 Semantic Knowledge

The knowledge related to semantic proximity between any information is referred to as semantic knowledge. The term semantic proximity between information considers or compares information that is semantically similar (which may not be syntactically similar). For example, if the adversary has information about a user's disease as a specific disease (e.g. bronchitis). The published social graph has information about the broader disease (e.g. respiratory disease), not a specific one. In that case, both the disease can be semantically compared but not syntactically. As a result, semantic knowledge will help disclose privacy where information is imprecise and inaccurate. In the context of social networks, information can be identification information, structural information and sensitive information. Moreover, social network data is inaccurate and imprecise. The adversary too, can have inaccurate knowledge. As a result, semantic knowledge can de - anonymize users in social networks against inaccurate and imprecise adversarial knowledge. Specifically, the adversary's inaccurate and imprecise knowledge considered are as follows:

- The adversary can have knowledge in the form of range rather than a spe-

cific value. For example, suppose the adversary knows zipcode of an individual in the form of a range, i.e., 13050 - 13065 instead of a specific one.

- The adversary can have knowledge that is syntactically dissimilar, but semantically similar. For example, the adversary has knowledge of the location of an individual in the form of the country rather than the city.
- Relations and identification information are distorted in the published social network graph. As a result, the social network data is inaccurate and incomplete.

This has motivated us to devise a de - anonymization technique [152] that can address the above issues. We discuss the proposed solution in the next section.

7.4 Preliminaries

7.4.1 Social Network

A social network is represented by $G(U, A, E)$ where U is a set of users, A is a set of attributes and E is a set of edges. Assume that the set of users U in social network is represented as $U = \{u_1, \ldots, u_n\}$, A is set of quasiidentifiers (QI) and sensitive attributes (SA) and is represented as $A = QI \cup SA$. Here, $QI = \{a_1, \ldots, a_j\}$ and $SA = \{a_{j+1}, \ldots, a_m\}$, E is set of user - user edges (E_{U-U}) and user - attribute edges (E_{U-A}) and is represented as $E = E_{U-U} \cup E_{U-A}$. E_{U-U} is set of edges that connects users while E_{U-A} is a set of edges that connects users with its attribute.

To extract knowledge from the available information, we need to model the social network in a tabular form. A social data table is represented by $T(U, A, E_{U-U})$, where U is a set of users, A is a set of attributes and E_{U-U} is a set of user - user edges.

7.4.2 Adversary's Background Knowledge

A publisher's goal is to publish social network in the public domain such that privacy of the user is preserved and usefulness of data is not distorted. To achieve the above goal, it uses anonymization techniques. The anonymized social graph is represented as $G'(U', A', E')$, where U' is a set of users (specifically user names are replaced with pseudonyms), A' is a set of attributes and E' is a set of edges. The attributes and edges can be anonymized by generalizing attributes, removing edges, adding edges and data sanitization. On the other hand, the adversary has the goal to disclose the privacy of the user or group of users. It achieves the above goal by extracting knowledge about the user along with manipulation capabilities. The knowledge is extracted from various sources like data aggregation, data crawling, data mining, published data, social profiles [7], [47], [49], [27], [1], [24]. The variants of knowledge that the adversary possesses are as follows:

- First, the adversary has access to any published social graph [24][47] other than the anonymized social graph. The published social graph can be a subset of the anonymized social graph.

- Second, the adversary has knowledge of personal information of the user or set of users like identification attribute information, sensitive attribute information, approximate sensitive attributes, approximate identification attributes (Note that, here, approximate attribute refers to an imprecise and semantically similar attribute).

- Third, the adversary has structural knowledge in terms of degree, ego network, neighbourhood network of the user [7].

- Fourth, different types of knowledge like correlational knowledge [7] [9], semantic knowledge, rule based knowledge [6] that can be helpful in predicting and inferring sensitive attribute information from the existing data (published graph, individual information).

As we have considered multiple variants of knowledge, we represent the adversary's background knowledge in the form of a table. It helps in extracting knowledge using manipulation capabilities due to its simple representation. The adversary's background knowledge table is represented by $T^k(U^k, A^k, E^k_{U-U})$, where U^k is set of users known to the adversary, A^k is a set of attributes known to the adversary and E^k_{U-U} is a set of edges known to adversary (As per definition 7.2). In T^k, each user and its corresponding attribute and edge information is represented in the form of a record. For example, suppose an adversary has knowledge of user u^k_1, its corresponding identification attributes named zipcode and age and connections with other users. It signifies as a record in adversary's background knowledge table T^k and is represented as $(u^k_1, (13054, 40-50), (u^k_2, u^k_3))$. Adversary's background knowledge can be represented as graph G^k.

7.4.3 Metrics

In this section, we propose metrics [152] used to compare semantically similar and imprecise knowledge. We consider two users, u^k and $u^{'}$, and calculate the distance $D(u^k, u^{'})$ between them by collectively considering attribute and structural distance.

Attribute Distance

Attributes are compared semantically as well as syntactically. Typically, there are two types of attributes in the social network data - numerical and categorical attributes. Example of numerical attributes are zipcode, age, salary etc., whereas categorical attributes are disease, gender etc.

- **Categorical Attribute Distance:** For any given two categorical attributes $a^k_i, a^{'}_i \in A$ and an ontology O_A, the distance between a^k_i and $a^{'}_i$ is as follows:

$$D^C_A(a^k_i, a^{'}_i) = \begin{cases} 0 & a^k_i = a^{'}_i \\ \frac{d_{O_A}(a^k_i, a^{'}_i)}{d_{max}} & a^k_i \neq a^{'}_i \end{cases} \tag{7.1}$$

Here, $d_{O_A}(a_i^k, a_i^{'})$ is the distance between two categorical attributes a_i^k and $a_i^{'}$ in ontology O_A. O_A is an ontology generated for a given domain of the attribute A. In this work, we do not focus on ontology generation. An example of an ontology is given in [31]. d_{max} is the maximum distance in the ontology O_A, where distance is the number of edges in this metric. If a_i^k and $a_i^{'}$ have syntactically the same values then the distance is 0; else, distance is calculated.

- **Numerical Attribute Distance:** The numerical attribute value can either be fixed or continuous. We represent continuous value as a range. For example, suppose a numeric attribute is a salary. The fixed value of salary is $40K$, whereas the continuous value is $40K - 80K$. These two values are different interpretations in terms of knowledge as the first one has more conviction than the latter one. We discuss different metrics for fixed as well as range values:

  Case 1: Fixed Numeric Attribute Value

  For any given two fixed numeric attributes $a_i^k, a_i^{'} \in A$, the distance between a_i^k and $a_i^{'}$ is as follows:

$$D_A^N(a_i^k, a_i^{'}) = \begin{cases} 0 & a_i^k = a_i^{'} \\ \frac{|a_i^k - a_i^{'}|}{|max_A - min_A|} & a_i^k \neq a_i^{'} \end{cases} \tag{7.2}$$

  Here, max_A and min_A is the maximum and minimum value of attribute A. If the two fixed attribute values are the same, then the distance is 0; else, distance is calculated using formula.

  Case 2: Numeric Attribute Value in terms of range

  For any given fixed numeric attribute $a_i^k \in A$ and numeric attribute in terms

of range $a_i^{'} \in A$, the distance between a_i^k and $a_i^{'}$ is as follows:

$$D_A^N(a_i^k, a_i^{'}) = \begin{cases} \frac{d(a_i^k, a_i^{'})}{(max_a - min_a) * L_G} & a_i^k \in a_i^{'} \\ 1 & a_i^k \notin a_i^{'} \end{cases} \tag{7.3}$$

Here, max_A and min_A are maximum and minimum value of attribute A respectively. $d(a_i^k, a_i^{'})$ is the difference of range of attribute $a_i^{'}$. L_G is the level of generalization that is present in the range. If $a_i^k \notin a_i^{'}$ then distance is 1; else, distance is calculated using formula.

Definition 7.1. Attribute Distance: For given users u^k, $u^{'}$ having j categorical attributes and m numerical attributes, the attribute distance between users u^k and $u^{'}$ are as follows:

$$D_A(u^k, u^{'}) = \frac{\Sigma_{i=1}^{j}(D_A^C(a_i^k, a_i^{'})) + \Sigma_{i=1}^{m}(D_A^N(a_i^k, a_i^{'}))}{j + m} \tag{7.4}$$

Structural Distance

The structural distance between users u^k and $u^{'}$ is measured by comparing edges in l hop structure of u^k and $u^{'}$. In some cases, a single - hop (1 hop) structure may not quantify the structural differences as edges can be removed or added randomly. As a result, we consider l hop structure for comparing structural differences. The structural distance is defined as follows:

Definition 7.2. Structural Distance: For given two users u^k, $u^{'}$ and its l hop structure, the structural distance is as follows:

$$D_S(u^k, u^{'}) = \frac{|n((E_{U-U}^k)_l) - n((E_{U-U}^{'})_l)|}{n((E_{U-U}^k)_l)} \tag{7.5}$$

Here, $n((E^k_{U-U})_l)$ and $n((E^{'}_{U-U})_l)$ are number of edges in l hop structure of user u^k and $u^{'}$, respectively.

Having defined the attribute and structural distance independently, we now define aggregate distance as follows:

Definition 7.3. Aggregate Distance D(u^k, $u^{'}$): For given users u^k and $u^{'}$, aggregate distance between u^k and $u^{'}$ is as follows:

$$D(u^k, u^{'}) = w_s(D_S(u^k, u^{'})) + w_a(D_A(u^k, u^{'})) \tag{7.6}$$

Here, w_a and w_s are the weights associated with attribute distance and structural distance, respectively where $0 \leq w_s, w_a \leq 1$.

7.5 DeSAN: De - anonymization Technique

We assume that users in graph G^k and G' do overlap i.e. $U^k \cap U^{'} \neq \phi$. We also assume that a large part of users of G^k are present in $G^{'}$. But, $U^k \not\subset U^{'}$ as $G^{'}$ is anonymized where users and relations can be added and/or removed randomly as per the privacy requirement. The main objective of de - anonymization is to map users U' in anonymized graph to users U^k in adversary's background knowledge graph correctly and accurately. We formulate the de - anonymization problem as follows.

Given an anonymized graph $G'(U', A', E')$ and an adversary's background knowledge graph $G^k(U^k, A^k, E^k)$, user $u'_i \in U'$ maps user $u^k_i \in U^k$ accurately if $D(u^k_i, u^{'}_i)$ is minimum. This notion can be extended to G' and G^k as follows:

$$\arg\min \Sigma^{k_s,n_s}_{i,j=1} D(u^k_i, u^{'}_j) \tag{7.7}$$

Here, k_s and n_s are the number of users in $u_i^k \in U^k$ of G^k and $u_j^{'} \in U'$ of G' respectively that are mapped uniquely (each user of U^k is mapped uniquely to user of $U^{'}$) such that each mapping of user u_i^k with $u_j^{'}$ has minimum distance. D is the function that calculate distance between users as per equation 7.6.

7.5.1 Overview

DeSAN [152] de - anonymizes users using imprecise and inaccurate adversary's background knowledge. In the proposed DeSAN technique, the main idea is to map users of the anonymized graph with users of the adversary's background knowledge graph, such that the distance between users is minimum. In other words, less is the distance between the adversary and anonymized user; more is the mapping conviction. The proposed DeSAN bifurcates into two phases: - 1) Initialization 2) Mapping. In the Initialization phase, users of the anonymized graph are clustered based on degree information present in the adversary's background knowledge graph. In the Mapping phase, the loosely clustered users (from the first phase) are mapped based on aggregate distance.

7.5.2 The Proposed Algorithm

Initialization

Initialization phase focuses on the initial mapping between users of anonymized graph and users of adversary's background knowledge graph. The criteria of initial mapping are the degree information of the adversary's background knowledge graph/table. A user - defined parameter named degree difference helps in the initial mapping. The parameter degree difference considers both exact degree information and varying degree information. For varying degree information, it takes degree information as exact degree$\pm\Delta$, where $\Delta \geq 1$. The parameter Δ is instrumental in capturing the imprecise structural information resulted due to anonymization. It will also help eliminate unnecessary comparisons and aggregate distance calculations between users. As a result, the initial mapping consists

of loosely mapped anonymized users and users in the adversary's background knowledge graph/table based on degree information.

The overview of Initialization_List Procedure is as follows: Line 2 - 4 extracts degree of set of users U^k. Line 5 - 10 generates set of degree Δd that incorporates degree difference Δ. Line 11 - 17 divides the users U' of anonymized table T' with respect to set of degrees Δd. Line 18 - 26 generates initialization list. In initialization list, each user u_i^k of U^k is linked with set of users $U_S^{'}$ based on set of degree Δd. Line 27 returns Initialization list IL to the main function.

Algorithm 6 Initialization_List Procedure

1: **procedure** INITIALIZATION_LIST(T', T^k, Δ)
2: **for** each i in U^k **do**
3: $d^k \leftarrow$ Add $n(E^k_{U-U})_i$
4: **end for**
5: **for** each i in d^k **do**
6: **for** each j in Δ **do**
7: $\Delta d \leftarrow d_i^k$
8: $\Delta d \leftarrow d_i^k \pm j$
9: **end for**
10: **end for**
11: **for** each i in Δd **do**
12: **for** each j in U' in T' **do**
13: **if** $n(E^{'}_{U-U})_j == \Delta d_i$ **then**
14: $d_i^{''} \leftarrow$ Add $u_j^{'}$
15: **end if**
16: **end for**
17: **end for**
18: $d^{'} \leftarrow$ Link Δd with $d^{''}$
19: **for** each i in U^k **do**
20: **for** each j in d' **do**
21: **if** $n(E^k_{(U-U)})_i == \Delta d_j$ **then**
22: $U_S^{'} \leftarrow$ Add $u_k^{'}$
23: **end if**
24: **end for**
25: **end for**
26: $IL \leftarrow$ Link U^k and $U_S^{'}$ with respect to Δd
27: **return** IL
28: **end procedure**

Mapping

Mapping phase focusses on mapping between users of anonymized graph and users of adversary's background knowledge graph present in the initial mapping (Initialization phase). Here, the mapping criteria are the aggregate distance that considers information of structure and attributes into the picture. The Initialization list is narrowed down to last list by calculating aggregate distance and mapping users U^k (in Adversary's background knowledge table/graph) with user $U_S^{'}$ (Anonymized users) having minimum aggregate distance. The aggregate distance should be less than δ. Here, δ is the upper bound of aggregate distance. Following the initial mapping, the mapping is propagated until all users are mapped.

The overview of algorithm 2 is as follows: The input of algorithm is anonymized social graph G', adversary's background knowledge graph G^k, degree difference Δ, and distance threshold δ. The output is set of de - anonymized users U^D. Line 1 captures anonymized graph G' to anonymized social data table T'. Line 2 captures adversary's background knowledge graph G^K to adversary's background knowledge table T^k. Line 3 calls procedure Initialization_List. Line 4 - 11 generates the Last List LL consist of the distance between U^k with $u_L^{'}$ in IL is minimum and doesn't exceed distance threshold δ. Line 12 adds LL to U^D.

7.6 Privacy - Preserving Technique against Background Knowledge in Social Networks

Various privacy - preserving techniques [39][56][57][58][59][60][61][62][63][64] have been proposed and widely discussed in the literature. Here, adversary considers different variants of knowledge like correlation knowledge [7][64], rule - based mining technique [6] and probabilistic knowledge [64] in social networks. However, previous literature has not considered all variants of knowledge collectively. Also, semantic knowledge is available with the adversary (as discussed in section

Algorithm 7 De - anonymization Technique

INPUT: Anonymized Social Graph G', Adversary's background knowledge table T^k / Adversary's background knowledge Graph G^k, Δ, δ.

OUTPUT: Set of De - anonymized Users U^D

1: Capture Anonymized Social Graph G' to Anonymized Social Data Table T'.
2: Capture Adversary's background knowledge graph G^k to Adversary's background Knowledge Table T^k.
3: $IL \leftarrow Initialization\_List(T', T^k, \Delta)$
4: **for** each u_i^k in IL **do**
5: **for** each $u_j^{'}$ in $U_S^{'}$ **do**
6: **if** $D(u_i^k, u_j^{'}) \leq \delta$ **then**
7: $D_I \leftarrow$ Add $D(u_i^k, u_j^{'})$
8: **end if**
9: **end for**
10: **end for**
11: $LL \leftarrow$ Link U^K and $U_L^{'}$ where $U_L^{'} \leftarrow$ user with $min(D_I)$
12: $U^D \leftarrow LL$.

7.3). As a result, the background knowledge assumption should be more comprehensive instead of selective adversarial assumption. This has motivated us to devise a privacy - preserving technique that considers comprehensive background knowledge (as discussed in section 7.4.2).

7.6.1 The Proposed Privacy - Preserving Technique

The proposed privacy - preserving technique preserves privacy against the adversary's background knowledge. It adds spurious records such that the adversary's background knowledge has negligible impact on the anonymized social graph.

Firstly, the privacy - preserving technique generates knowledge using knowledge generator K_G. Knowledge Generator K_G takes as an input a social data table T and information I. It generates a set of knowledge K. Knowledge Generator can be any function that matches the information I with the identification attributes, sensitive attributes and structural attributes present in the social data table T. The set of knowledge K is filtered out into set of adversary's background knowledge A_K such that each k_i exceeds knowledge threshold t_k (Here, t_k is set

by the data publisher). Lastly, $A_K^{'}$ spurious records are added to preserve privacy. $A_K^{'}$ spurious records are added such that anonymized social graph G' fulfils $|P(G'|A_K) - P(G')| \leq \Delta$. Here, Δ is negligible.

Algorithm 8 Privacy - Preserving Technique against Background Knowledge

INPUT: Social Graph G, Information I, knowledge threshold t_k.

OUTPUT: Anonymized Social Graph G'

1: Capture Social Graph G to Social data Table T.
2: $K \leftarrow K_G(T, I)$
3: **for** each i in K **do**
4: **if** $P(K_i) \geq t_k$ **then**
5: $A_K \leftarrow$ Add k_i
6: **end if**
7: **end for**
8: $T' \leftarrow$ Add A'_K spurious records
9: Capture Anonymized Social Data Table T' to Anonymized Social Graph G'.

The working of algorithm is as follows: The input of algorithm is a Social graph G, Information I and knowledge threshold t_k. The output of algorithm is an anonymized social graph G'. Line 1 captures the social graph G to social data table T. Line 2 generates set of knowledge K using the knowledge generator K_G. Line 3 - 7 filters Knowledge set K where each knowledge exceeds knowledge threshold t_k. Line 8 adds spurious records. Line 9 recaptures anonymized social data table T' to anonymized social graph G'.

7.6.2 Analysis of the proposed privacy - preserved technique

Adversarial Model

In social networks, data publisher and adversary are two important entities in an adversarial model. The role of data publisher in social networks is to anonymize data such that user(s) privacy is preserved while usefulness is not distorted. On

the contrary, the adversary's role is to disclose the privacy of user/s using its manipulation capabilities. We explain the mechanism of the adversarial model as follows:

Mechanism: Consider an anonymized social graph G' which consists of n unique users. Each user has its attributes A' (anonymized) and edges E' (anonymized). Here, A' consists of identification attributes and sensitive attributes whereas E' consists of relation and attribute edges. The adversarial model consists of two blocks i.e. Knowledge Generator block (K_Gen) and Priv ($Priv$) block. Knowledge Generator block generates adversary's background knowledge. The input to the Knowledge Generator block (K_Gen) block is the anonymized social graph G' and Information I. The output of the block is Adversary's Background Knowledge A_K. Here, $A_K = \{k_1, \ldots, k_m\}$, where $m \geq 1$. The Priv block ($Priv$) links the user(s) with their sensitive attribute using Adversary's Background Knowledge A_K. The output of Priv block is $P(G'|A_K)$. $P(G'|A_K)$ is the probability of linking the user with their sensitive attribute in an anonymized social graph G' when Adversary's Background Knowledge A_K is present.

Capability of Adversary: The adversary has access to the anonymized social graph G'. The adversary has access to adversary's background knowledge as discussed in section 7.4.2.

- $\delta(G') \rightarrow T'$: The function δ takes input as an anonymized social graph G' and the output is anonymized social data table T'.
- $K\_Gen(T', I) \rightarrow A_K$: The function K_Gen takes as an input an anonymized social data table T' and Information I, whereas the output is Adversary's Background Knowledge A_K.
- $Priv(T', A_K) \rightarrow P$: The function $Priv$ takes as an input an anonymized social data table T' and Adversary's Background Knowledge A_K, whereas the output is probability Pr.

- $Result(Pr) \rightarrow s, u$: The function $Result$ compares the probability $P(G'|A_K)$ with $P(G')$. If the difference is non - negligible output is s, otherwise u.

Privacy Disclosure in Social Network: Adversary's background knowledge A_K has the potential to disclose the sensitive attribute of the user(s) in the anonymized social graph G'. We define privacy disclosure concerning adversarial background knowledge A_K in social networks as follows:

Definition 7.4. Privacy against Adversary's Background Knowledge: For a given anonymized social graph G' and adversary's background knowledge A_K, G' preserves privacy against adversary's background knowledge A_K if the difference between the probability of linking user(s) with its sensitive attribute in G' in the presence of A_K and absence of A_K is negligible.

$$|P(G'|A_K) - P(G')| \leq \Delta \tag{7.8}$$

Here, Δ is negligible. A privacy - preserving technique preserves privacy against adversary's background knowledge if it fulfils equation 7.8. This acts as a privacy guarantee against adversary's background knowledge A_K in social networks.

Theorem 7.1. *Proposed Privacy - Preserving technique preserves privacy against* Adv_{A_K}.

Proof. Let G' be an anonymized social graph as per section 7.4.1. G' consists of n' users. Here, $A_K^{'}$ spurious records are added in the anonymized social graph G'. The initial probability of linking the user(s) with their sensitive attribute in the anonymized social graph G' when adversary's background knowledge is not present is $P(G') = \left(\frac{1}{n'}\right)^{n'}$, where $n' = n + A_K^{'}$. The adversary Adv_{A_k} has access to the functions described in the adversarial model, and it applies on G' as follows:

- $\delta(G') \rightarrow T'$: This function takes as an input G' and captures social data table T' as an output.

- $K\_Gen(T', I) \rightarrow A_{Ks}$: The function *K_Gen* takes as an input T' and Information I. The output is A_{K_s}.

- $Priv(T', A_{K_s}) \rightarrow Pr$: The *Priv* function takes as an input T' and A_{K_s}, whereas the output is Pr i.e. $P(G'|A_{K_s})$. We have added spurious records in T' such that adversary's background knowledge has negligible impact on T'. As a result, $P(G'|A_{K_s}) = \left(\frac{1}{n'-k}\right)^{n'-k} \approx \left(\frac{1}{n'}\right)^{n'}$. Here, k is the number of users linked to their sensitive attribute using adversary's background knowledge, and it is negligible as spurious records are added.

- $Result(Pr) \rightarrow u$: The function *Result* compares the probability $P(G'|A_{K_s})$ with $P(G')$ and output is u as $|P(G'|A_{K_s}) - P(G')|$ is negligible. As a result, the adversary is unsuccessful in disclosing privacy.

Therefore, the proposed privacy - preserving technique preserves privacy against Adv_{A_K}. ■

7.7 Evaluation and Experimental Results

7.7.1 Dataset

The Facebook [50] dataset has a high resemblance to the social network scenario, is freely available and extensively used in the de - anonymization literature [49][6]. We use Facebook dataset [50] for the evaluation of the proposed DeSAN technique. The Facebook dataset consists of user identification information as well as its social relation. Broadly, we have considered attributes as education degree, hometown, language, location, work employee, work position and gender. Here, gender is a sensitive attribute, whereas remaining are the identification attributes. In Facebook [50] dataset, the value of each attribute is $1/0$, which signifies the presence/absence of that particular attribute. Statistics of the dataset is shown in Table 7.1. We have implemented the de - anonymization technique in python.

Information	Facebook
Number of Nodes	1045
Number of Edges	53498
Number of Attributes	153
Number of Sensitive Attributes	2
Average degree	51.19

Table 7.1: Statistical Information: Facebook Social Network Data Set

7.7.2 Prerequisites

Before evaluating the de - anonymization technique, we generate the adversary's background knowledge graph/table and anonymized dataset. For the adversary's background knowledge graph/table, we randomly select a set of users from the dataset as this will help in calculating the correctness of the technique. We assume adversaries have imprecise and inaccurate knowledge; thus, we have taken incomplete adversary's background knowledge by randomly removing information. For anonymization, we have considered two anonymization techniques. First is naive anonymization, where user names are replaced with pseudonyms. Second is, randomly removing social links, users and attribute information. We apply both the above techniques to the Facebook dataset. We name dataset $D1$, which incorporates the first anonymization technique while dataset $D2$ incorporates the second one. We measure the accuracy of the DeSAN by $\frac{n(U_m^k)}{n(U^k)}$, where $n(U_m^k)$ is the set of users that are successfully matched and $n(U^k)$ is the total number of users in the adversary's background knowledge graph/table. We have considered different values of w_a, w_s and Δ to demonstrate its effect on the dataset $D1$ and $D2$. The DeSAN is also compared against $D1$ and $D2$, where the same adversarial information is used for de - anonymization against $D1$ and $D2$.

7.7.3 Experimental Results

We evaluate the effect of w_a, w_s and Δ on $D1$ and $D2$. In other words, it shows the impact of structural and attribute information present in the adversary's background knowledge graph/table to de - anonymize users in $D1$ and $D2$. Figure.

7.1 , Figure 7.2 and Figure 7.3 shows the accuracy in de - anonymizing users in the datasets $D1$ and $D2$ with different combinations of w_a and w_s when degree difference Δ is 1, 2 and 3, respectively.

In Figure 7.1, for the dataset $D1$, either the structural or attribute information is required to de - anonymize users. In Figure 7.1, for the dataset $D2$, both structural and attribute information are required to de - anonymize users. Note that, here, all users will not get de - anonymized as the adversary's background knowledge is incomplete.

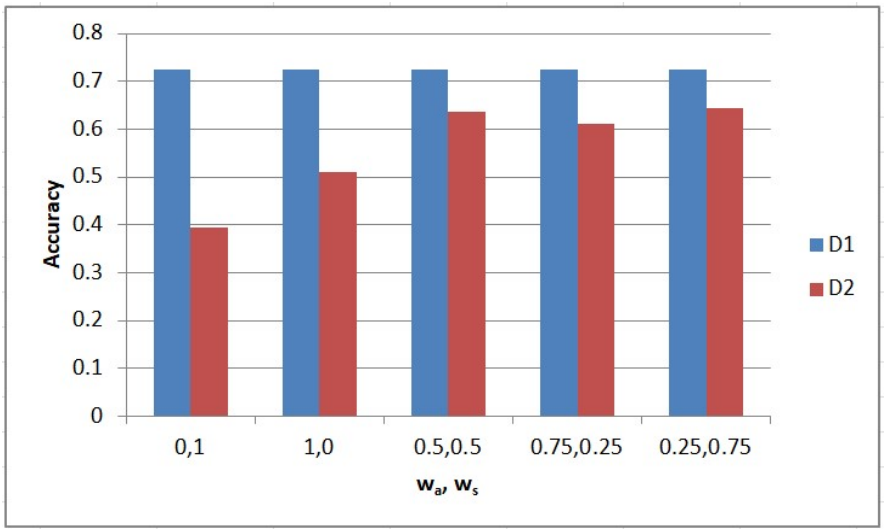

Figure 7.1: Effect of w_a and w_s on accuracy of $D1$ and $D2$ when $\Delta = 1$.

In Figure 7.2 and Figure 7.3, in the dataset $D1$, attribute information has a higher impact on de - anonymizing users. In Figure 7.2 and Figure 7.3, in the dataset $D2$, both structural and attribute information are required to de - anonymize users. Note that, here too, all users will not get de - anonymized as the adversary's background knowledge is incomplete. In a nutshell, structural and attribute information are collectively required for higher de - anonymization accuracy, where the adversary's background knowledge is incomplete. Note that, here, higher de - anonymization accuracy means more users are de - anonymized.

We evaluate the DeSAN based on de - anonymization accuracy. More is the de

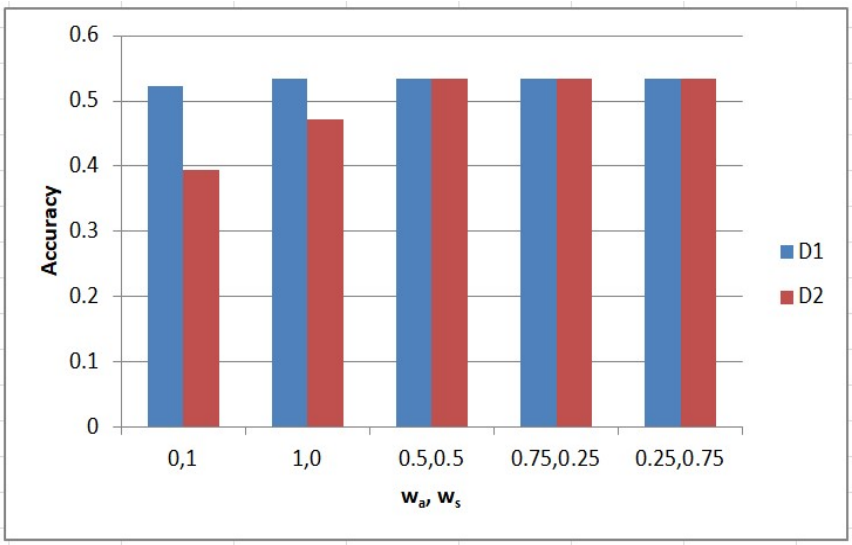

Figure 7.2: Effect of w_a and w_s on accuracy of $D1$ and $D2$ when $\Delta = 2$.

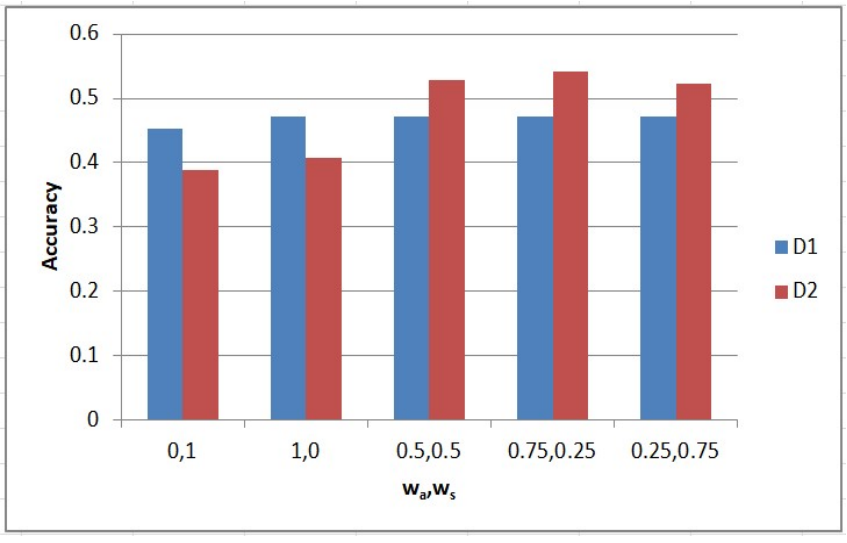

Figure 7.3: Effect of w_a and w_s on accuracy of $D1$ and $D2$ when $\Delta = 3$.

- anonymization accuracy; more are the users whose privacy discloses. We take the values of $w_a = 0.5$ and $w_s = 0.5$, as both structural and attribute information collectively gives better de - anonymization accuracy. In Figure 7.4, the DeSAN on the dataset $D2$ performs better than $D1$ when Δ increases. Here, $D2$ contains less accurate information as compared to $D1$. If the Δ increases, more users will be considered and will capture the inaccurate data, which gives high de - anonymization accuracy. As a result, DeSAN gives encouraging results in the distorted dataset against incomplete adversary's background knowledge in terms of de - anonymization accuracy.

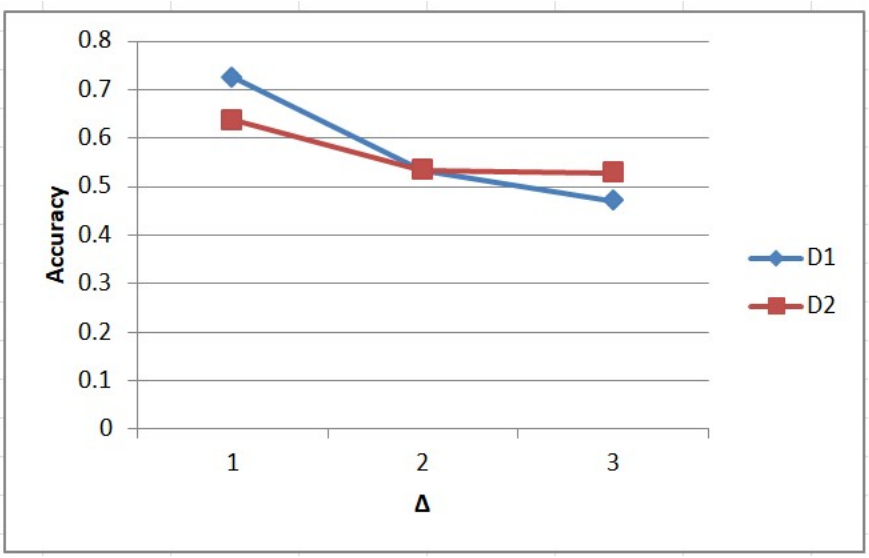

Figure 7.4: De - anonymization accuracy of $D1$ and $D2$ when $w_a = 0.5$ and $w_s = 0.5$.

7.8 Conclusion

Social network de - anonymization has been seen as a prominent demand in many areas that allows users and/or service providers to make the best out of this powerful tool while preserving users' privacy. We proposed a de - anonymization technique, DeSAN, that is able to identify users in a distorted graph. The proposed DeSAN technique assumes a strong adversarial model in which the adversary has comprehensive knowledge. We have also come up with a privacy -

preserving technique against background knowledge. We have implemented the DeSAN on a real dataset, and the experimental results show a promising outcome in terms of de - anonymization accuracy. Future scope of the proposed scheme is to come up with a privacy-preserved scalable clustering algorithm that can defeat background knowledge adversaries.

CHAPTER 8

Conclusion and Future Scope

8.1 Conclusion

Data privacy is important as disclosure of sensitive data can have a detrimental effect on the reputation and trust of the individual(s). Sensitive data can be disclosed due to the adversary's access to tons of information as well as its competent manipulation capabilities. As a consequence, the privacy of the individual(s) gets threatened by the adversary's background knowledge. This makes background knowledge an important concern and a research challenge that needs to be addressed.

Privacy - Preserving Data Publishing domain addresses data privacy concerns when publishing data. Prominent data privacy solutions like k - anonymity, l - diversity, t - closeness are discussed in the Privacy - Preserving Data Publishing literature. However, background knowledge has been instrumental for privacy attacks in the existing privacy solutions in the Privacy - Preserving Data Publishing domain. Certainly, background knowledge anticipates an imperative threat to data privacy owing to its extensibility and accessibility to the diverse knowledge variants.

Our research work focuses on designing and analyzing privacy solutions against background knowledge in the Privacy - Preserving Data Publishing domain. We study the existing privacy models in the literature and highlight their strengths

and weaknesses. Our study emphasizes that background knowledge needs to be studied and analyzed from a broader perspective in the Privacy - Preserving Data Publishing domain. Moreover, more comprehensive and strong data privacy solutions need to be designed against background knowledge. Our research work proposes three data privacy models against the background knowledge.

Our first work focuses on modelling and designing privacy solutions for published dataset against background knowledge in the Privacy - Preserving Data Publishing domain. We studied background knowledge in Privacy - Preserving Data Publishing domain and observed that background knowledge has different interpretations for different privacy models and needs to be modelled. Further, our research work defined background knowledge as a set of diverse knowledge variants and proposed an adversarial model against background knowledge. The proposed adversarial model has been evaluated against standard privacy approaches like k - anonymity, l - diversity and t - closeness. This work has motivated a need to design a strong privacy notion that considers comprehensive and realistic background knowledge assumptions. We come up with a strong privacy notion named $(\theta, [lb, ub]^{+sp}, \alpha)$ Privacy against background knowledge that is comprehensive. The proposed privacy model $(\theta, [lb, ub]^{+sp}, \alpha)$ Privacy protects against an adversary who has comprehensive background knowledge. The experimental results of the proposed model outperform its privacy strength in comparison to related approaches.

In the last decade, social networks have grown tremendously, which generates massive user data. Social Networks is a prominent application of Privacy - Preserving Data Publishing where data privacy is a cause of concern. Our second work focuses on designing privacy solutions for inference attack using rule - based mining techniques in social networks. We proposed an adversarial model against rule - based mining capability and analyzed it against the existing literature. Further, we proposed a privacy model named Rule - Anonymity against rule - based mining techniques. The proposed privacy model is incorporated in

an anonymization technique named the Rule - Based Anonymization technique. The proposed technique is analyzed against an adversary with rule - based mining capabilities. The experimental results demonstrate a significant decrease in prediction accuracy against existing literature.

Our third work focuses on modelling comprehensive background knowledge in social networks. Besides, we observe that semantic knowledge in social networks can help de - anonymize users against inaccurate and imprecise knowledge. We have assumed a strong adversary that has access to comprehensive background knowledge which also considers inaccurate and imprecise knowledge. To capture the semantic knowledge, we have proposed a distance metric named aggregate distance that considers attribute as well as structural properties. Our proposed de - anonymization technique DeSAN uses distance metric for mapping users in a distorted graph. The experimental results show encouraging outcomes in terms of de - anonymization accuracy against the adversary having inaccurate and imprecise background knowledge. Further, a privacy - preserving technique is proposed to address the disclosure.

To summarize the thesis, our research work addresses data privacy issues resulting due to background knowledge in a larger prospect in the privacy-preserving data publishing domain. Our research work proposed stringent privacy solutions against an adversary with powerful and realistic capabilities. The algorithmic solutions have incorporated the privacy definitions. Moreover, the experiments were performed on real data sets, which are extensively used in the privacy - preserving data publishing domain and are freely accessible. The experimental results have imbibed more confidence in the proposed privacy solutions with encouraging results. They are supported by theoretical analysis against stronger adversarial assumptions.

8.2 Future Scope

In the era of digitization, data being the focal point is vulnerable to various privacy attacks. To overcome attacks, data privacy needs to be incorporated in applications where sensitive data is published. Our work can be incorporated in a wide variety of applications as it considers realistic and comprehensive adversarial background knowledge.

Precisely, the future scope of the proposed work is as follows:

- The proposed works give stringent data privacy solutions to protect the individual's privacy against extensive and exhaustive information accessible by the adversary. Nevertheless, different applications have different privacy requirements for publishing data, and not all applications need to incorporate stringent data privacy solutions to protect data. The utility of data is equally necessary while protecting data against comprehensive background knowledge. The proposed work aims to be extended in privacy - utility trade - offs to provide practical and application - specific data privacy solutions.

- The adversary has become powerful and resourceful due to its access to background knowledge. Moreover, the data publisher needs to consider the adversary with its fullest capabilities as information is freely available and accessible. Rigorous mathematical definitions will imbibe the trust and confidence of individuals, and at the same time, help data publishers protect the data while publishing. The proposed data privacy solution against background knowledge can be extended in the essence of a stronger mathematical definition similar to differential privacy.

References

[1] L. Sweeney. *k*-anonymity: a model for protecting privacy. In International Journal on Uncertainty Fuzziness and Knowledge-based Systems, 10(5), pp. 557 – 570, 2002.

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

[3] N. Li, T Li, and S. Venkatasubramanian. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity In Proceedings of IEEE International Conference on Data Engineering, pp. 106 – 115, 2007.

[4] Knowledge. https://en.wikipedia.org/wiki/Knowledge

[5] T. Li, and N. Li. Injector: Mining Background Knowledge for Data Anonymization. In Proceedings of IEEE International Conference on Data Engineering, pp. 446 – 455, 2008.

[6] Z. Cai, Z. He, X. Guan, and Y. Li. Collective Data-Sanitization for Preventing Sensitive Information Inference Attacks in Social Networks. In IEEE Transactions on Dependable And Secure Computing 15(4), 2016.

[7] J. Qian, X. Li, C. Zhang, L. Chen, T. Jung, and J. Han. In Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model. In IEEE Transactions on Dependable and Secure Computing, 2017.

[8] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. Worst-case background knowledge in privacy. In Proceedings of IEEE International Conference on Data Engineering, pp. 126 – 135, 2007.

[9] T. Li, N. Li, and J. Zhang. Modeling and Integrating Background Knowledge in Data Anonymization. In Proceedings of IEEE International Conference on Data Engineering, pp. 6 – 17, 2009.

[10] B. Chen, K. LeFevre, and R. Ramakrishnan. Privacy skyline: privacy with multidimensional adversarial knowledge. In Proceedings of International Conference on Very Large Data Bases, pp. 770 – 781, 2007.

[11] W. Du, Z. Teng, and Z. Zhu. Privacy-maxent: integrating background knowledge in privacy quantification. In ACM SIGMOD International Conference on Management of data, pp. 459 – 472, 2008.

[12] M. E. Nergiz, M. Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In Proceedings of the ACM SIGMOD international conference on Management of data, pp. 665 – 676, 2007.

[13] 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. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and data Mining, pp. 754 – 759, 2006.

[14] T. M. Truta, and B. Vinay. Privacy Protection: p-Sensitive k-Anonymity Property. In International Conference on Data Engineering Workshops, pp. 94 – 94, 2006.

[15] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate Query Answering on Anonymized Tables. In International Conference on Data Engineering, pp. 116 – 125, 2007.

[16] R. Wong, A. Fu, K. Wang, and J. Pei. Minimality Attack in Privacy Preserving Data Publishing. In International Conference on Very Large Data Bases, pp. 543 – 554, 2007.

[17] K. Wang, and B. C. M. Fung. Anonymizing sequential releases. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 414 – 423, 2006.

[18] J. Li, Y. Tao, and X. Xiao. Preservation of proximity privacy in publishing numerical sensitive data. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 473 – 486, 2008.

[19] K. Wang, Y. Xu, A. W. C. Fu, and R. C. W. Wong. FF-Anonymity: When Quasi-identifiers Are Missing. In IEEE 25th International Conference on Data Engineering, pp. 1136 – 1139, 2009.

[20] X. Xiao, and Y. Tao. *m*-invariance: Towards privacy preserving re-publication of dynamic datasets. In ACM SIGMOD International Conference on Management of Data, pp. 689 – 700, 2007.

[21] X. Xiao, and Y. Tao. Personalized privacy preservation. In Proceedings of the ACM SIGMOD international conference on Management of data, pp. 229 – 240, 2006.

[22] C. Dwork. Differential Privacy: A Survey of Results. In International Conference on Theory and Applications of Models of Computation, pp. 1 – 19, 2008.

[23] C. Dwork, and A. Roth. The algorithmic foundations of differential privacy. In Foundations and Trends in Theoretical Computer Science, 9, pp. 211 – 407, 2014.

[24] A. Narayanan, and V. Shmatikov. De-anonymizing Social Networks. In IEEE Symposium on Security and Privacy, pp. 173 – 187, 2009.

[25] B. Zhou, J. Pei, and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. In ACM SIGKDD Explorations, pp. 12 – 22, 2008.

[26] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. In VLDB Journal 19, pp. 797 – 823, 2010.

[27] H. Li, Q. Chen, H. Zhu, D. Ma, H. Wen, and X. Shen. Privacy Leakage via De-Anonymization and Aggregation in Heterogeneous Social Networks. In IEEE Transactions on Dependable and Secure Computing, 17(2), pp. 350 – 362, 2020.

[28] Noticeable and Negligible Functions, https://people.eecs.berkeley.edu/sanjamg/classes/cs276-fall14/scribe/lec02.pdf.

[29] O. Goldreich. Foundations of Cryptography. Cambridge University Press, 2004.

[30] X. Xiao, and Y. Tao. Anatomy: simple and effective privacy preservation. In Proceedings of International Conference on Very Large Data Bases, pp. 139 – 150, 2006.

[31] Disease Ontology, http://disease-ontology.org/.

[32] B. Fung, K. Wang, A. Fu, and P. Yu. Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

[33] D. Lambert. Measures of disclosure risk and harm. In Journal of Official Statistics, 9(2), pp. 313 – 331, 1993.

[34] U. C. Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Adult.

[35] ARX Data Anonymization Tool, https://arx.deidentifier.org/anonymization-tool/.

[36] The HIPAA Privacy Rule, https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.

[37] P. Samarati. Protecting respondents identities in microdata release. in IEEE Transactions on Knowledge and Data Engineering, 13(6), pp. 1010 – 1027, 2001.

[38] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. In ACM Computing Survey, 42(4), 2010.

[39] B. Zhou, and J. Pei. Preserving Privacy in Social Networks Against Neighborhood Attacks. In International Conference on Data Engineering, pp. 506 – 515, 2008.

[40] C. Clifton, and T. Tassa. On syntactic anonymity and differential privacy. In IEEE 29th International Conference on Data Engineering Workshops, pp. 88 – 93, 2013.

[41] N. Holohan, S. Antonatos, S. Braghin, and P. Aonghusa. (k, ϵ) - anonymity: k - anonymity with ϵ - differential privacy. In arXiv:1710.01615 [cs.CR], 2017.

[42] N. Li, W. Qardaji, and D, Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In Proceedings of the ACM Symposium on Information, Computer and Communications Security, pp. 32 – 33, 2012.

[43] L. Alfantoukh, and A. Durresi. Techniques for Collecting data in Social Networks. In International Conference on Network-Based Information Systems, pp. 336 – 341, 2014.

[44] Data Breaches, https://www.expresscomputer.in/security/386-mn-user-records-from-18-companies-stolen-in-data-breaches/61730/.

[45] M. Al-garadia, M. Khan, K. Varathan, G. Mujtaba, and A. Al-Kabsi. Using online social networks to track a pandemic: A systematic review. In Journal of Biomedical Informatics, vol. 62 pp. 1 – 11, 2016.

[46] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In Proceedings of International Conference on World Wide Web, pp. 181 – 190, 2007.

[47] P. Pedarsani, and M. Grossglauser. On the privacy of anonymized networks. In Proceedings of International Conference on Knowledge discovery and data mining, pp. 1235 – 1243, 2011.

[48] S. Nilizadeh, A. Kapadia, and Y. Ahn. Community-enhanced deanonymization of online social networks. In Proceedings of ACM SIGSAC Conference on Computer and Communications Security, pp. 537 – 548, 2014.

[49] S. Ji, T. Wang, J. Chen, W. Li, P. Mittal, and R. Beyah. De-SAG: On the de-anonymization of structure-attribute graph data. In IEEE Transactions on Dependable and Secure Computing, 16(4), pp. 594 – 607, 2019.

[50] Stanford Large Network Dataset Collection, https://snap.stanford.edu/data/index. html.

[51] Z. Pawlak. Rough set theory and its applications to data analysis. In Journal of Cybernetics and Systems, 29(7), pp. 661 – 688, 1998.

[52] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In International Conference on Very Large Data Bases, pp. 487 – 499, 1994.

[53] C. Aggarwal. Data Mining: The Textbook. In Springer International Publishing, 2015.

[54] R. Farahbakhsh, X. Han, A. Cuevas and N. Crespi. Analysis of publicly disclosed information in facebook profiles. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 699 – 705, 2013.

[55] Y. Dong, Y. Yang, J. Tang, Y. Yang and N. Chawla.Inferring User Demographics and Social Strategies in Mobile Social Networks. In ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 15 – 24, 2014.

[56] B. Zhou and J. Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. In Knowledge and Information Systems, 28, pp. 47 – 77, 2011.

[57] C. Tai, P. Yu, D. Yang and M. Chen. Privacy preserving social network publication against friendship attacks. In ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 1262 – 1270, 2011.

[58] J. Cheng, A. Fu and J. Liu. K-isomorphism: privacy preserving network publication against structural attacks. In ACM SIGMOD International Conference on Management of data, pp. 459 – 470, 2010.

[59] K. Liu and E. Terzi. Towards identity anonymization on graphs. In ACM SIGMOD International Conference on Management of data, pp. 93 – 106, 2008.

[60] L. Zou, L. Chen and M. Özsu. K-automorphism: A general framework for privacy preserving network publication. In PVLDB, 2(1), pp. 946 – 957, 2009.

[61] M. Yuan, L. Chen, P. Yu and T. Yu. Protecting Sensitive Labels in Social Network Data Anonymization. In IEEE Transactions on Knowledge and Data Engineering, 25(3), pp. 633 – 647, 2013.

[62] P. Mittal, C. Papamanthou and D. Song. Preserving link privacy in social network based systems. In Network and Distributed System Security Symposium, pp. 1 – 15, 2013.

[63] S. Ji, P. Mittal and R. Beyah. Graph Data Anonymization, De-Anonymization Attacks, and De-Anonymizability Quantification: A Survey. In IEEE Communications Surveys & Tutorials, 19(2), pp. 1305 – 1326, 2017.

[64] X. Li, C. Zhang, T. Jung, J. Qian and L. Chen. Graph-based privacy-preserving data publication. In International Conference on Computer Communications, pp. 1 – 9, 2016.

[65] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and Zhu. An Approximation Algorithm for k-Anonymity. In Proceedings of the International Conference on Database Theory, 2005.

[66] V. Vazirani. Approximation Algorithms: http://tocs.ulb.tu-darmstadt.de/98805355.pdf.

[67] A. Meyerson and R. Williams. On the complexity of optimal k-Anonymity. In Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 223 – 228, 2004.

[68] P. Samarati. Protecting respondants' identities in microdata release. In IEEE Transactions on Knowledge and Data Engineering, 13(6), 2001.

[69] C. C. Aggarwal. On *k*-anonymity and the curse of dimensionality.In Proceedings of International Conference on Very Large DataBases, pp. 901 – 909, 2005.

[70] M. E. Nergiz, C. Clifton and A. E. Nergiz. Multirelational k-Anonymity. In IEEE Transactions on Knowledge and Data Engineering, 21(8), pp. 1104 – 1117, 2009.

[71] A. Appari and M. Johnson. Information security and privacy in healthcare: current state of research. In Int. J. Internet and Enterprise Management, 6(4), pp. 279 – 314, 2010.

[72] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi and M. Saadi. Big data security and privacy in healthcare: A Review. In Procedia Computer Science, vol. 113, pp. 73 – 80, 2017.

[73] W. Price and G. Cohen. Privacy in the age of medical big data. In Nat Med, 25, pp. 37 – 43, 2019.

[74] Sankar P, Mora S, Merz JF, Jones NL. Patient perspectives of medical confidentiality: a review of the literature. J Gen Intern Med., 18(8), pp. 659 – 669, 2003.

[75] A. Beresford and F. Stajano. Location privacy in pervasive computing. In IEEE Pervasive computing, no. 1, pp. 46 – 55, 2003.

[76] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li. Achieving k-anonymity in privacy-aware location-based services" in INFOCOM, pp. 754 – 762, 2014.

[77] L. Liu. From data privacy to location privacy: models and algorithms. In Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 1429 – 1430, 2007.

[78] H. Chow and M. Mokbel. 2011. Trajectory privacy in location-based services and data publication. In SIGKDD Explor. Newsl. 13(1), pp. 19 – 29, 2011.

[79] S. Jajodia, P. Samarati, S. Wang. Privacy in Location-Based Applications Research Issues and Emerging Trends. In Springer Berlin Heidelberg, 2009.

[80] M. Smith, C. Szongott, B. Henne and G. von Voigt. Big data privacy issues in public social media. In 6th IEEE International Conference on Digital Ecosystems and Technologies, pp. 1 – 6, 2012.

[81] C. Zhang, J. Sun, X. Zhu and Y. Fang. Privacy and security for online social networks: challenges and opportunities. In IEEE Network, 24(4), pp. 13 – 18, 2010.

[82] E. Zheleva and L. Getoor. Privacy in Social Networks: A Survey. In Aggarwal C. (eds) Social Network Data Analytics. Springer, 2011.

[83] R. Gross and A. Acquisti. 2005. Information revelation and privacy in online social networks. In Proceedings of the ACM workshop on Privacy in the electronic society, pp. 71 – 80, 2005.

[84] https://healthitanalytics.com/news/big-data-analytics-show-covid-19-spread-outcomes-by-region.

[85] M. Mondal, S. Bharati, P. Podder, and P. Podder. Data analytics for novel coronavirus disease. In Informatics in medicine unlocked, 20, 2020.

[86] https://www.geospatialworld.net/blogs/popular-apps-covid-19/.

[87] A. Mavragani, and K. Gkillas. COVID-19 predictability in the United States using Google Trends time series. In Scientific Reports, 10, 20693, 2020.

[88] T. Alanzi. A Review of Mobile Applications Available in the App and Google Play Stores Used During the COVID-19 Outbreak. In Journal of multidisciplinary healthcare, 14, pp. 45 – 57, 2021.

[89] R. Gupta, M. Bedi, P. Goyal, S. Wadhera, and V. Verma. Analysis of COVID-19 Tracking Tool in India: Case Study of Aarogya Setu Mobile Application. In Digital Government: Research and Practice, 1(4), 28, 2020.

[90] S. Gao, J. Rao, Y. Kang, Y. Liang, J. Kruse, D. Doepfer, A. Sethi, J. Reyes, J. Patz, and B. Yandell, Mobile phone location data reveal the effect and geographic variation of social distancing on the spread of the COVID-19 epidemic. arXiv:2004.11430 [cs.SI]

[91] N. Elgendy, and A. Elragal. Big Data Analytics: A Literature Review Paper. In Advances in Data Mining, Applications and Theoretical Aspects, pp. 214 – 227, 2014.

[92] K. Kambatla, G. Kollias, V. Kumar, and A. Grama. Trends in big data analytics. In Journal of Parallel and Distributed Computing, 74(7), pp. 2561 – 2573, 2014.

[93] N. Ghani, S. Hamid, I Hashem, and E. Ahmed. Social media big data analytics: A survey. In Computers in Human Behavior, 101, pp. 417 – 428, 2019.

[94] C. Tsai, C. Lai, H. Chao, and A. Vasilakos. Big data analytics: a survey. In Journal of Big Data, 2(21), 2015.

[95] C. Aggarwal. An Introduction to Social Network Data Analytics. In Social Network Data Analytics, Springer, 2011.

[96] C. Bousquet. Mining Social Media Data for Policing, the Ethical Way, https://datasmart.ash.harvard.edu/news/article/mining-social-media-data-policing-ethical-way

[97] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning, The MIT Press, 2016.

[98] P. Ongsulee, Artificial intelligence, machine learning and deep learning. In International Conference on ICT and Knowledge Engineering, pp. 1 – 6, 2017.

[99] G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. García, I. Heredia, P. Malík, and L. Hluchý. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. In Artificial Intelligence Review, 52, pp. 77 – 124, 2019.

[100] https://www.ibm.com/cloud/learn/machine-learning.

[101] C. Holsapple, A. Lee-Post, and R. Pakath. A unified foundation for business analytics. In Decision Support Systems, 64, pp. 130 – 141, 2014.

[102] https://www.cio.com/article/3601764/what-is-business-analytics-using-data-to-predict-business-outcomes.html.

[103] https://analyticsindiamag.com/how-big-data-is-the-game-changer-for-indian-government-in-e-governance/.

[104] M. Sarker, M. Wu, and M. Hossin. Smart governance through bigdata: Digital transformation of public agencies. In International Conference on Artificial Intelligence and Big Data, pp. 62 – 70, 2018.

[105] https://2021.ai/covid-19-public-datasets-important/.

[106] K. Verbert, N. Manouselis, H. Drachsler, and E. Duval. Dataset-Driven Research to Support Learning and Knowledge Analytics. In Educational Technology & Society, 15(3), pp. 133 – 148, 2012.

[107] S. Dash, S. Shakyawar, M. Sharma, and S. Kaushik. Big data in healthcare: management, analysis and future prospects. In Journal of Big Data, 6(54), 2019.

[108] S. Akter, and S. Wamba. Big data and disaster management: a systematic review and agenda for future research. In Annals of Operations Research, 283, pp. 939 – 959, 2019.

[109] https://safetymanagement.eku.edu/blog/4-ways-big-data-is-revolutionizing-emergency-management/.

[110] T. Li, and N. Li. On the tradeoff between privacy and utility in data publishing. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 517 – 526, 2009.

[111] V. Rastogi, D. Suciu, and S. Hong. The boundary between privacy and utility in data publishing. In Proceedings of the 33rd international conference on Very large data bases, VLDB Endowment, pp. 531 – 542, 2007.

[112] L. Sankar, S. Rajagopalan, and H. Poor. Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach. In IEEE Transactions on Information Forensics and Security, 8(6), pp. 838 – 852, 2013.

[113] R. Wong, A. Fu, K. Wang, P. Yu, and J. Pei. Can the Utility of Anonymized Data be Used for Privacy Breaches? In ACM Transactions on Knowledge Discovery from Data, 5(3), 2011.

[114] F. Sangogboye, R. Jia, T. Hong, C. Spanos, and M. Kjærgaard. A Framework for Privacy-Preserving Data Publishing with Enhanced Utility for Cyber-Physical Systems. In ACM Transactions on Sensor Networks, 14(30), 2018.

[115] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu. Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 785 – 790, 2006.

[116] J. Brickell, and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 70 – 78, 2008.

[117] D. Kifer, and J. Gehrke. Injecting utility into anonymized datasets. In Proceedings of ACM SIGMOD international conference on Management of data, pp. 217 – 228, 2006.

[118] C. Aggarwal. On k-anonymity and the curse of dimensionality. In Proceedings of the 31st international conference on Very large data bases, VLDB Endowment, pp. 901 – 909, 2005.

[119] Location data sharing: What are the concerns around privacy?, https://www.geospatialworld.net/blogs/is-your-location-data-safe/.

[120] R. Ness. Influence of the HIPAA Privacy Rule on health research. In JAMA, 298(18), pp. 2164 – 2170 2007.

[121] S. Hoffman. Electronic Health Records and Medical Big Data. In Cambridge University Press, New York, 2016.

[122] N. Terry. Existential challenges for healthcare data protection in the United States. In Ethics, Medicine and Public Health, 3(1), pp. 19 – 27, 2017.

[123] P. Pedarsani, D. Figueiredo, and M. Grossglauser. A Bayesian method for matching two similar graphs without seeds. In 51st Annual Allerton Conference on Communication, Control, and Computing, pp. 1598 – 1607, 2013.

[124] K. Li, G. Lu, G. Luo, and Z. Cai. Seed-free Graph De-anonymiztiation with Adversarial Learning. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 745 – 754, 2020.

[125] M. Srivatsa, and M. Hicks. Deanonymizing mobility traces: using social network as a side-channel. In Proceedings of the ACM conference on Computer and communications security, pp. 628 – 637, 2012.

[126] S. Ji, W. Li, M. Srivatsa, J. He, and R. Beyah. Structure based data de-anonymization of social networks and mobility traces. In Proceedings of International Conference on Information Security, pp. 237 – 254, 2014.

[127] D. Yin, Y. Shen, and C. Liu. Attribute Couplet Attacks and Privacy Preservation in Social Networks. In IEEE Access, 5, pp. 25295 – 25305, 2017.

[128] Y. Shao, J. Liu, S. Shi, Y. Zhang, and B. Cui. Fast De-anonymization of Social Networks with Structural Information. In Data Science and Engineering, 4, pp. 76 – 92, 2019.

[129] J. Qian, X. Li, T. Jung, Y. Fan, Y. Wang, and S. Tang. Social Network De-anonymization: More Adversarial Knowledge, More Users Re-identified?. In ACM Transactions on Internet Technology, 19(3), pp. 1 – 22, 2019.

[130] L. Nie, L. Zhang, M. Wang, R. Hong, A. Farseev, and T. Chua. Learning User Attributes via Mobile Social Multimedia Analytics. In ACM Transactions on Intelligent Systems and Technology,8(3), pp. 1 – 19, 2017.

[131] D. Jurgens. That's what friends are for: inferring location in online social media platforms based on social relationships. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp.273 – 282, 2013.

[132] J. He, W. Chu, and Z. Liu. Inferring privacy information from social networks. In Intelligence and Security Informatics, Springer, pp.154 – 165, 2006.

[133] N. Gong, and B. Liu. Attribute Inference Attacks in Online Social Networks. In ACM Transactions on Privacy and Security, 21(1), pp. 1 – 30, 2018.

[134] Two Billion Files Leaked in US Data Breaches in 2017, https://www.infosecurity-magazine.com/news/two-billion-files-leaked-in-us-data/.

[135] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In ACM Transactions on Algorithms, 6(3), 2010.

[136] T. Li, N. Li, J. Zhang, and I. Molloy. Slicing: A New Approach for Privacy Preserving Data Publishing. In IEEE Transactions on Knowledge and Data Engineering, 24(3), pp. 561 – 574, 2012.

[137] https://analyticsindiamag.com/10-biggest-data-breaches-that-made-headlines-in-2020/.

[138] https://www.identityforce.com/blog/2021-data-breaches.

[139] https://datareportal.com/social-media-users.

[140] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[141] K. Nagarajan, M. Muniyandi, B. Palani, and, S. Sellappan. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. In BMC Medical Research Methodology, 20(233), 2020.

[142] A. Karaivanov. A social network model of COVID-19. In PLOS ONE, 15(10), pp. 1 – 33, 2020.

[143] https://www.securitymagazine.com/articles/94327-million-facebook-instagram-and-linkedin-users-scraped-data-exposed.

[144] https://www.cshub.com/attacks/articles/iotw-facebook-data-leak-impacts-533-million-users.

[145] J. Cabañas, Á. Cuevas, and R. Cuevas. Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. In USENIX Security Symposium, pp. 479 – 495, 2018.

[146] https://en.wikipedia.org/wiki/List_of_data_breaches.

[147] E. Ryu, Y. Rong, J. Li, and A. Machanavajjhala. CURSO: Protect Yourself from Curse of Attribute Inference. In Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks, pp. 13 – 18.

[148] Y. Zhong, N. Jing Yuan, W. Zhong, F. Zhang, and X. Xie., You are where you go: Inferring demographic attributes from location check-ins. In WSDM, 2015.

[149] A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. In Proceedings of the third ACM international conference on Web Search and Data Mining, pp. 251 – 260, 2010.

[150] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. In Proceedings of the National Academy of Sciences of the United States of America, 107(52), pp. 22436 – 22441, 2010.

[151] N. Desai, and M.L. Das. Rule Based Anonymization Against Inference Attack in Social Networks. In International Journal of Social Computing and Cyber-Physical Systems, Inderscience, 2(3), pp. 212 – 228, 2021.

[152] N. Desai, and M.L. Das. DeSAN: De-anonymization against Background Knowledge in Social Networks. In International Conference on Information and Communication Systems (ICICS), pp. 99 – 105, 2021.

[153] K. Hemantha, N. Desai, and M. L. Das. On Minimality Attack for Privacy-Preserving Data Publishing. In Proceedings of International Symposium on Security in Computing and Communication, Springer, pp. 324 – 335, 2019.

[154] N. Desai, and M.L. Das. Privacy-Preserving Scheme Against Inference Attack in Social Networks. In Proceedings of International Conference on Computing, Communications, and Cyber-Security, Springer, pp. 819 – 830, 2020.

[155] X. Ding, C. Wang, K. Choo, and H. Jin. A Novel Privacy Preserving Framework for Large Scale Graph Data Publishing. In IEEE Transactions on Knowledge and Data Engineering, 33(2), pp. 331 – 343, 2021.

[156] Q. Liu, G. Wang, F. Li, S. Yang, and J. Wu. Preserving Privacy with Probabilistic Indistinguishability in Weighted Social Networks. In IEEE Transactions on Parallel and Distributed Systems, 28(5), pp. 1417 – 1429, 2017.

[157] R. Trujillo-Rasua, and I. Yero. k-Metric antidimension: A privacy measure for social graphs. In Information Sciences, 328, pp. 403 – 417, 2016.

[158] S. Mauw, Y. Ramírez-Cruz, and R. Trujillo-Rasua. Conditional adjacency anonymity in social graphs under active attacks. In Knowledge and Information Systems, 61, pp. 485 – 511, 2019.

CHAPTER A

Publications of the Thesis Work

- N. Desai, M. L. Das, and N. Kumar. $(\theta; D; \alpha)$ - Privacy: A Privacy Model Against Adversarial Background Knowledge. IEEE Transactions on Computational Social Systems, (under review).

- N. Desai and M.L. Das. Rule Based Anonymization Against Inference Attack in Social Networks. International Journal of Social Computing and Cyber-Physical Systems, Inderscience, 2(3), pages 212-228, 2021.

- N. Desai and M.L. Das. DeSAN: De-anonymization against Background Knowledge in Social Networks. International Conference on Information and Communication Systems, IEEE, pages 99-105, 2021.

- K. Hemantha, N. Desai, M. L. Das. On Minimality Attack for Privacy - Preserving Data Publishing. In Proceedings of International Symposium on Security in Computing and Communication, Springer, pages 324-335, 2019.

- N. Desai and M.L. Das. Privacy - Preserving Scheme Against Inference Attack in Social Networks. In Proceedings of International Conference on Computing, Communications, and Cyber-Security, Springer, pages 819-830, 2020.

In - Progress Work

- N. Desai and M.L. Das. On Background Knowledge Attacks in Privacy - Preserving Data Publishing Models.

- N. Desai and M.L. Das. Preserving Privacy against Background Knowledge in Social Networks.