

PRIVACY PRESERVING DATA PUBLISHING
WITH MULTIPLE SENSITIVE ATTRIBUTES

by

Ahmed Abdalaal

Submitted to the Graduate School of
Engineering and Natural Sciences
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

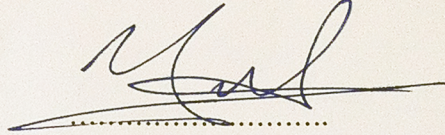
Sabanci University

August 2012


PRIVACY PRESERVING DATA PUBLISHING
WITH MULTIPLE SENSITIVE ATTRIBUTES

APPROVED BY

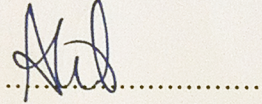
Assoc. Prof. Yücel Saygın
Sabancı Üniversitesi (Thesis Supervisor)



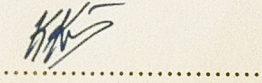
Assoc. Prof. Albert Levi
Sabancı Üniversitesi



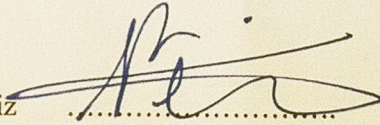
Asst. Prof. Ali İnan
Işık Üniversitesi



Asst. Prof. Kağan Kurşungöz
Sabancı Üniversitesi



Asst. Prof. Mehmet Ercan Nergiz
Zirve Üniversitesi



DATE OF APPROVAL:

August 2, 2012

© Ahmed Abdalaal 2012

All Rights Reserved

ABSTRACT

Data mining is the process of extracting hidden predictive information from large databases, it has a great potential to help governments, researchers and companies focus on the most significant information in their data warehouses. High quality data and effective data publishing are needed to gain a high impact from data mining process. However there is a clear need to preserve individual privacy in the released data. Privacy-preserving data publishing is a research topic of eliminating privacy threats. At the same time it provides useful information in the released data. Normally datasets include many sensitive attributes; it may contain static data or dynamic data. Datasets may need to publish multiple updated releases with different time stamps. As a concrete example, public opinions include highly sensitive information about an individual and may reflect a person's perspective, understanding, particular feelings, way of life, and desires. On one hand, public opinion is often collected through a central server which keeps a user profile for each participant and needs to publish this data for researchers to deeply analyze. On the other hand, new privacy concerns arise and user's privacy can be at risk. The user's opinion is sensitive information and it must be protected before and after data publishing. Opinions are about a few issues, while the total number of issues is huge. In this case we will deal with multiple sensitive attributes in order to develop an efficient model. Furthermore, opinions are gathered and published periodically, correlations between sensitive attributes in different releases may occur. Thus the anonymization technique must care about previous releases as well as the dependencies between released issues.

This dissertation identifies a new privacy problem of public opinions. In addition it presents two probabilistic anonymization algorithms based on the concepts of k -anonymity [1, 2] and ℓ -diversity [3, 4] to solve the problem of both publishing datasets with multiple sensitive attributes and publishing dynamic datasets. Proposed algorithms provide a heuristic solution for multidimensional quasi-identifier and multidimensional sensitive attributes using probabilistic ℓ -diverse definition. Experimental results show that these algorithms clearly outperform the existing algorithms in term of anonymization accuracy.

ÖZET

Veri madenciliği tahmin edilebilir gizli bilgiyi büyük veri tabanlarından çıkarma işlemidir. Devletlere, araştırmacılara ve şirketlere veri ambarlarındaki en önemli bilgilere odaklanmaları konusunda yardım etmek gibi büyük bir potansiyele sahiptir. Veri madenciliğinin yüksek bir etki sağlayabilmesi için yüksek kaliteli veriye ve etkin veri yayıncılığına ihtiyaç duyulur. Buna karşın, yayınlanan veri için kişisel mahremiyetin korunması da açık bir ihtiyaçtır. "Mahremiyet koruyan veri yayıncılığı" yayınlanan veriden faydalı bilgiler elde ederken mahremiyet ihlaline yol açabilecek tehlikeleri önlemenin yollarını inceleyen bir araştırma konusudur. Normalde veri kümelerinin birçok hassas özelliği vardır; durağan veya devingen veri içerebilirler. Veri kümeleri farklı zaman damgalı birden çok güncellenmiş sürümü yayınlamak durumunda kalabilirler. Somut bir örnek vermek gerekirse, kamuoyu bireyler hakkında yüksek hassaslıkta bilgi içerir ve bireylerin görüş açısını, anlayışını, duygularını, yaşam tarzını ve arzularını yansıtabilir. Bir yandan, kamuoyu her katılımcı için bir kullanıcı profilinin tutulduğu merkezi sunucular tarafından toplanır. Öte yandan, yeni mahremiyet sorunları ortaya çıkar ve kullanıcının mahremiyeti tehlikeye girebilir. Kullanıcının görüşü hassas bir bilgidir ve veri yayıncılığından önce ve sonra da korunmalıdır. Görüşler genelde birkaç mevzu hakkındadır ama, toplam mevzu sayısı çok fazladır. Bu durumda, etkili bir model geliştirebilmek için birden çok hassas özellikte başı çıkılmalıdır. İlaveten, görüşler belirli aralıklarla toplanıp yayınlandığında, hassas özelliklerin farklı sürümleri arasında ilişkiler ortaya çıkabilir. Bu yüzden, isimseleştirme yöntemi yayınlanan konular arasındaki bağımlılığı incelediği gibi önceki sürümleri de göz önüne almalıdır.

Bu tez kamuoyu hakkında yeni bir mahremiyet problemi tespit ediyor. Bunun yanında, devingen veri kümelerini yayınlamak ve birden çok hassas özellik içeren veri kümelerini yayınlamak problemlerini çözmek için k-isimsizleştirme [1, 2] ve ell-çeşitlilik [3, 4] kavramlarına dayanan iki olasılıksal algoritma sunuyor. Önerilen algoritmalar, olasılıksal ell-çeşitlilik tanımını kullanarak çok boyutlu belirteçimsiler ve çok boyutlu hassas özellikler için sezgisel bir çözüm sağlıyor. Deneysel sonuçlar bu algoritmaların isimsizleştirme doğruluk payı açısından var olan diğer algoritmaları geride bıraktığını gösteriyor.

Dedicated to my family

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Prof. Yücel Saygın for his help with this work as well as my graduate study. He has always been understanding and supportive and given very good advice on any matter.

Dr. Mehmet Ercan Nergiz has been practically my co-advisors. I am indebted to him for helping the security analysis and reviewing my work very carefully.

Also I owe a Great many thanks to Prof. ErKay Savaş, Prof. Albert Levi and Prof. Kemal İnan for their helpful support during my study.

As well, I would like to thank Ms. Evrim Güngör, from International Relations Office and Mrs. Gülin Karahüseyinoğlu, from Student Resources Office for their administrative support.

Although, I don't have words to express my gratitude and thanks. I dedicate a special thanks to my family for their love and support. I hope to return the favor someday soon.

TABLE OF CONTENTS

ABSTRACT	iv
ACKNOWLEDGEMENTS	vii
LIST OF FIGURES	xi
LIST OF TABLES	xii
1. INTRODUCTION	1
1.1 Motivations.....	1
1.2 Contributions.....	7
1.3 Structure of the Dissertation.....	8
2. BACKGROUND AND RELATED WORKS.....	10
2.1 Privacy of Public Opinions	10
2.2 Privacy-Preserving Data Publishing.....	11
2.3 Privacy-Preserving Data Publishing Models.....	13
2.3.1 Statistical Methods.....	14
2.3.2 Partitions-Based Anonymization	15
2.3.3 Probabilistic Model.....	21
2.4 Complexity of finding optimal k-anonymity	22
2.5 Privacy-Preserving Data Publishing Possible Attacks	24
2.5.1 Linking Attack	24
2.5.2 Homogeneity Attack	25
2.5.3 Background Knowledge Attack.....	25
2.5.4 Skewness Attack	26
2.5.5 Similarity Attack.....	26
2.5.6 Membership Disclosure	27
2.5.7 Multiple Release Attack.....	27
2.5.8 Minimality Attack.....	28
2.5.9 Inference Attack.....	28
2.5.10 deFinetti Attack.....	29
2.6 INFORMATION LOSS METRICS.....	29
2.6.1 Discernibility Metric.....	29

2.6.2	Loss Metric	30
2.6.3	Average Query Error.....	30
3.	PRIVACY-PRESERVING FOR MULTIPLE SENSITIVE ATTRIBUTES.....	31
3.1	Naïve Approach.....	32
3.2	Machanavajjhala’s et al. Approach	33
3.3	Li and Ye Approach	35
3.4	Gal’s et al Model	36
3.5	Xiao-Chun et al Model.....	38
3.6	Ye et al Model	40
3.7	Fang et al Model.....	40
4.	PRIVACY-PRESERVING FOR DYNAMIC RELEASES.....	41
4.1	SAs Independent Approach.....	42
4.2	SAs Dependent Approaches.....	44
4.2.1	Record-linking Attack.....	44
4.2.2	Value-association Attack	45
4.2.3	Correspondence Attack	46
4.2.4	Value-equivalence Attack.....	48
4.2.5	Tuple-equivalence Attack	52
4.3	ρ -different Approach	53
5.	MSA DIVERSITY ALGORITHM	59
5.1	Adversary Model and Privacy Standard.....	59
5.2	Problem Formulation.....	62
5.3	Data preprocessing	62
5.4	Checking for MSA Diversity	62
5.5	Generalization Algorithm.....	63
5.5.1	Mapping multi-dimensional QI to one-dimension.....	64
5.5.2	The MSA-diversity Heuristic Algorithm	69
6.	EXPERIMENTAL RESULTS	70
6.1	Utility - varying ℓ and d	71
6.2	Comparison with Previous Work	71
6.2.1	Utility comparison - varying ℓ	72
6.2.2	Utility comparison - varying d	73
6.2.3	Probability of disclosure comparison.....	75

7. CONCLUSIONS.....	77
APPENDIX A: LIST OF ACRONYMS.....	78
REFERENCES.....	79
VITA.....	86

LIST OF FIGURES

Figure 1 : Public opinions on acceptance homosexuality in different countries[10].....	2
Figure 2 : Privacy-Preserving Data Publishing general process	11
Figure 3 : Age generalization tree.....	18
Figure 4 : 3-diversity groups using Gal’s et al. model.....	37
Figure 5 : Hilbert curve mapping for T1 and T2	55
Figure 6 : 3-diversity groups using our model.....	61
Figure 7 : Hilbert curve mapping.....	63
Figure 8 : Different types of space-filling curves	64
Figure 9 : 3D scatter plot for Table 43.....	66
Figure 10 : Groups construction process	66
Figure 11 : Permutation matrices for 3 elements	67
Figure 12 : One matrix of Costas arrays for 3 elements	67
Figure 13 : Pseudo code for our heuristic algorithm.....	69
Figure 14 : LM, Information loss with varying ℓ and d	70
Figure 15 : DM, Information loss with varying ℓ and d	71
Figure 16 : DM comparison, varying ℓ and $d=2$	72
Figure 17 : Query accuracy with varying ℓ and $d=2$	72
Figure 18 : DM comparison, varying ℓ and $d=5$	73
Figure 19 : Query accuracy with varying ℓ and $d=5$	73
Figure 20 : DM, varying number of sensitive attributes and $\ell=2$	74
Figure 21 : Query accuracy with varying number of sensitive attributes and $\ell = 2$	74
Figure 22 : DM, varying number of sensitive attributes and $\ell = 5$	75
Figure 23 : Query accuracy with varying number of sensitive attributes and $\ell = 5$	75
Figure 24 : Probability of disclosure for each tuple, $d=2$ and $\ell = 5$	76
Figure 25 : Probability of disclosure for each tuple, $d=5$ and $\ell=2$	76

LIST OF TABLES

Table 1 : The microdata sample T	3
Table 2 : Anonymized data T*	3
Table 3 : Public data P	4
Table 4 : Private data T	5
Table 5 : Gal's et al released data [11]	6
Table 6 : Microdata Table MT	13
Table 7 : Suppression mechanism.....	16
Table 8 : Generalization mechanism.....	17
Table 9 : Bucketization mechanism.....	18
Table 10 : ℓ -diversity model	19
Table 11: Global-recoding and local-recoding	23
Table 12 : Linking Attack	24
Table 13 : Homogeneity and background knowledge attacks	25
Table 14 : Skewness attack	26
Table 15 : Similarity attack.....	27
Table 16 : Multiple release attack.....	28
Table 17 : Microdata Table with (d) SA.....	31
Table 18 : The microdata and anonymized data sample T	32
Table 19 : Microdata of Machanavajjhala et al. approach for MSA	33
Table 20 : Anonymized data of Machanavajjhala et al. approach for MSA.....	34
Table 21 : Li and Ye approach.....	35
Table 22 : Gal's et al released data T*1	36
Table 23 : Microdata for Xiao-Chun et al. model.....	38
Table 24 : Anonymized data of Xiao-Chun et al (MMDCF algorithm)	39
Table 25 : Ye et al model example	40
Table 26 : The microdata of two independent issues.....	43
Table 27 : The anonymized data for two independent issues	43
Table 28 : The microdata for Join Attack example.....	45
Table 29 : First Release for the Value-association Attack.....	46
Table 30 : Second Release for the Value-association Attack	46
Table 31 : An anonymized data for R1 in the Correspondence Attack	47
Table 32: An anonymized data for R2 in the Correspondence Attack	47

Table 33 : 2-diversity anonymization	48
Table 34 : A naive 2-diversity anonymized data at R2.....	49
Table 35 : 2-invariance anonymized data at R2.....	50
Table 36 : 2-invariance, 2-value equivalence anonymized data	51
Table 37 : Example of diseases correlations (C)	54
Table 38 : Two datasets releases.....	54
Table 39 : 2-diversity anonymized data at R1	56
Table 40 : 2-invariance anonymized data at R2.....	57
Table 41 : 2-different anonymized data at R2	58
Table 42 : MSA-diversity released data T2 *	60
Table 43 : Microdata with one dimension QI	65

1. INTRODUCTION

1.1 Motivations

Governments, political parties, social associations, etc., need to stay in touch with their audiences. Understanding public opinion is essential for a democratic process. Public opinion helps political decision-makers to understand underlying issues that are of utmost importance to them. Issues such as discrimination, gay rights, abortion, cloning, capital punishment, affirmative action, euthanasia, and national security are examples of hot public opinion topics governments need a comprehensive analysis of [5-8]. Social research and opinion polls give people the opportunity to express their views regularly on different topics and provide an efficient way to measure public opinion. Since 1973, the European Commission has been monitoring the evolution of public opinion in the Member States [9], information which helps in the preparation of texts, decision-making and the evaluation of its work.

A user profile needs to be constructed for individuals to participate in the public opinion process. These profiles contain valuable data about the user, such as nation, gender, city, and so on. These data may also contain Name, address, User's social ID, Date of birth and Sex. Due to the rapid developments in computer and network technologies, many on-line public opinion polls and mobile-based public opinion systems are used in the opinion process, thus enabling greater participation. Therefore, the public opinion process must guarantee that individuals can express their preferences freely without any threats to their own privacy. Polls done under the risk of identification may not be accurate.

For example, Figure 1 shows that in Africa, Asia and the Middle East, attitudes toward homosexuals are generally negative while the European and American voters are generally positive [10]. Voters with Yes/No from an opposing/supporting country may receive public pressure from majority of their countryman if their identities are revealed. If voters are not convinced that such a risk is small, they may not want to reveal their true opinion causing a bias towards the more common attitudes.

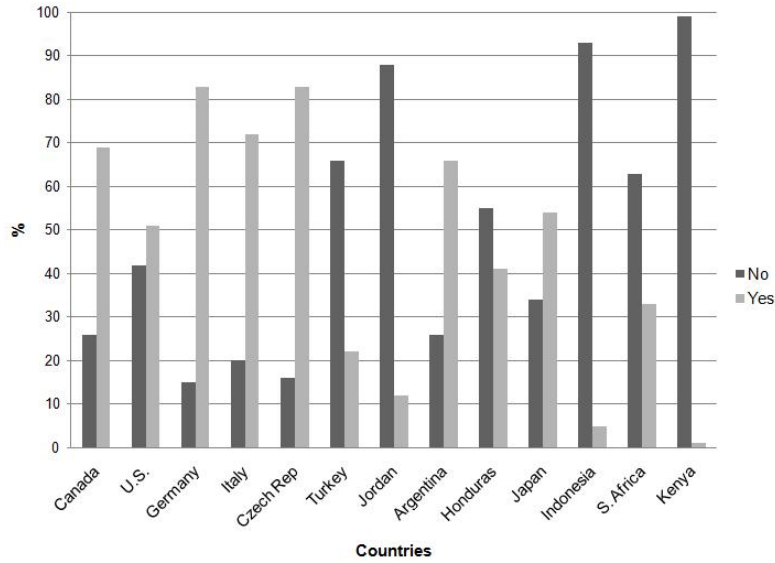


Figure 1 : Public opinions on acceptance homosexuality in different countries[10]

Public opinion privacy means that neither the organizing authorities nor any other third party can link an opinion to the individual who has cast it. This requires achieving some degree of anonymity. As a naive approach, anonymity can be achieved by removing the attributes which uniquely identify individual users such as name, SSN, address, phone number. However, as shown in [11-14], this approach will not be enough to ensure anonymity due to the existence of quasi-identifier attributes (QI) which can be used together to identify individuals based on their profile information. Attributes like birth date, gender and ZIP code, when used together, can accurately identify individuals. [15]

In this dissertation, we examine a case in which we have a large number of opinions and the data holder needs to publish this data. Adversaries can launch an attack based on user profile and public opinion. We focus on the protection of the relationship between the quasi-identifiers and multiple sensitive attributes. Many works like k-anonymity, ℓ -diversity, t-closeness, etc., have been proposed as a privacy protection model for micro data [3], [4]. However, most of models only deal with data with a single sensitive attribute [3], [16], [17], [18], [19],[20]. In addition, we aim to preserve privacy when there are correlations between sensitive attributes within same release or different releases.

Various techniques can be employed to provide anonymity in a public opinion process. Most electronic voting schemes like the Blind signature scheme [21],[22] the Homomorphism scheme [2] and the Randomization-Based scheme [23] are based on cryptography techniques. These provide on-line privacy preservation for voters, which is also

suitable for use in the public opinion process. Also, k-Anonymous message transmission protocol [24] preserves user privacy during the voting process, and does not require the existence of trusted third parties. This technique tries to protect a user’s privacy during the voting process; however, in public opinion polls we need to provide anonymity after the opinions are collected and more specifically when the central servers want to publish this data.

To limit sensitive information disclosure in data publishing, ℓ -Diversity [3] has been proposed. One definition of ℓ -diversity requires that there are at least ℓ values of sensitive attributes in each equivalence class. It has been shown in [11], [25], [12] that under non-membership information ℓ -diversity fails to protect privacy. As an example, Table 1 shows some voter’s records, where age and zip code are the quasi-identifiers and Issue1 and Issue2 are the sensitive attributes. The anonymization in Table 2 satisfies 3-diversity on Issue1 alone and Issue2 alone. Consider an adversary who has the background knowledge that Amy will not vote for (c) on Issue1, thus the adversary can exclude the tuples with (c) on Issue1. Since the remaining tuples all have (w) on Issue2, the adversary will conclude Amy has voted (w) on Issue 2.

Tuple ID	Quasi-Identifiers (QI)		Sensitive Attributes (SA)	
	Age	Zip code	Issue1(I ₁)	Issue2(I ₂)
Amy	30	1200	b	w
Bob	20	2400	c	x
Che	23	1500	a	w
Dina	27	3400	c	y

Table 1 : The microdata sample T

Quasi-Identifiers (QI)		Sensitive Attributes (SA)	
Age	Zip code	Issue1(I ₁)	Issue2(I ₂)
[20-30]	[1200-3400]	b	w
[20-30]	[1200-3400]	c	x
[20-30]	[1200-3400]	a	w
[20-30]	[1200-3400]	c	y

Table 2 : Anonymized data T*

It has been shown in [11] that direct application of the techniques proposed for these models creates anonymizations that fail to protect privacy under additional background on non-memberships. As an example, take ℓ -diversity which ensures that each individual can at best be mapped to at least ℓ sensitive values and suppose a data holder has the microdata given in Table 1. Directly applying a single-sensitive attribute ℓ -diversity (SSA-diversity) algorithm on the microdata would result in Table 2 which provides 3-diversity. (E.g., an adversary knowing the public table and seeing Table 2 can at best map, say Amy, to 3 distinct values a, b, and c for issue 1, and to w, x, and y for issue 2.) However, if the adversary also knows that Amy does not vote for c for issue 1, she can easily conclude that Amy voted for w for issue 2. Note that public opinion polls collect votes on many issues and it is easy to obtain such non-membership knowledge (compared to membership knowledge) making such attacks a threat in the domain of public opinions.

Tuple ID	Explicit-Identifiers (EI)		Quasi-Identifier (QI)	
	SSN	Name	Age	Zip code
t1	2502	Bob	20	3000
t2	2353	Ken	25	3500
t3	2453	Peter	25	4000
t4	1564	Sam	30	6500
t5	5021	Jane	35	4500
t6	9432	Linda	40	5500
t7	5024	Alice	45	6000
t8	1304	Mandy	50	5000
t9	1202	Tom	55	6500

Table 3 : Public data P

Work in [11] extended the definition of ℓ -diversity to provide protection against non-memberships attacks. Their model ensures that an individual can at best be linked to at least ℓ distinct sensitive values and under i bits of non-membership knowledge, the released data should still satisfy $(\ell-i)$ -diversity.

For example in Table 4 and Table 5, each anonymization group satisfies 3-diversity that is every individual can at best be mapped to at least 3 sensitive values. Even if an adversary knows that, say Linda (t6), does not vote for c on issue1, the adversary will still not be sure whether Linda votes for y or x thus the model ensures 2-diversity within the group under one bit of non-membership knowledge. However, this work does not offer a probabilistic model. That is there is little relation between the privacy parameter ℓ and the probability of disclosure. For example, the table in Table 5 is considered 3-diverse however the probability that Alice (t7) votes for c on issue1 is 1/2. This makes it difficult to make risk/benefit/cost analysis of publishing private data under a privacy parameter ℓ [11, 12, 25].

Tuple ID	Quasi-Identifier (QI)		Sensitive Attributes (SA)	
	Age	Zip code	Issue1 (I ₁)	Issue2 (I ₂)
Bob(t1)	20	3000	a	w
Ken(t2)	25	3500	b	z
Peter(t3)	25	4000	d	x
Sam(t4)	30	6500	a	x
Jane(t5)	35	4500	b	y
Linda(t6)	40	5500	a	y
Alice(t7)	45	6000	c	z
Mandy(t8)	50	5000	a	x
Tom(t9)	55	6500	c	w

Table 4 : Private data T

Tuple ID	Quasi-Identifier (QI)		Sensitive Attributes (SA)	
	Age	Zip code	I ₁	I ₂
t1	[20-25]	[3000-4000]	a	w
t2	[20-25]	[3000-4000]	b	z
t3	[20-25]	[3000-4000]	d	x
t6	[40-55]	[5000-6500]	a	y
t7	[40-55]	[5000-6500]	c	z
t8	[40-55]	[5000-6500]	b	x
t9	[40-55]	[5000-6500]	c	w
t4	*	*	*	*
t5	*	*	*	*

Table 5 : Gal's et al released data [11]

1.2 Contributions

In this dissertation, we combine the best of the two worlds and propose two probabilistic models, MSA-diversity to preserving privacy for data with multiple sensitive attributes, and ρ -different to preserve privacy for dynamic data, which

- protects against identification and non-membership attacks even when we have multiple sensitive attributes,
- and bounds the probability of disclosure allowing risk analysis on the publisher side.

More precisely, MSA- ℓ diversity ensures that the probability of mapping an individual to a sensitive value is bounded by $1/(\ell-i)$ under i bits of non-membership knowledge. As an example, given $\ell=3$, our technique generates the anonymization in Table 42 (page 60) in which the probability of disclosure is bounded by $1/3$ for all individuals. If an adversary knows that, say Bob (t_1), does not vote for d on issue1, the probability that he votes for, say a , on issue1; or say w , on issue2 is still bounded by $1/2$. Our contribution in this thesis can be summarized as follows:

- 1) Formally define probabilistically MSA-diversity privacy protection model for datasets with multiple sensitive attributes.
- 2) Formally define probabilistically ρ -different privacy protection model for dynamic datasets.
- 3) Design a heuristic anonymization algorithm for MSA-diversity. We borrow ideas from state of the art anonymization techniques such as Hilbert curve anonymization [26, 27] to increase utility.
- 4) Moreover, a formally definition of a new attack for publishing dataset with fully dependent sensitive attributes. More details will be discussed in Chapter 4.

1.3 Structure of the Dissertation

Unless otherwise stated, the dissertation examples will be on public opinion data. The data are practically organized as a table of rows (or records, or tuples) and columns (or fields, or attributes). The dissertation has seven chapters.

Chapter 1 “INTRODUCTION”

It provides an introduction to public opinion polls and its relation with privacy-preserving data publishing. There is a clear demand for gathering and sharing public opinions without compromising the participant privacy. We demonstrate an example of public opinion polls and another example of challenges appears when publishing public opinions. Furthermore we declare contributions of this dissertation.

Chapter 2 “BACKGROUND AND RELATED WORKS”

This chapter presents some anonymization models for preserving privacy. In addition it explains a variety of attacks that can be used to disclose the released data, and the related privacy models proposed for preventing such attacks. All discussed models and attacks are applicable to one sensitive attribute. It is also presents three types of information loss metrics which will be used in the experiments part. These metrics are recently used by most of similar models and approaches in the privacy preserving data mining.

Chapter 3 “PRIVACY-PRESERVING FOR MULTIPLE SENSITIVE ATTRIBUTES”

It discusses most of the published work for preserving privacy for data with multiple sensitive attributes. In addition it explains the weaknesses and the attacks still applicable for the released data.

Chapter 4 “PRIVACY-PRESERVING FOR DYNAMIC RELEASES”

It explains recent work for preserving privacy for dynamic data releases and its relations with public opinion polls problem. As will it presents possible attacks applicable to the released data. ρ -different model will be present to preserve participants' privacy.

Chapter 5 “MSA DIVERSITY ALGORITHM”

It explains in detail our MSA-diversity model. Also data preprocessing, Problem formulation, and constructing a probabilistic definition to preserve privacy are discussed.

Chapter 6 “EXPERIMENTAL RESULTS”

It presents results of employing real data set to the MSA-model and Gal’s model. The experiments focus on the variation of the number sensitive attributes. In addition experiments show the effects of diversity variations. For case of comparison we show how MSA-model provides more accurate results than Gal et al’s model, what is more that MSA-model also presents the most accurate released data than other models described in chapter 3.

Chapter 7 “CONCLUSIONS”

It describes the overall conclusions and future work for releasing data with multiple sensitive attributes.

2. BACKGROUND AND RELATED WORKS

2.1 Privacy of Public Opinions

Public opinion is a psychological and social process to collect the individual views, attitude and beliefs about a specific topic. Public opinion has a significant impact on policy making process. A country president, parliament members, political parties, social groups, businessmen, human rights associations, journalists and consultants as well as candidate presidents and candidate parliament members, frequently ask the same question “How does the public think about a certain topic?”. Public opinion is an indicator of the opposition and problems that may be faced in implementation of policies. Such information can be used by policy makers to device party, company or government policies to be realistic rather than idealistic. Politicians need to know public opinions to keep people trust and win reelection. Also, in private sources organizations as the Political Action Committees (PACs) raise money for or against elect specific candidates. These groups can be very effective in policy decisions. Social groups may form interest groups to directly work to raise awareness and actively involved in everything from environmental issues to social issues, all having an impact on policy.

Opinion polling is a way to understand public opinion. It tells us how a population thinks and feels about any given topic. It may use a survey, a questioner, electronic devices, web based polls or a mobile base polls. It categories individuals view about a specific viewpoint. Social scientists and scholars use polls results to explain why respondents still believe or change their minds about the poll topic. Opinion polling is usually designed to represent the opinions of people by conducting a series of questions and then conclude conceptions in ratio or within confidence intervals. These quantitative data often reveal citizens’ preferences, and tell us a sense of how people feel about policy issues, social practices, or lifestyle issues. Opinion polling was an important factor for Unites States 43rd president George W. Bush decision to attack Iraq in 2003 [28]. Bush conclude that American citizen support military actions. This example gives us how public opinion polling leads to critical decisions.

Paper-form polling is traditional way to collect public opinion. A company organizing these polling needs to print many polling forms then destitute it in many places. This need a

large number of equipments and stuff, furthermore it's time cost. The rapid developments in mobile, computer and network technologies change the whole polling process. Nowadays a company is able to use online systems such as web-based polling or social sites polling or even SMS messaging. Participants can use their own computers, tablet or smart phones to give her/his opinion. In order to implement web-based opinion polling, many companies construct a profile for each participant. This profile may contain important information about the participant such as user location, age, gender, occupation or marital status.

The collection of public opinion information facilitates large-scale data mining and data analysis. The information holders such as governments, individual associations and companies have mutual benefits to sharing data among various parties. Moreover, some regulations may require certain data to be released. For example, Netflix, a popular online movie and television rental service, aimed to improve the movie recommendations accuracy therefore released a data set contains anonymous movie ratings of 500,000 subscribers [29].

Public opinion data contain sensitive information about individuals, and sharing such data immediately may reveal individual privacy. As a practical solution data holders may write an agreement, guidelines or general polices with other parties to restrict usage and storage of sensitive data. However to assume a high level trust is impractical solution. Such agreements cannot guarantee careless or misuse of sensitive data, which may lead to violate an individual privacy. A key point is to develop a practical approach keeps data useful and at same time protects individual privacy.

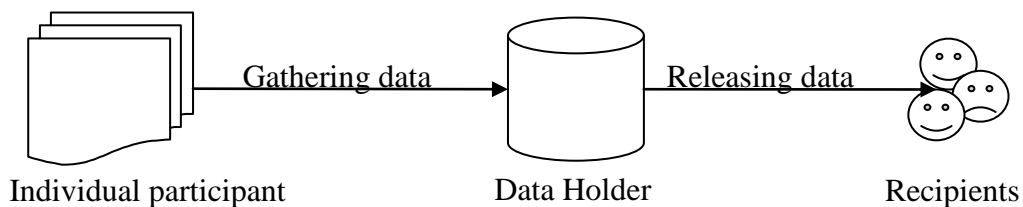


Figure 2 : Privacy-Preserving Data Publishing general process

2.2 Privacy-Preserving Data Publishing

The privacy-preserving data publishing (PPDP) aims protecting the private data and preserving the data utility as much as possible. In PPDP process we have three main users:

1. Individual participant: In public opinion polling, voter will participate and give her/his opinion in a certain topic.
2. Data holder: such as a corporation who organizes the data collection and then anonymizes it. Data holder may be untrusted and gathering information to his own purposes. The voter should be responsible to untrusted data holder scenario and has the ability to decide if it's possible to vote or not. Another scenario might be happen when there is a non-expert data holder. This may leads to publishing a mis-anonymized data. Therefore it's necessary to find a PPDP model to be used in this scenario.
3. Data recipient: researchers who need the data to perform demographic research. Or might be an adversaries use the data to reveal individual privacy.

A common type of the data gathered by data holders is a table form. Many data holders use this table for its simplicity to voters; also data holders can analyze it fast. A table attributes can be categorized as following:

- Explicit Identifiers: provide a means to directly identify a participant, such as name, phone number, and social security number (SSN).
- Quasi Identifiers: attributes can be used together to identify individuals based on their profile information. Attributes like birth date, gender and ZIP code, when used together, can accurately identify individuals.
- Sensitive Attributes: contain personal privacy information like participants' opinion or vote.
- Non-Sensitive Attributes: which when be released will not affect participant directly or indirectly.

The PPDP mechanism namely anonymization or sanitization, seeks to protect participants privacy by hiding the identity of each participant and/or the sensitive data.

Sanitization mechanism represents the variety of all possible data publishing in an application of privacy-preserving data publishing. An anonymization algorithm may use Randomization, Generalization, Suppression, Swapping or Bucketization mechanism to publish a useful and safe data. [30].

2.3 Privacy-Preserving Data Publishing Models

Removing Explicit-Identifiers attributes may not protect participant privacy. [13] shows a real-life privacy threat by linking a combination of attributes (zip code, date of birth, gender) from public voter table with released table. This combination of attributes called the Quasi-Identifiers. Research [31] showed that 87% of the U.S. population had reported characteristics that made them unique based on only such quasi-identifiers. For example, removing SSN and Name from Table 6 will produce Table 4, however it's easy to re-identify participants by check the common Age and Zip code from Table 3 which publicity available and Table 4.

	Explicit-Identifiers (EI)		Quasi-Identifiers (QI)		Sensitive Attributes (SA)	
Tuple ID	SSN	Name	Age	Zip code	Issue1 (I ₁)	Issue2 (I ₂)
t1	2502	Bob	20	3000	a	w
t2	2353	Ken	25	3500	b	z
t3	2453	Peter	25	4000	d	x
t4	1564	Sam	30	6500	a	x
t5	5021	Jane	35	4500	b	y
t6	9432	Linda	40	5500	a	y
t7	5024	Alice	45	6000	c	z
t8	1304	Mandy	50	5000	a	x
t9	1202	Tom	55	6500	c	w

Table 6 : Microdata Table MT

Various privacy models have been proposed in literature. We can categories it in to three main types, Statistical models, Partition-based anonymization models and Probabilistic models. Some of often models will be described in the following sections.

2.3.1 Statistical Methods

Some PPDP models use statistical methods to preserve individual privacy. In the following sections there will be a discussion about the randomization and swapping methods.

2.3.1.1 The Randomization Method

The randomization method has emerged as an important approach for data disguising in Privacy-Preserving Data Publishing (PPDP). It uses data distortion methods in order to create private representations of the records [32, 33]. The randomization method adds noise to the sensitive data so the participants' records are anonymized and at same time it preserves statistical information such as average or mean values. In most cases, it's possible to reconstruct aggregate answers from the data distribution by subtracting the noise from the noisy data, however participant records cannot be recovered. The randomization method could be classified in to two main classes;

- Random Perturbation method, which creates anonymized data by randomly perturbing the attribute values.
- Randomized Response method, which samples anonymized data from a probability distribution, given that the added noise is drawn from a fixed distribution.

Work in [34] showed that the addition of public information makes the randomization method vulnerable in unexpected ways. Moreover the randomization method is unable to guaranty privacy in the high dimensional case.

2.3.1.2 Swapping Method

Data swapping is to anonymize a dataset by exchanging values of sensitive attributes among data tuples [35]. It provides protection from identity disclosure and it's a value-invariant technique. Data swapping perfectly maintains univariate statistics and partially maintains lower-order multivariate statistics [36]. It can be used to preserve privacy for both numerical attributes and categorical attributes. Data protection level depends on the anonymization level induced in the data. Predefined criteria needed to specify tuples or values to be swapped. Often, a most rare tuples cause more data disclosure risk, therefore swapping method is commonly applies in this case. The key point is to find a suitable data

swapping algorithm which preserves released data as well as preserves dataset statistics. Data swapping method is done globally or locally. Globally swapping causes high impact on data utility, while locally or rank-based data swapping causes high error rates for aggregate queries. [37] work showed an example of privacy breach when an adversary has a prior belief on a unique attribute.

2.3.2 Partitions-Based Anonymization

Many models are designed to prevent disclosure of sensitive information by dividing data into groups of anonymous records. k -anonymity, ℓ -diversity, t -closeness and other models will be discussed in the following sections.

2.3.2.1 The k -anonymity Model

The basic idea of k -anonymity is to reduce the granularity of representation of the quasi-identifier attributes such a way each record contained in the released data cannot be distinguished from at least $k-1$ participants whose information also appears in the released data [13].

k -anonymity firstly removes explicit-identifier attributes, and then suppresses, generalizes or bucketizes quasi-identifier attributes. k -anonymity thus prevents quasi-identifier linkages. At worst, the data released narrows down an individual entry to a group of k individuals. Unlike randomization models, k -anonymity assures that the data released is accurate. Many methods have been proposed for achieving k -anonymity. In addition proposed methods use many mechanisms as suppression, generalization and bucketization to represent anonymized data.

Suppression mechanism:

It refers to replace certain attribute with the most general value, which means not releasing a value at all. Table 7 shows a released table satisfy 2-anonymity. t_1 and t_8 has been totally suppressed which means totally data loss. For t_2 and t_3 the zip code attribute has been suppressed. In t_4 and t_9 the age attribute has been suppressed. There are many suppression types like:

- Tuples suppression: one or more tuples will be suppressed. It's useful for outlier tuples.
- Cell suppression: one or more cells will be suppressed, where a cell represents an attribute value for a tuple.

- Attribute Suppression: one or more attributes will be suppressed. It's often used to suppress the explicitly identifier attributes.

Work [13] showed a model which combines generalization and suppression to achieve k-anonymity.

Tuple ID	(QI)		(SA)	
	Age	Zip code	Issue1 (I ₁)	Issue2 (I ₂)
t1	*	*	*	*
t2	25	*	b	z
t3	25	*	d	x
t4	*	6500	a	x
t5	*	*	b	y
t6	*	*	a	y
t7	*	*	c	z
t8	*	*	*	*
t9	*	6500	c	w

Table 7 : Suppression mechanism

Generalization mechanism:

It refers to replace a value with a less specific value based on a predefined domain hierarchy trees. For instance generalize Age value 35 to Age range of values [30-45]. Table 8 represents a released table satisfying 3-anonymity. There are 3 identical tuples for each quasi-identifier. Using the hierarchy tree in Figure 3 (a), the age value for t1 and t2 have been generalized from 20 and 25 values to range of values 2* which equivalent to [20 - 29] and the (*) icon means all possible values in its position. After generalizing some attribute values, the set of quasi-identifier (QI) attributes (age and zip code) of tuples t1 and t2 become identical. Each group of tuples that have identical QI attribute values is called an equivalence class.

Figure 3 (b) represents a range-based example of constructing a hierarchy tree. Generalization is created by generalizing all values in an attribute to a specific level of generalization. Obviously more generalization decreases data utility therefore a generalization mechanism must generalize the data not more than needed.

Attribute Generalization: It is applied at the level of column. When we perform generalization on column, it generalizes all values which belong to that column.

Cell Generalization: We can perform generalization on any particular cell of any attributes rather than whole column. Using this we can generalize only those cells that need generalization. Disadvantage of this approach is that it will increase complexities to manage values which are generalized at various levels.

Tuple ID	(QI)		(SA)	
	Age	Zip code	Issue1 (I ₁)	Issue2 (I ₂)
t1	[20-25]	[3000-4000]	a	w
t2	[20-25]	[3000-4000]	b	z
t3	[20-25]	[3000-4000]	d	x
t4	[30-45]	[4500-6500]	a	x
t5	[30-45]	[4500-6500]	b	y
t6	[30-45]	[4500-6500]	a	y
t7	[45-55]	[5000-6500]	c	z
t8	[45-55]	[5000-6500]	a	x
t9	[45-55]	[5000-6500]	c	w

Table 8 : Generalization mechanism

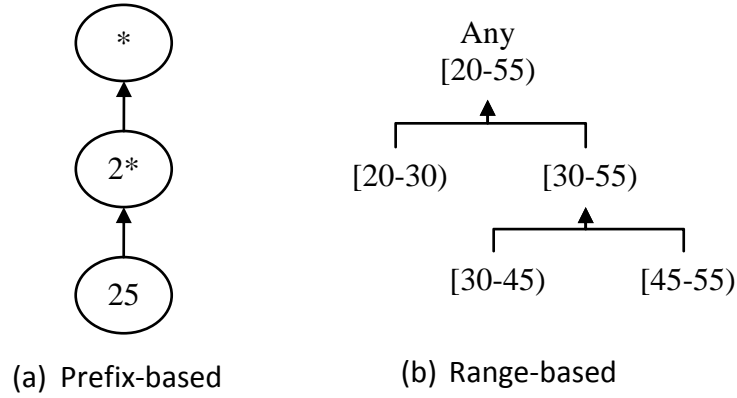


Figure 3 : Age generalization tree

Bucketization mechanism:

Instead modifying QI attributes and sensitive attributes, it divides the tuples into non-overlapping groups (buckets) and assigns a GID for each group. Then it publishes two tables, the first table with QI and the corresponding group GID and the second table with sensitive attributes and the corresponding group GID. Here each group works as a quasi-identifier and the sensitive attribute value of any participant would not be distinguish from any other participant in the same group. Table 9 shows two tables as result of bucketization mechanism. The first table represents QI tuples and the second represents SA tuples. However bucketization mechanism suffers from membership disclosure. Adversary can use the QI from the first table to check if a certain participant in this data.

Tuple ID	Age	Zip code	GID
t1	20	3000	1
t2	25	3500	1
t3	25	4000	1
t4	30	6500	3
t5	35	4500	2
t6	40	5500	2
t7	45	6000	3
t8	50	5000	2
t9	55	6500	3

GID	Issue1 (I ₁)
1	a
1	b
1	d
3	a
2	b
2	a
3	c
2	a
3	c

(a) QI table
(b) SA table

Table 9 : Bucketization mechanism

However k-anonymity does not provide full privacy due to the lack of diversity in the SA values (Homogeneity Attack) and if the adversary has additional background knowledge (Background knowledge Attack) these attacks will be discussed in details in section 2.3.3.2.

2.3.2.2 ℓ -diversity Model

ℓ -diversity is an effective model to remedy k-anonymity drawbacks. It's not only preventing identification of a tuple but also it preventing inference of the sensitive values of the attributes of that tuple. The ℓ -diversity model for privacy requires that there are at least ℓ "well-represented" values of sensitive attributes in each equivalence class. Work [3] presented a number of different instantiations for the ℓ -diversity definition which differ the meaning of being "well-represented". Simply it can mean ℓ distinct values. Table 10 (b) shows a released table satisfies 2-diversity. There are three groups where t1, t2 and t3 are in the same group and have identical QI values. In each group, there are at least two distinct SA values.

Tuple ID	(QI)		(SA)	(QI)		(SA)
	Age	Zip code	Issue1 (I ₁)	Age	Zip code	Issue1 (I ₁)
t1	20	3000	a	[20-25]	[3000-4000]	a
t2	25	3500	b	[20-25]	[3000-4000]	b
t3	25	4000	d	[20-25]	[3000-4000]	d
t4	30	6500	a	[30-40]	[4500-6500]	a
t5	35	4500	b	[30-40]	[4500-6500]	b
t6	40	5500	a	[30-40]	[4500-6500]	a
t7	45	6000	c	[45-55]	[5000-6500]	c
t8	50	5000	a	[45-55]	[5000-6500]	a
t9	55	6500	c	[45-55]	[5000-6500]	c

Table 10 : ℓ -diversity model

However as shown in Table 10 the third group t7, t8 and t9 has two sensitive attributes where (c) value more frequent than (a) value. Therefore distinct ℓ -diversity cannot prevent

probabilistic inference attacks. Moreover ℓ -diversity does not consider semantic meaning of SA values therefore it cannot prevent similarity attack.

2.3.2.3 t -closeness Model

t -closeness model [4] bounds distance between the distribution of a sensitive attribute in any equivalence class and the distribution of a sensitive attribute in the overall dataset by a predefined threshold t . t -closeness model can prevent skewness attack (will be discussed in Section 2.5.4). Consider a voter table where 90% of tuples have (c) SA value and 10% of tuples have (a) SA value. Assume that we released a table satisfies 2-diversity. This group has 50% of (c) and 50% of (a). However, this group presents a serious privacy risk because any tuple in the group could be inferred as having (a) with 50% confidence, compared to 10% in the overall table. Such attack called skewness attack and t -closeness model can prevent it. The Earth Mover's Distance (EMD) method [38] is used in order to quantify the distance between the two distributions of SA values. Many distance metric methods have been proposed. Kullback-Leibler, Weighted-Mean-Variance and Chi Square but these don't take into account ground distance (semantic distance), but EMD considers it. The EMD is based on the minimum amount of work required to transform one finite distribution into another one by moving distribution mass between each other [39].

Due to [40] the EMD function cannot prevents attribute linkage on numerical sensitive attributes. Moreover t -closeness forcing all released groups to have close distribution to the original data which negatively affects the data utility. Also t -closeness generalizes each attribute independently which causes loss correlation between different attributes [41].

2.3.2.4 Other Models

δ -Presence: work [42] presented δ -Presence metric to prevent table linkage threat. It concerns the case where a participant presence in the database causes a serious privacy risk. δ -Presence bounds the probability of inferring the presence of any participant within a range $\delta = (\delta_{\min}, \delta_{\max})$.

Personalized Privacy: work [43] presented personalized privacy metric to allow each participant to specify her/his own privacy level based on a predefined taxonomy tree for SA. For example a participant may be does not mind if others know that she/he have been voted positively/negatively for a certain topic. A table satisfies personalized anonymity with a

certain threshold if no adversary can infer the privacy requirement of any tuple with a probability above the threshold.

(X, Y)-Linkability, (α , k)-Anonymity, LKC-Privacy and more proposed to give a general privacy preserving.

2.3.3 Probabilistic Model

Recently some probabilistic models [44-47] are designed to prevent disclosure of sensitive information by providing ability to statistical queries. ϵ -differential, (c, t)-isolation and (d, γ)-privacy will be discussed in the following sections.

2.3.3.1 Differential Privacy

As an alternative of the partition-based models, differential privacy allows only statistical queries like sum or count queries. [46] proposes ϵ -differential privacy model to preserve privacy. It shows that the risk of addition or removal of a tuple doesn't affect the released data privacy. Consequently the computations will be insensitive to any changes in any tuple. Moreover the adversary will gain nothing. A random function \mathcal{F} will be used to generate the data to be released, such that \mathcal{F} is not very sensitive to any tuple in the data set. Formally, A randomized function \mathcal{F} gives ϵ -differential privacy if for all data sets D and D' differing on at most a single user, and all $T \subseteq \text{Range}(\mathcal{F})$, where ϵ is a positive real constant.

$$\frac{\Pr[\mathcal{F}(D) \in T]}{\Pr[\mathcal{F}(D') \in T]} \leq \exp(\epsilon)$$

The key point is to add random noise to the queries answers so that the answer changes but not the overall statistics. Therefore more queries means more noise needed to be added. This noise depends on ϵ and the sensitivity of the function \mathcal{F} .

Differential privacy has two kind of interaction, non-interactive and interactive approaches. In the non-interactive approach all queries have to be known in advanced. After that a perturbed version of the data created. While the interactive approach answers only a sub linear number of queries [48]. In differential privacy model there is no assumption about adversary's belief or tuples dependency [49].

2.3.3.2 (c, t)-Isolation

An adversary may try to isolate or to eliminate a tuple (a participant) from a dataset. PPDP requires that, using released data and background information should not increase the adversary ability to isolate any tuple. Work [15] has proposed a privacy model (c, t)-isolation to prevent tuple isolation in a statistical database. Suppose a data set D has been anonymized and released. Let D has n tuples. Suppose those tuples are represented as points in a certain space, where p is a point in D space and q is a point in D' space. The adversary is able to know the q point. Let δ be the distance between p and q . Let $B(q, c\delta)$ is a ball of radius $c\delta$ around point q . Then the point q (c, t)-isolates point p if $B(q, c\delta)$ contains fewer than t points in the table. where c is an isolation parameter and t is a threshold of privacy. (c,t)-isolation can be viewed as a record linkage problem and is suitable for problems with numerical attributes.

2.3.3.3 (d, γ)-Privacy

Work [47] presented a probabilistic privacy model (d, γ)-privacy, which relates the adversary's prior belief $P(t)$ for a given tuple t , with the posterior belief $P(t|D)$ for the same tuple. (d, γ)-privacy shows that when the $P(t)$ is small, there is a reasonable trade-off between privacy and utility. The privacy definition requires that the posterior belief $P(t|D) \leq \gamma$ and $\frac{P(t|D)}{P(t)} \geq \frac{d}{\gamma}$.

2.4 Complexity of finding optimal k-anonymity

In [50] work, authors have considered the complexity of finding an optimal value of k which ensure the anonymity of tuples up to a group of size k , while minimizing the amount of information loss. They showed that optimal k -anonymization for multi-dimensional QI is NP-hard under the suppression model. Therefore to minimize the number of suppressed tuples, a greedy approximate model has been proposed. Two approximation algorithms were propose: the first algorithm runs in time $O(n^{2k})$ and achieves an approximation bound of $O(k \log k)$, the second algorithm runs in a polynomial running time. Recently many improved models has been proposed and showed an approximation bound of $O(\log k)$ [51]. In [18] work, authors point up that suppression model is a special case of generalization model; furthermore they show that k -anonymization is also NP-hard under generalization model.

Data recording is a way to achieve k -anonymity based on generalization. There are two kinds of recording: global-recording and local-recording. In global-recording, same value in an attributes must generalize to the same level. In local-recording, same value in an

attribute may generalize to different levels. Global-recording may cause a higher information loss than local-recording. For example, Table 11(a) shows a generalization for Age and Zip code attributes, where the first generalization in Table 11(b) is global-recording based and the second generalization in Table 11(c) is local-recording. It's clear that in Table 11(b) the tuples t7, t8 and t9 are generalized more than the corresponding tuples in Table 11(c).

In multi-dimensional generalization, recording may work in each attribute separately or mapping the Cartesian product of all attributes. Work [26] showed that applying recording process in the Cartesian product is more accurate than the separated manner. Most of recent research like [23, 52] proposed algorithms for one dimension and global-recording.

Specialization is the reverse operates of generalization. It is a top-down process, which starts from the most general value and dividing data based on predefined conditions.

Tuple ID	(QI)		(QI)		(QI)	
	Age	Zip code	Age	Zip code	Age	Zip code
t1	20	3000	[20-25]	[3000-4000]	[20-25]	[3000-4000]
t2	25	3500	[20-25]	[3000-4000]	[20-25]	[3000-4000]
t3	25	4000	[20-25]	[3000-4000]	[20-25]	[3000-4000]
t4	30	6500	[30-40]	[4500-6500]	[30-40]	[4500-6500]
t5	35	4500	[30-40]	[4500-6500]	[30-40]	[4500-6500]
t6	40	5500	[30-40]	[4500-6500]	[30-40]	[4500-6500]
t7	45	6000	[45-55]	[4500-6500]	[45-55]	[5000-6500]
t8	50	5000	[45-55]	[4500-6500]	[45-55]	[5000-6500]
t9	55	6500	[45-55]	[4500-6500]	[45-55]	[5000-6500]

Table 11: Global-recoding and local-recoding

In Chapter 5 a local-recording, multi-dimensional generalization algorithm will be presented.

2.5 Privacy-Preserving Data Publishing Possible Attacks

Many PPDP algorithms have been proposed in order to protect data after publishing it and at same time preserve maximum utility. However many attacks have been proposed to reveal participant privacy. One of the most cited example of this type of privacy breach is the AOL search data leak. In 2006, AOL researchers recently published the search logs of about 650,000 members. The release intended for research purposes. Unfortunately, AOL did not notice that users' searches may potentially identify individual users. Using search engines to find an individual's name, address or a telephone number, could then leads to a specific individual. The release replaced users' names with persistent pseudonyms. It did not take much inspecting for The New York Times to conclude that searched words belong to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga. [53] In the next section the often privacy attacked will be discussed.

2.5.1 Linking Attack

Simply removing Explicitly-identifier (EI) attributes not enough. Using linking attack; an adversary still be able to identify individual participant by linking external data to anonymized data [13].

k-anonymity model provide a solution to avoid linking attacks. It requires that each record in the released data is identical to at least k-1. Table 12 shows an example how the adversary compares QI values in the anonymized table (a) and public data (b). It obvious that t1 has the same QI values in both tables which conclude with high probability that they are the same participant.

Tuple ID	(QI)		(SA)		(EI)	(QI)		
	Age	Zipcode	Issue1	Issue2		SSN	Name	Age
t1	20	3000	a	w	2502	Bob	20	3000
t2	25	3500	b	z	1304	Mandy	50	5000
t3	25	4000	d	x	1202	Tom	55	6500
t4	30	6500	a	x	1564	Sam	30	6500

(a) anonymized data (b) Public data

Table 12 : Linking Attack

2.5.2 Homogeneity Attack

Appears when an anonymous data groups lack of diversity. Table 13 shows a 2-anonymity anonymized data, for the first group there are two tuples with same SA value. Therefore an adversary can easily reveal participant's privacy. k-anonymity requires each tuple in anonymized data to appear at least k times, but does not say anything about the SA values. If a SA values in a QI group are same then it violate privacy requirements.

ℓ -diversity suggests that as improvement to k-anonymity, the anonymized groups should diverse the SA values for each QI attribute.[3]

(QI)		(SA)
Age	Zip code	(I ₁)
[20-30]	[1200-3400]	c
[20-30]	[1200-3400]	c
[30-40]	[5600-6600]	a
[30-40]	[5600-6600]	b

Table 13 : Homogeneity and background knowledge attacks

2.5.3 Background Knowledge Attack

An adversary has background knowledge about the SA values. For example if the adversary knows that certain city supports certain party with very high confidence. In Table 13 if the city has the zipcode 6600 and support choice (a) and appears in the second group, then the adversary can concludes that participants from city with zipcode 5600 has been voted for I₁ by (b). k-anonymity does not protect against background knowledge attack.

ℓ -diversity provides a solution by increasing the diversity of SA values for each anonymized group.

2.5.4 Skewness Attack

Adversary can reveal participants privacy if anonymized groups have a non-uniform distribution of SA values. ℓ -diversity model prevents direct attribute disclosure; however it doesn't provide a sufficient distribution for sensitive attribute values. Table 14 shows an anonymized data which satisfies 2-diversity. The SA values include four (a) values and one (b) value. This implies that the participants have been voted for (a) choice by probability 80%. This type of privacy threats called Skewness attack. t-closeness [4] model provides a solution for this attack. It bounds the data set distribution distance between the distribution of SA values in the original data set and the released data for each group.

(QI)		(SA)
Age	Zip code	(I ₁)
[20-30]	[1200-3400]	a
[20-30]	[1200-3400]	a
[20-30]	[1200-3400]	a
[20-30]	[1200-3400]	a
[20-30]	[1200-3400]	b

Table 14 : Skewness attack

2.5.5 Similarity Attack

Participant's privacy may be at risk if sensitive attribute values of anonymized groups are similar. An anonymization algorithm must consider semantic meanings of SA values. In public opinion polls such attack is rarely happen. Table 15 shows a 3-diversity anonymized data. Assume choices (a) and (b) have closed mining (first opinion) and choice (f) has a totally opposite meaning (second opinion). Then the similarity between (a) and (b) will implies that voters have choose 80% of the first opinion.

(QI)		(SA)
Age	Zip code	(I ₁)
[20-30]	[1200-3400]	a
[20-30]	[1200-3400]	b
[20-30]	[1200-3400]	a
[20-30]	[1200-3400]	b
[20-30]	[1200-3400]	f

Table 15 : Similarity attack

2.5.6 Membership Disclosure

An adversary can discover whether a participant presence in the released data or not. People have the right to hide their participation in any public opinion process. Bucketization mechanism for instant does not prevent this attack as we mentioned in 2.3.2.1. Generalization and slicing mechanisms [54] prevent membership attack.

2.5.7 Multiple Release Attack

A microdata often has to perform many operations for its tuples. Insertions, deletions and updates operations may leads to republishing a new anonymized version. However multiple releases open to be linked together which may compromise data privacy (will be discussed in more details in chapter 4). A suggested solution is to consider all of the released data before publishing the new one. But it's not always the case. Data publisher may not notice that another release may happen in future; also other data holders are able to release some data. Table 16 shows a 3-anonymity and 2-diversity for the first release R1 and the second release R2. Assume an adversary knows that a voter presented in both releases and she/he is 40 years old and living in a city with 3000 zipcode. Examining R1 and R2 together, the adversary can eliminates r1, r2 and r6 tuples. Also the adversary can eliminate tuple r3 or r4 due to the distribution of SA values in R1.

Preventing such attacks, called Multiple Release or correspondence attacks, needs to consider all changes occurred to the data moreover to consider the anonymization models used in previous releases [55-58].

TID	(QI)		(SA)
	Age	Zipcode	Issue1
t1	[20-40]	30**	a
t2	[20-40]	30**	a
t3	[20-40]	30**	b

(a) Release 1 (R1)

TID	(QI)		(SA)
	Age	Zipcode	Issue1
r1	2*	3***	a
r2	2*	3***	a
r3	4*	3***	b
r4	4*	3***	b
r5	4*	3***	a
r6	2*	3***	b

(b) Release 2 (R2)

Table 16 : Multiple release attack

2.5.8 Minimality Attack

In addition to background knowledge and anonymized data, adversaries may have access to algorithms used to anonymize data. Based in this knowledge, work [59] presented a minimality attack which may be used by adversaries to breach participants' privacy. Using a probabilistic formula, an adversary eliminates impossible cases in order to launch elimination attack. In general, the minimality principle state that a generalization algorithm should not synthesized data more than its necessary to achieve its requirement.

2.5.9 Inference Attack

Inference attack occurs when an adversary is able to infer a sensitive data with high confidence. The adversary deduces the sensitive data using trivial information. Even if the QI is not fully released it may be possible to infer missing QI values from other information. It's possible to infer gender or religion from name, birth year from graduation year [60]. Several works [61-63] have proposed solutions for inference attack.

2.5.10 deFinetti Attack

Using a statistical theorem known as deFinetti's theorem, [64] showed that using a group of anonymized data deFinetti attack can build a classifier to predict the SA value associate with this group. Firstly, deFinetti attack guessing a random permutation for each QI in order to assign a SA value to each tuple. This step produces a set of conditional distributions as following $\Pr(QI_i|SA_j)$ where i is $|QI|$ and j is the number of distinct SA values in each QI. The set of conditional distributions can be described as a classifier. Secondly, in each QI deFinetti attack uses the classifier to check the relative likelihood for the guessed permutation. Iteratively the process will construct a precise classifier. Thirdly, Anonymized data and the constructed classifier can be used to reveal participant privacy. [65].

As showed in [64] deFinetti attack can be used against any model uses tuple-independent model, the random-worlds method [66], or independent and identically distributed model.

2.6 INFORMATION LOSS METRICS

As mentioned in Chapter 1, one of the primary goals of data publishing utility is the quality of the released dataset. Unfortunately, de-identification of datasets degrades the utility of the dataset giving us a trade-off between privacy and utility. A good anonymization not only satisfies the underlying privacy standard but also minimizes information loss due to generalizations. To achieve such an anonymization, we first need a metric to measure the level of utility of a given anonymized dataset. A typical utility metric measures the data quality in the released data with respect to the data quality in the original dataset. This chapter describes three commonly used information loss metrics:

2.6.1 Discernibility Metric

The Discernibility Metric (DM) penalizes by a value of the anonymized group size each unsuppressed tuple and assigned a penalty of the input dataset size for each suppressed tuple. In addition each suppressed tuple incurs a cost $|T|$. Given that we did not perform any tuple suppression then the DM error is the normalized sum of all assigned penalties [23], [24]. The certainty loss is the sum of intervals size on all attributes of the generalized tuples [41]. DM can be mathematically stated as follows:

$$DM(T^*) = \sum_{t \in T} |EC(t)|^2$$

Where t is a tuple from T and $EC(t)$ is the Equivalence Class of T^* indistinguishable from t . (EC defined in section 5.1)

2.6.2 Loss Metric

The information loss metric (LM) is the sum of all normalized information loss for each column in the anonymized table. LM is between 0 and 1, where 0 means no information loss and 1 means total information loss. We use the [67] definition, which described mathematically as follows: For attributes with numerical values, assume for each interval the lower and upper values be L and U respectively. The information loss can be calculated by

$$LM(T^*) = \sum_{t \in T} (U_i - L_i) / (U - L)$$

For attributes with categorical values, assume N is the total number of leaf nodes in R . and NP is the total number of leaf nodes in a sub tree rooted in P . consider the generalization based on the domain generalization structure \mathcal{R} (as shown in Figure 3).

2.6.3 Average Query Error

The *Average Relative Error* (AvRE) [68] measures the distortion by comparing the counts of the randomly generated SQL queries over T^* to the counts over T . AvRE for each tuple can be measured as $|act - est|/act$. Where *act* means the actual results driven from the original data and *est* means the estimated results driven from the sanitized data.

In the chapter 6 we will show a comparison between MSA-diversity model and Gal's et al model in term of the Discernibility Metric, Loss Metric and *Average Relative Error*.

3. PRIVACY-PRESERVING FOR MULTIPLE SENSITIVE ATTRIBUTES

Typical public opinion polls cover a very large number of issues. Each issue is recognized as a sensitive attribute. Therefore we will deal with multiple sensitive attributes in order to develop an efficient model. The data holder needs to publish a large number of opinions about many issues. Adversaries use participants' profiles and public opinion to launch an attack. The main idea is to protect and secure the relationship between the quasi-identifiers and multiple sensitive attributes. Many works like k-anonymity, ℓ -diversity, t-closeness, etc., have been proposed as a privacy protection model for microdata [3], [4]. However, most of models assume there is one single sensitive attribute in the microdata table.

Tuple ID	Explicit Identifiers (EI)		Quasi Identifiers (QI)		Sensitive Attributes (SA)				
	SSN	Name	Age	Zip code	I ₁	I ₂	I ₃	...	I _d
t1	2502	Bob	20	3000	a	w	e		k
t2	2353	Ken	25	3500	b	z	e		m
t3	2453	Peter	25	4000	d	x	f		k
t4	1564	Sam	30	6500	a	x	e		k
t5	5021	Jane	35	4500	b	y	g		n
t6	9432	Linda	40	5500	a	y	f		l
t7	5024	Alice	45	6000	c	z	f		m
t8	1304	Mandy	50	5000	a	x	h		l
t9	1202	Tom	55	6500	c	w	g		n

Table 17 : Microdata Table with (d) SA

Table 17 shows an example for a microdata with multiple sensitive attributes. I₁, I₂, to I_d are opinions for different issues. Each issue has a number of distinct choices and a

participant can choose one. Issues can be related to each other, and have a strong dependency. In this case their joint distribution will be similar; therefore we can consider them as one issue. Also issues can be independent from each other, thus their joint distribution are different. In this case we consider them as different issues.

At present, few multiple sensitive attributes models have been proposed in the literature to prevent re-identification risks caused by external knowledge. In the following subsections will discuss it.

3.1 Naïve Approach

It has been shown in [11] [12], [25], that under non-membership information ℓ -diversity fails to protect privacy. Simply using ℓ -diversity for MSA will cause privacy breach.

As an example, Table 18 (i) shows some voter's records. The anonymization in Table 18 (ii) satisfies 3-diversity on I_1 alone and I_2 alone. Consider an adversary who has the background knowledge that Amy will not vote for (c) on I_1 , thus the adversary can exclude the tuples with (c) on I_1 . Since the remaining tuples all have (w) on I_2 , the adversary will conclude Amy has voted (w) on I_2 .

Tuple ID	(QI)		(SA)	
	Age	Zip code	I_1	I_2
Amy	30	1200	b	w
Bob	20	2400	c	x
Che	23	1500	a	w
Dina	27	3400	c	y

i. Microdata

Tuple ID	(QI)		(SA)	
	Age	Zip code	I_1	I_2
Amy	[20-30]	[1200-3400]	b	w
Bob	[20-30]	[1200-3400]	c	x
Che	[20-30]	[1200-3400]	a	w
Dina	[20-30]	[1200-3400]	c	y

ii. Anonymized data

Table 18 : The microdata and anonymized data sample T

3.2 Machanavajjhala’s et al. Approach

According to work by [3], if we have two sensitive attributes, the main idea is to treat the first sensitive attribute as part of the quasi-identifier when checking for diversity in the second sensitive attribute (and vice versa). Thus we can ensure the diversity principle is held for the entire dataset. However, this solution is impractical for use in public opinion data because of the huge number of opinions each participant is expected to express. Also the sensitive attribute which treated as part of QI may be generalized, which will rise the information utility loss.

Tuple ID	(QI)			(SA)
	Age	Zip code	I ₁	I ₂
Amy	30	1200	b	w
Bob	20	2400	c	x
Che	23	1500	a	w
Dina	27	3400	c	y

i. Treat I₁ as a part of QI

Tuple ID	(QI)			(SA)
	Age	Zip code	I ₂	I ₁
Amy	30	1200	w	b
Bob	20	2400	x	c
Che	23	1500	w	a
Dina	27	3400	y	c

ii. Treat I₂ as a part of QI

Table 19 : Microdata of Machanavajjhala et al. approach for MSA

Consider the row microdata in Table 19. As first phase, suppose I₁ treated as a part of QI, then checking the diversity of I₂ will produce the anonymized data in Table 20(i). Second phase treat I₂ as part of QI and check I₁ diversity. As result shown in Table 20(ii) to produce 2-diversity data, this approach generalized the QI and SA which causes utility loss.

Tuple ID	(QI)			(SA)
	Age	Zip code	I ₁	I ₂
Amy	[20-30]	[1200-3400]	[a,b]	w
Che	[20-30]	[1200-3400]	[a,b]	w
Bob	[20-30]	[1200-3400]	c	x
Dina	[20-30]	[1200-3400]	c	y

i. First phase

Tuple ID	(QI)		(SA)	
	Age	Zip code	I ₂	I ₁
Amy	[20-30]	[1200-3400]	w	[a,b]
Che	[20-30]	[1200-3400]	w	[a,b]
Bob	[20-30]	[1200-3400]	[x,y]	c
Dina	[20-30]	[1200-3400]	[x,y]	c

ii. Second phase (2-diversity)

Table 20 : Anonymized data of Machanavajjhala et al. approach for MSA

3.3 Li and Ye Approach

Work by [25] provided a two-step greedy generalization algorithm, which is used to carry out the multiple sensitive attributes processing. First phase: quasi-identifiers are generalized using a top-down specialization greedy algorithm. It starting with the whole data set as a single group and then trying to split it to smaller groups until further split will violate α -QI condition, where α -QI is the diversity requirement for QI. α -QI is predefined by data holder. Second phase: sensitive attributes are masked (generalized) using a bottom-up local recording algorithm. It checks the α -SA condition for each equivalence class - which constructed in the first phase-, where α -SA is the diversity requirement for SA. α -SA is predefined by data holder.

However, in the public opinion case we have few choices for each sensitive attribute which leads to a huge information loss if we apply the masking step of this solution. Moreover [25] doesn't construct groups in a probabilistic manner, which may leads to have one SA value more frequent than other SA values within a group. For example, Table 21 shows the process to satisfy α -QI = α -SA = 2. In the first phase there are two equivalence classes. In the second phase due to the local generalization step there are suppressions to two values. w value in the first equivalence class will be suppressed and c value in the second equivalence class will be suppressed.

Tuple ID	(QI)		(SA)		(QI)		(SA)	
	Age	Zip code	I ₁	I ₂	Age	Zip code	I ₁	I ₂
Amy	30	1200	b	w	[23-30]	[1200-1500]	b	*
Che	23	1500	a	w	[23-30]	[1200-1500]	a	*
Bob	20	2400	c	x	[20-27]	[2400-3400]	*	x
Dina	27	3400	c	y	[20-27]	[2400-3400]	*	y

i. First phase ii. Second phase

Table 21 : Li and Ye approach

3.4 Gal's et al Model

Work by [11] mentions that the table T satisfies both k-anonymity and ℓ -diversity if T is divided into a partition and each group contains at least k records, and to delete all rows in the group, at least ℓ distinct values need to be deleted to delete all rows in the group. T is also anatomized or generalized. However this work isn't the most appropriate approach to this problem. [11] model works top down, it starting with the whole data set as a single group and then trying to split it to smaller groups until further split will violate k-anonymity and ℓ -diversity conditions.

Using dataset in Table 4, as shown in Figure 4 it starts by select the attribute with widest normalized range. In this example it's possible to choose Zip code or Age. By choosing the Age attribute and split it in to two groups 3000-5000 and 5500-6500 look to (Figure 4). Then we will have t1, t2, t3, t5 and t8 in G1 and t4, t6, t7 and t9 in G2. Even the first part t1, t2, t3, t5 and t8 can construct G1 the second part cant construct any group. Therefore it's only possible to make one partition and exclude the rest.

It's also possible choosing Age attribute to split the data to two groups. As results [11] will give us two groups t1, t2 and t3 as G1, t6, t7, t8 and t9 as G2, t4 and t5 will be excluded. Table 22 shows the anonymized data.

Tuple ID	(QI)		(SA)	
	Age	Zip code	I ₁	I ₂
t1	20-25	3000-4000	a	w
t2	20-25	3000-4000	b	z
t3	20-25	3000-4000	d	x
t6	40-55	5000-6500	a	y
t7	40-55	5000-6500	c	z
t8	40-55	5000-6500	b	x
t9	40-55	5000-6500	c	w
t4	*	*	*	*
t5	*	*	*	*

Table 22 : Gal's et al released data T*1

Figure 4 shows a 2D space, where the x-dimension is Zipcode And y-dimension is Age. Each cell indicates a tuple in the dataset of Table 4. Dashed line represent [11] partitioning process. Assuming normal distribution, the probability that a tuple with age = 25 and Zipcode = 4000 is 1/64. Where 64 is the total number of data cells. Clearly [11] model causes a large information loss. [11] model focuses on both k-anonymity and ℓ -diversity anonymization. However, [3] work mentions the k-anonymity drawbacks and how it fails to preserve privacy practically. k-anonymity still can't prevent homogeneity attack and background knowledge attack. Therefore focusing on ℓ -diversity will provide stronger privacy preserving.

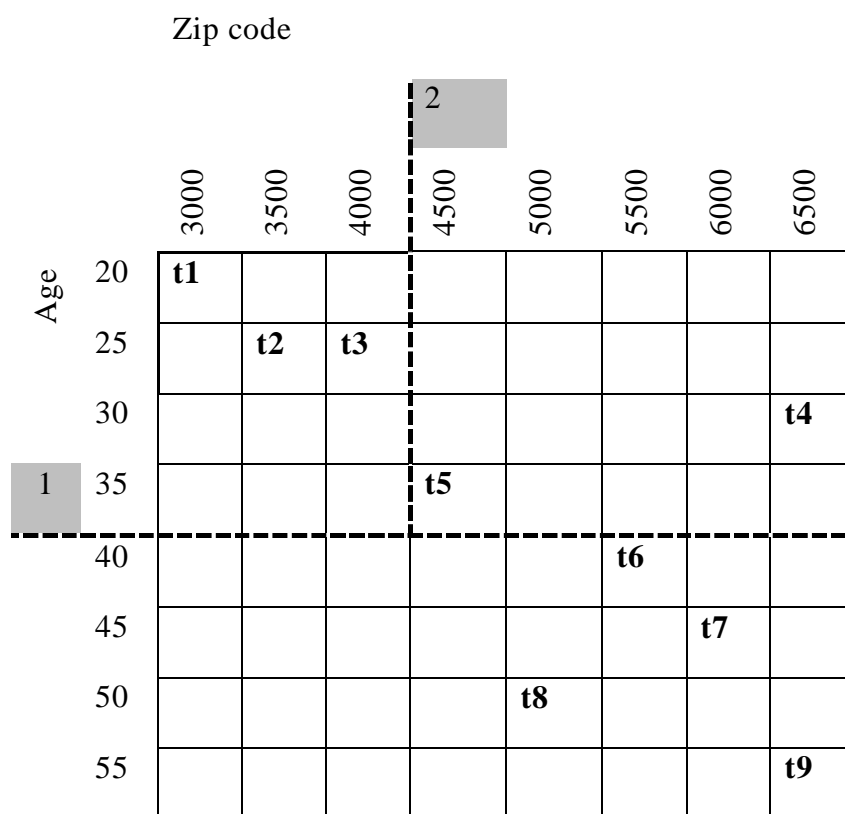


Figure 4 : 3-diversity groups using Gal's et al. model

3.5 Xiao-Chun et al Model

Work by [12] shows a multi-dimensional bucket grouping for SA with multiple attributes. Table 24 shows the anonymized data of Table 23 based on the maximal multi-dimension-capacity first algorithm. However it's clear that [12] can't prevent membership attack. Furthermore it doesn't present a probabilistic model. For example, the G3 in Table 24 has the probability that Mandy (t8) votes for z on issue2 is 2/3. This makes it difficult to make risk/benefit/cost analysis of publishing private data.

Tuple ID	Quasi-Identifier (QI)		Sensitive Attributes (SA)	
	Age	Zip code	Issue1 (I ₁)	Issue2 (I ₂)
Bob(t1)	20	3000	a	f
Ken(t2)	25	3500	a	w
Peter(t3)	25	4000	a	x
Sam(t4)	30	6500	b	f
Jane(t5)	35	4500	b	w
Linda(t6)	40	5500	c	y
Alice(t7)	45	6000	c	y
Mandy(t8)	50	5000	d	z
Tom(t9)	55	6500	e	f

Table 23 : Microdata for Xiao-Chun et al. model

Tuple ID	(QI)			(SA)	
	Age	Zip code	Group ID	Group ID	I ₁ , I ₂
t1	20	3000	G1	G1	c, y
t2	25	3500	G2	G1	b, w
t3	25	4000	G3	G1	a, f
t4	30	6500	G2	G2	b, f
t5	35	4500	G1	G2	a, w
t6	40	5500	G1	G2	c, y
t7	45	6000	G2	G3	e, z
t8	50	5000	G3	G3	d, z
t9	55	6500	G3	G3	a, x

Table 24 : Anonymized data of Xiao-Chun et al (MMDCF algorithm)

3.6 Ye et al Model

Work by [69] proposes decomposition model for MSA privacy preserving problem. Basically, it decomposes the dataset into SA-groups. For example, Table 25 (ii) shows an anonymized data for (i) table. For each group of tuples, instead of generalizing QI attributes it composes or bucketizes the SA values. firstly, [69] model constructs tuples equivalence classes based on ℓ -diversity model and using one sensitive attribute each time. However this model can't prevent membership attack, all tuples are released as it is without anonymization process. Additionally, it doesn't provide a probabilistic model; therefore a tuple may be more likely to appear than other tuples within same equivalence class.

Tuple ID	(QI)		(SA)		(QI)		(SA)	
	Age	Zip code	I ₁	I ₂	Age	Zip code	I ₁	I ₂
Amy	30	1200	b	w	30	1200	b, c	w, x
Bob	20	2400	c	x	20	2400		
Che	23	1500	a	w	23	1500	a, c	w, y
Dina	27	3400	c	y	27	3400		

i. Microdata

ii. Anonymized data

Table 25 : Ye et al model example

3.7 Fang et al Model

Work by [70] provides a new model, CODIP, as a privacy preserving model for data with multiple sensitive attributes. CODIP projects the microdata on to SA groups, where each group satisfies t-closeness or any other anonymization algorithm. The anonymized data will be the projected data, where projected data isolates disjoint SA groups in separate tables. All tuples in the anonymized data will have a randomize order. However CODIP prevents intersection and minimality attacks it still doesn't avoid attacks based on probability.

4. PRIVACY-PRESERVING FOR DYNAMIC RELEASES

Recent research has been devoted to study the privacy preserving for multiple data publishing [28, 29, 44, 45, 57, 58, 61, 71, 72]. They refer to this kind of study as sequential releases [28, 29, 71], serial data publishing [44, 55], dynamic anonymization [58, 72] and multiple data releases depending on the difference between data releases.

The dynamic data releases occur when data holder needs to publish new information for same data, another instance for same data or/and updated data. For example, data holder published data for an issue 1 (R1), another issue 2 needed to be published therefore dataset holder makes a new release (R2). These releases may differ with quasi-identifier attributes /sensitive attributes or/and with period of time. The first release R1 released at timestamp1. The second release R2 will be released at timestamp2, and so on so forth, where timestamp1 < timestamp2. Moreover, when anonymizing R2 we cannot modify R1; simply R1 becomes part of history. In public opinion polls privacy preserving area, polls may organize in a different periods of time. One may suggest accumulating all releases, and then anonymizing it together. The first public opinion issue will postponed until the second, third, or more ones are collected. This idea is a time consuming task. Its not be acceptable in public opinion problem.

k-anonymity, ℓ -diversity, t-closeness, etc [3], [16], [17], [18], [19], have provided a number of valuable privacy-protecting techniques. However, they are only deal with one-time data release (static- release). This implies a significant limitation, as in many applications data have many releases which collected continuously.

Simply using k-anonymity or other static-release approaches to anonymize the new releases independently without considering the previous releases may cause privacy violations. The relations between QI and SA in the releases data will give adversaries great opportunities to reveal individuals' privacy. Illustrations of these threats will be present through some examples in the following sections.

The key point is how to anonymize the current release so that it cannot be linked to previous releases, and still remains useful for its own release purpose.

Many types of dynamic data releases have been discussed, work by [73] discussed the sequential data releases when the sensitive attributes for new release doesn't has any relation with the previous release. In [73] a global guarantee (across all releases) has been declared to give more accurate privacy measurement than the local guarantee (for each release). Work [71] studied the problem of releasing different attributes' subsets for same microdata where Quasi-identifiers can be reconstructed from several releases. Work [58] defined a new model (m-invariance) for dynamic releases for same dataset with updated tuples. In [58] both insertions and deletions have been discussed. Both of [57] and [72] presented remedy models for m-invariance model. While [74] and [75] concerns only the insertion of new tuples.

4.1 SAs Independent Approach

ℓ -diversity ensures that the probability of mapping an individual to a sensitive value is bounded by $1/\ell$. Therefore, it guarantees that every equivalence class contains at least ℓ distinct SA values. Assuming there are no associations between issues SAs, which means that the first SA values are independent from the second SA values, and so on so forth.

As an example, Table 26 shows two voters' records where first issue SA doesn't has any relation with second issue SA. The anonymization in Table 27 (i) satisfies 3-diversity on I_1 and Table 27 (ii) satisfies 3-diversity on I_2 . Consider an adversary who has the background knowledge that Bob (t1) will vote for (a) on I_1 , thus the adversary can only exclude t1 tuple in the first group. Since there are no relations between the two SAs, the first group that includes t1 in Table 27 (ii) will not be affected. The adversary will not gain any new information and Table 27 will still satisfy 3-diversity on I_1 and I_2 .

Tuple ID	(QI)		(SA)
	Age	Zip	I ₁
Bob(t1)	20	3000	a
Ken(t2)	25	3500	b
Peter(t3)	25	4000	d
Sam(t4)	30	6500	a
Jane(t5)	35	4500	b
Linda(t6)	40	5500	a
Alice(t7)	45	6000	c
Mandy(t8)	50	5000	b
Tom(t9)	55	6500	c

i. First Issue T1

Tuple ID	(QI)		(SA)
	Age	Zip	I ₂
Bob(t1)	20	3000	w
Ken(t2)	25	3500	z
Peter(t3)	25	4000	x
Sam(t4)	30	6500	x
Jane(t5)	35	4500	y
Linda(t6)	40	5500	y
Alice(t7)	45	6000	z
Mandy(t8)	50	5000	x
Tom(t9)	55	6500	w

ii. Second Issue T2

Table 26 : The microdata of two independent issues

TID	(QI)		(SA)
	Age	Zip code	R ₁
t1	[20-25]	[3000-4000]	a
t2	[20-25]	[3000-4000]	b
t3	[20-25]	[3000-4000]	d
t4	[30-45]	[4500-6500]	a
t5	[30-45]	[4500-6500]	b
t7	[30-45]	[4500-6500]	c
t6	[40-55]	[5000-6500]	a
t8	[40-55]	[5000-6500]	b
t9	[40-55]	[5000-6500]	c

a) First Release R1

TID	(QI)		(SA)
	Age	Zip code	R ₂
t1	[20-25]	[3000-4000]	w
t2	[20-25]	[3000-4000]	z
t3	[20-25]	[3000-4000]	x
t4	[30-45]	[4500-6500]	x
t5	[30-45]	[4500-6500]	y
t7	[30-45]	[4500-6500]	z
t6	[40-55]	[5000-6500]	y
t8	[40-55]	[5000-6500]	x
t9	[40-55]	[5000-6500]	w

b) Second Release R2

Table 27 : The anonymized data for two independent issues

The non-membership attack doesn't work in the independent association case. Even if an adversary knows i bits of non-membership knowledge, he/she cannot link it with the next release SA.

However, in real life there are partially or fully associations among sensitive attributes. For instance, public opinion polls like "Do you approve or disapprove of Barack Obama's performance as president?" have been published five times during 2011 [62]. In this case, there are five polls where SAs are fully associated. Another partially association may also happens. In dynamic data releases it's not possible to deal with the SAs as one SA, simply when anonymizing the second release; it's not possible to change the first release. In the next section, techniques prepared to this type of sensitive values will present.

4.2 SAs Dependent Approaches

Correlations between sensitive attributes in sequentially released data may leads to serious disclosure scenarios. It may happen that one SA work as identifier to another SA. In public opinion polls, a poll question about head of a party performance and another question about the party performance have a high association. The following sections will represent the prepared approaches and possible attacks for SAs associations.

4.2.1 Record-linking Attack

Data holder release new data as it's available. Different releases perhaps for same dataset with different attributes which may yield to record-linking across multiple anonymous releases. For instance, first release may contain Age, Zip code and Issue1 attributes, second release may contain participant's name, Age and Zip code attributes. An adversary can launch a record-linking attack by joining the identical set of attributes from the two releases. In Table 28, the adversary may join the first tuple in Table 28 (i) with second tuple in Table 28 (ii) to conclude that Bob has voted by (a) for issue1.

(QI)		(SA)
Age	Zip code	I ₁
25	3500	b
20	3000	a
40	5500	a
45	6000	c

i. First release R1

	(QI)	
Name	Age	Zip code
Bob	20	3000
Ken	25	3500
Linda	40	5500
Alice	45	6000

ii. Second release R2

Table 28 : The microdata for Join Attack example

The work in [71] considers record-linking attack and provide a remedy for this type of dynamic releases definition. It presents the concept of (X, Y)-privacy as a top-down specialization approach to prevent record-linking attack.

4.2.2 Value-association Attack

Another definition for dynamic data release state that: the same data and same attributes may anonymize differently for different purposes. m-invariance model [58] supports dynamic data release in both new tuple insertions and deletions scenarios. Value-association attack happens if an adversary knows that a certain participant appears in both releases. Table 29 and Table 30 depict an anonymized data with two releases, where R1 is the first release and R2 is the second release. An adversary may know that Bob participates in R1 and R2 and his age is 20, then by looking to the association between the first group in R1 and R2, it is easy to deduce that he has (a) as SA value. The m-invariance model effectively limits the risk of privacy disclosure caused by this attack. It guarantees that each anonymized group has at least m tuples, each with a unique set of sensitive values. In order to make these groups m-invariance may insert some counterfeit tuples. Due to the sensitive values' consistency, a value-equivalence attack may be used to breach privacy; this attack will be presented in Section 4.2.4.

Name	QI		SA
	Age	Zip	
Bob	25	3500	a
Ken	20	3000	b
Linda	40	5500	a
Alice	45	6000	c

i. Microdata T1

QI		SA
Age	Zip	
[20-30]	[3000-4000]	a
[20-30]	[3000-4000]	b
[40-50]	[5000-6000]	a
[40-50]	[5000-6000]	c

ii. First release (R1)

Table 29 : First Release for the Value-association Attack

Name	QI		SA
	Age	Zip	
Bob	25	3500	a
<i>Peter</i>	23	4000	c
Linda	40	5500	a
<i>Mandy</i>	50	6000	b

i. Microdata T2

QI		SA
Age	Zip	
[20-30]	[3000-4000]	a
[20-30]	[3000-4000]	c
[40-50]	[5000-6000]	a
[40-50]	[5000-6000]	b

ii. Second release (R2)

Table 30 : Second Release for the Value-association Attack

4.2.3 Correspondence Attack

In dynamic data release, even if released data met the anonymization requirement; an adversary may focus on the groups' relation and correspondence between released data. Possible scenario as presented in Table 31 and Table 32: an adversary may know that Jane is in R1 and R2. Therefore in R1, Jane may has $\langle a, a, b \rangle$ as sensitive values, while in R2 she may has $\langle b, b, a \rangle$. This allow adversary to eliminate t1 or t2, also to eliminate r3 or r4. As result Jane has probability $\frac{1}{2}$ to has a (or b) as SA.

Without delaying tuples or inserting counterfeit tuples, [57] work presented an example of correspondence attack. In addition, it presents BCF-anonymity method as a new generalization method secured from correspondence attack. However, [57] work concerns only in the tuple insertion scenario. It doesn't provide a solution for new releases with updated tuple.

[61] Work mentioned a possible approach to anonymize only new tuples for new releases. In addition it also declared such approach will cause a low quality data. [61] work proposes incremental anonymization techniques for insertion scenario.

TID	QI		SA
	Age	Gender	
t1	35	*	a
t2	35	*	a
t3	35	*	b

Table 31 : An anonymized data for R1 in the Correspondence Attack

TID	QI		SA
	Age	Gender	
r1	[20-40]	M	a
r2	[20-40]	M	a
r3	[20-40]	F	b
r4	[20-40]	F	b
r5	[20-40]	F	a
r6	[20-40]	M	b

Table 32: An anonymized data for R2 in the Correspondence Attack

4.2.4 Value-equivalence Attack

The m-invariance model [58] provides protection against value-association attack. It keeps the same sensitive values for each group while changes the QI tuples. A serious privacy breach may appear if an adversary considers changes in QI tuples. Table 33 represents a microdata and anonymized data for first release R1 at time T1. Table 34 presents an updated microdata, where 3 tuples from Table 33 deleted and 4 new tuples added. Table 34 (b) presents a 2-diversity anonymized data.

If an adversary knows that Bob is in the first group in all releases R1 and R2, also knows that Ken voted by (b) then the adversary will reveal all votes for all participants in the first group. Furthermore if Bob is an adversary then he will reveal all other privacy this known as value-association attack.

	(QI)		(SA)
TID	Age	Zip code	I ₁
Bob	20	3000	a
Ken	25	3500	b
Sam	30	6500	a
Jane	35	4500	b
Linda	40	5500	a
Alice	45	6000	c

(a) Microdata T1

	(QI)		(SA)
TID	Age	Zip code	I ₁
Bob	[20-25]	[3000-3500]	a
Ken	[20-25]	[3000-3500]	b
Sam	[30-35]	[4500-6500]	a
Jane	[30-35]	[4500-6500]	b
Linda	[40-45]	[5500-6000]	a
Alice	[40-45]	[5500-6000]	c

(b) Anonymized data at R1

Table 33 : 2-diversity anonymization

TID	QI		SA
	Age	Zip code	I ₁
Bob	20	3000	a
Sam	30	6500	a
Alice	45	6000	c
Mandy	50	5000	b
David	40	5000	a
Tom	55	6500	c
Carol	45	5000	a

(a) Microdata T2

TID	QI		SA
	Age	Zip code	I ₁
Bob	[20-50]	[3000-6000]	a
Alice	[20-50]	[3000-6000]	c
Sam	[30-50]	[5000-6500]	a
Mandy	[30-50]	[5000-6500]	b
David	[40-60]	[5000-6500]	a
Tom	[40-60]	[5000-6500]	c
Carol	[40-60]	[5000-6500]	a

(b) Anonymized data at R2

Table 34 : A naive 2-diversity anonymized data at R2

The m-invariance model provides a solution in Table 35, m-invariance preserves same signature for the sensitive attribute values and uses a counterfeit tuples in order to keep it. In Table 35 (b) there are 4 groups, comparing the first group G1 with the same one in Table 33 (b) it has the same sensitive values. In G2 there is no new tuple with sensitive value b, therefore a counterfeit value c1 has been used. However if an adversary considers the changes in G1 for R1 and R2, He/she will notice from R1 that Bob and Ken voted by a and b, while from R2 Bob and Mandy voted by a and b, thus Kan and Mandy had same vote. This type of attack known as value-equivalence attack [72].

TID	QI		SA
	Age	Zip code	I ₁
Bob	20	3000	a
Sam	30	6500	a
Alice	45	6000	c
Mandy	50	5000	b
David	40	5000	a
Tom	55	6500	c
Carol	45	5000	a

(c) Microdata T2

TID	QI		SA
	Age	Zip code	I ₁
Bob	[20-50]	[3000-5000]	a
Mandy	[20-50]	[3000-5000]	b
Sam	[30-40]	[5000-6500]	a
c1	[30-40]	[5000-6500]	b
Alice	[40-50]	[5000-6000]	c
David	[40-50]	[5000-6000]	a
Tom	[40-60]	[5000-6500]	c
Carol	[40-60]	[5000-6500]	a

(d) Anonymized data at R2

Table 35 : 2-invariance anonymized data at R2

He et al. model [72] presents value-equivalence attack. In addition it provides a graph-based technique based on m-invariance model to protect sequential data releases against both value-association and value-equivalence attacks. Table 36 present three releases for microdata using He et al. model. It published same signatures in one bucket.

TID	QI		SA
Bob	[20-35]	[3000-6500]	a
Ken	[20-35]	[3000-6500]	b
Sam	[20-35]	[3000-6500]	a
Jane	[20-35]	[3000-6500]	b
Linda	[40-45]	[5500-6000]	a
Alice	[40-45]	[5500-6000]	c
c1	[40-45]	[5500-6000]	a
c2	[40-45]	[5500-6000]	c

(a) Anonymized data at R1

TID	QI		SA
Bob	[20-50]	[3000-5000]	a
Mandy	[20-50]	[3000-5000]	b
Sam	[30-40]	[5000-6500]	a
c3	[30-40]	[5000-6500]	b
Alice	[40-50]	[5000-6000]	c
David	[40-50]	[5000-6000]	a
Tom	[40-60]	[5000-6500]	c
Carol	[40-60]	[5000-6500]	a

(b) Anonymized data at R2

TID	QI		SA
Bob	[20-50]	[3000-5000]	a
Olga	[20-50]	[3000-5000]	b
Sam	[30-40]	[5000-6500]	a
Nic	[30-40]	[5000-6500]	b
Tom	[40-50]	[5000-6000]	c
Carol	[40-50]	[5000-6000]	a
c4	[40-60]	[5000-6500]	c
c5	[40-60]	[5000-6500]	a

(c) Anonymized data at R3

Table 36 : 2-invariance, 2-value equivalence anonymized data

4.2.5 Tuple-equivalence Attack

Table 36 represents an example of the most recent work in sequential releases problem. m-invariance [58] and He et al. [72] models provide solutions for value-association attack and value-equivalence attack respectively. The adversary scenario has been assumed that an adversary knows that a certain participant presents in two or more releases.

Although these models successfully preserve data against for the known attacks discussed in 4.2, we observe that there is a new possible attack threatens the anonymized data; we call it Tuple-equivalence attack. An adversary may know that certain participants in a release will vote for same party. This background knowledge will breach privacy for participants in the following releases. For example, in Table 36 there are four tuples in the first group in each release, in R1 $\langle \text{Bob, Ken, Sam, Jane} \rangle$, in R2 two tuples deleted and two new tuples added to become $\langle \text{Bob, Mandy, Sam, c3} \rangle$, in R3 two tuples deleted and two new tuples added to become $\langle \text{Bob, Olga, Sam, Nic} \rangle$. Let the adversary knows that $\langle \text{Ken, Jane} \rangle$ will vote by same SA value. Even if the adversary doesn't know that they have voted by a or b, s/he can compare first group in R1, R2 and R3 and learns that $\langle \text{Mandy, c3} \rangle$ in R2 have voted by same value that $\langle \text{Ken, Jane} \rangle$ in R1 and $\langle \text{Olga, Nic} \rangle$ in R3 voted. Therefore the adversary will reveals all sensitive values if he/she knows only one sensitive value in any release.

4.3 ρ -different Approach

The dynamic data releases occur when dataset holder needs to publish updated view for same dataset, also when dataset holder publishes dataset after certain time another dataset holder publishes an updated version for same or part of data participants. For example, dataset holder published data for an issue 1 (R1), another issue 2 needed to be published – by same dataset holder or other one- therefore dataset holder makes a new release (R2). These releases may differ with quasi-identifier attributes /sensitive attributes or/and with period of time. The first release R1 released at time stamped $time_1$. The second release R2 will be released at time stamped $time_2$, and so on so forth, where $time_1 < time_2$. Some tuples appeared in R1 will be removed in R2 also new tuples will be added to R2. In addition some values in R1 will be updated and appear in R2. When anonymizing R2 we cannot modify R1.

As mentioned in the introduction of chapter 4, recent researches on data anonymization focus on static datasets, which perform one-time release and did not support republication of updated data. In real datasets, there are many scenarios where dynamic datasets with multiple releases are published. For example, public opinions polling deal with multiple issues in a different timestamp. Hospitals may need to reveal data periodically to present the diseases changes for research purposes. An anonymization model must consider correlations between updated values. Many dynamic data studies consider only tuples insertion and deletion operations and did not offer value update operation (see section 4.2). Few works addressed the update operation but did not take in account the correlations between released data. To best of our knowledge there is no work consider the correlation between sensitive values.

In this work, we identify the privacy problem regarding dynamic dataset publishing and propose a new probabilistic privacy model ρ -different, specifically defined on datasets with continually updated attributes' values. We also present a heuristic anonymization technique to enforce ρ -different.

An adversary may combine released datasets to breach participants' privacy. He/she may compare or link tuples across released tables to identify a specific participant or to raise the likelihood for a certain participant with a certain sensitive value. We illustrate these threats through the following example.

A hospital periodically releases patient's diagnosis data to public in order to allow medical researchers to find valuable data. Moreover, the hospital publishes the correlation information between all diseases such as the following:

P(Deafness Rubella)	80%
P(Deafness Meningitis)	70%
P(Migraine Meniere)	50%
P(Fever Diarrhea)	70%
P(Fever Sore Throat)	50%
P(Diarrhea Gastrointestinal Infections)	50%
P(Aids Aids)	100%
P(Cancer Cancer)	100%

Table 37 : Example of diseases correlations (C)

P(Deafness | Rubella), 80%, which means that with 80% likelihood Rubella disease, will cause Deafness disease. P(Aids |Aids) with 100% likelihood means that Aids disease recognized as permanent diseases.

In the following table there are two releases R1 and R2 which represent anonymized version of dataset T1 and dataset T2 respectively. Ken's tuple was deleted in the second dataset T2 and Frank's tuple was inserted. The sensitive values for Bob and Linda were updated with new values.

TID	QI(Age)	SA1	TID	QI(Age)	SA2
Alice	[20-45]	Aids ¹	Alice	[20-45]	Aids ²
Bob	[20-45]	Diarrhea	Bob	[20-45]	Fever
Linda	[20-45]	GI	Linda	[20-45]	Diarrhea
Ken	[20-45]	Fever	Frank	[20-45]	Cancer

First release (R1)

Second release (R2)

Table 38 : Two datasets releases

R1 represent one group with 4 sensitive values. An adversary can infer that the probability for each sensitive value in R1 is 0.25. However s/he may use the correlations between sensitive values as described in Table 37 to calculate the probability for each sensitive value in R2. For instance the probability that any participant in R2 got Aids² is 0.366. This can be calculated as follows:

The correlation probability $X \rightarrow Y$ can be expressed as Conditional probability $\Pr(Y|X)$, where X is a sensitive value in the previous release $R1$ and the Y is a sensitive value in $R2$.

The Joint probability for dependent values can be calculated using the following equation:

$$\Pr(YX) = \Pr(Y|X).P(X)$$

While for independent values can be expressed as:

$$\Pr(YX) = \Pr(Y).P(X)$$

Then we need to calculate all combination of values in the two groups. It's clear that the probabilities of inferring sensitive values in the new release have been affected by the correlations between these sensitive values. The probability of Aids disease rose. To protect participant's privacy, we would like the probability of linking a sensitive attribute to a certain participant in one or more data releases to be at most ρ .

We assume the following adversary model in the case of public opinions scenario:

- The adversary has access to an external dataset P that contains EI and QI attributes.
- The adversary may know $R1$, $R2$, and correlations (C) between sensitive values in $R1$ and sensitive values in $R2$.

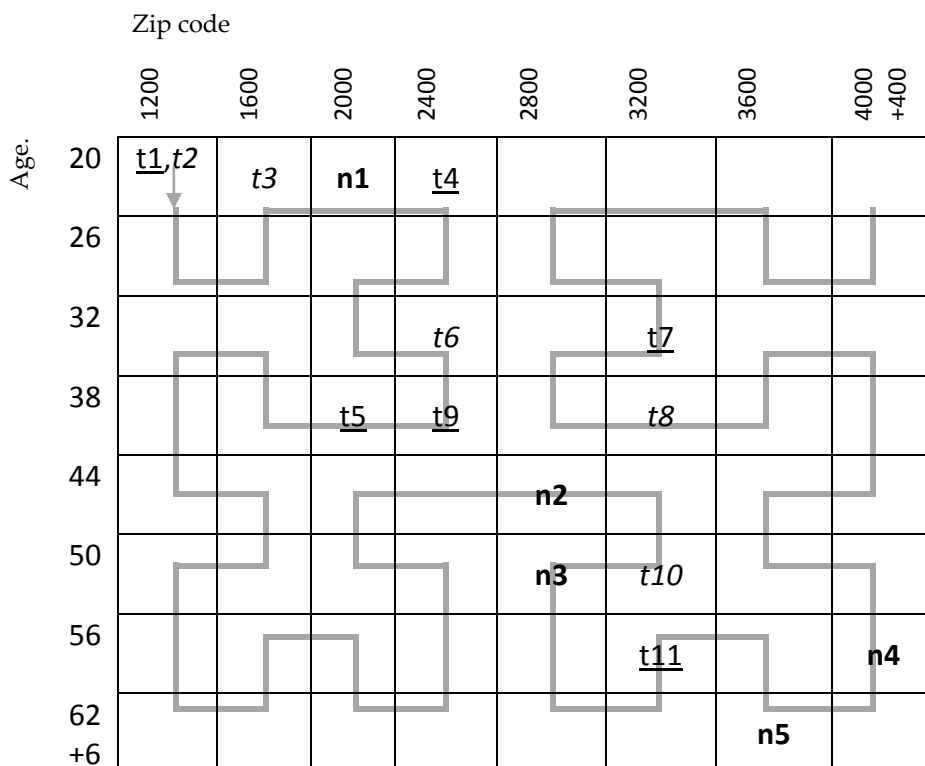


Figure 5 : Hilbert curve mapping for T1 and T2

Italic letters: tuples will be removed in R2, underline letters: tuples will be in both releases

Bold letters: tuple will be added in R2

Consider a hospital that releases medical data for researchers' project every six months. The hospital tried to preserve patients' data. Table 39 depicts the microdata T1 and its first anonymized release R1. R1 guarantee 2-diversity therefore, an adversary cannot identify any patient with probability more than 1/2. Table 39 (iii) represents the correlations between sensitive values from R1 and R2, where $g(R1) \rightarrow g(R2)$ 100% means that 100% likelihood that $g(R2)$ occurs knowing that $g(R1)$ has occurred.

ID	QI		SA	ID	GID	QI		SA	
	Age	Zip	l1			Age	Zip	l ₁	
<u>t1</u>	21	12k	d	<u>t1</u>	1	[21-22]	[12k, 14k]	d	P(g g) 100%
<i>t2</i>	22	14k	b	<i>t2</i>	1	[21-22]	[12k, 14k]	b	P(d d) 50%
<i>t3</i>	24	18k	f	<i>t3</i>	2	[23-24]	[18k, 25k]	f	P(c c) 100%
<u>t4</u>	23	25k	g	<u>t4</u>	2	[23-24]	[18k, 25k]	g	P(m m) 50%
<u>t5</u>	41	20k	f	<u>t5</u>	3	[36-41]	[20k, 27k]	f	
<i>t6</i>	36	27k	g	<i>t6</i>	3	[36-41]	[20k, 27k]	g	
<u>t7</u>	37	33k	d	<u>t7</u>	4	[37-43]	[26k, 35k]	d	
<i>t8</i>	40	35k	f	<i>t8</i>	4	[37-43]	[26k, 35k]	f	
<u>t9</u>	43	26k	g	<u>t9</u>	4	[37-43]	[26k, 35k]	g	
<i>t10</i>	52	33k	d	<i>t10</i>	5	[52-56]	[33k, 34k]	d	
<u>t11</u>	56	34k	g	<u>t11</u>	5	[52-56]	[33k, 34k]	c	

i. Microdata (T1)

ii. First Release (R1)

iii. Sensitive values correlations (C)

Table 39 : 2-diversity anonymized data at R1

The microdata at second released has been updated as follows, all underlined IDs <t1,t4,t5,t7,t9 and t11> will stay in R2, while all italicized IDs <t2,t3,t6,t8 and t10> will be deleted in R2, moreover tuples <n1, n2, n3, n4 and n5> with bold letters and start with n character will be added to R2.

ID	QI		SA
	Age	Zip	I2
<u>t1</u>	21	12k	d
<u>t4</u>	23	25k	g
n1	25	21k	f
<u>t7</u>	37	33k	d
<u>t9</u>	43	26k	g
<u>t5</u>	41	20k	f
n2	46	30k	g
n3	54	31k	d
<u>t11</u>	56	34k	g
n4	60	44k	g
n5	65	36k	f

ID	GID	QI		SA	Pr
		Age	Zip	I2	
<u>t1</u>	1	[21-22]	[12k-14k]	d	50
c1	1	[21-22]	[12k-14k]	b	50
n1	2	[23-25]	[21k-25k]	f	25
<u>t4</u>	2	[23-25]	[21k-25k]	g	75
<u>t7</u>	3	[37-43]	[26k-33k]	d	27
c2	3	[37-43]	[26k-33k]	f	20
<u>t9</u>	3	[37-43]	[26k-33k]	g	53
<u>t5</u>	4	[41-56]	[31k-34k]	f	25
n2	4	[41-56]	[31k-34k]	g	75
n3	5	[54-56]	[31k-34k]	d	25
<u>t11</u>	5	[54-56]	[31k-34k]	c	75
n4	6	[60-65]	[36k-44k]	g	50
n5	6	[60-65]	[36k-44k]	f	50

- i. Microdata (T2) ii. Second Release (R2) (2-invariance)

Table 40 : 2-invariance anonymized data at R2

Table 40 depicts 2-invariance anonymization for T2. It keeps same signature for the sensitive values. For that it uses a counterfeit tuples in order to keep it. EC_1^1 is the first Equivalence Class in R1 which has t1 and t2 tuples with d and b as sensitive values. EC_2^1 is the first Equivalence Class in R2 which has t1 and c1 tuple with d and b as sensitive values. Therefore for each EC_j^i , the j value represent the release number and the i value represent the equivalence class number within each release.

However the requirement of m-invariance model met in Table 40 (ii), it still violate patients privacy. The last column of Table 40 (Pr) represents the probability of each sensitive value based on the correlations between sensitive values. It shows that in EC_2^2 , EC_2^3 , EC_2^4 and EC_2^5 the calculated probabilities are more than the allowed one 50%. Therefore an adversary can identify that t4 in EC_2^2 has g disease with probability 75%.

The update operation and the correlations between sensitive values allow the adversary to enhance his likelihood to identify a certain participants. Furthermore, it shows the in effectiveness of existing solutions in privacy preserving.

Definition 1: We say a released table T^* is ρ -different if for all individual P (individual has any specific sensitive value| F, C) $< \rho$ where F is the frequency(priori belief) and C is the correlation information(likelihood).

Table 41 depicts ρ -different approach, which considers tuples' deletion and insertion operations as well as value update operation. What's more, ρ -different considers the full range of correlations between sensitive values. In Table 41 each group constructed based on the conditional probabilities caused from previous release. For instance EC_2^2 has two tuples t_4 and c_1 with two sensitive values g and c based on the correlations between sensitive values we have $P(g|g) = 100\%$ and $P(c|c) = 100\%$. The probabilities for sensitive values in this group will be 50% which met the ρ -different requirement. The adversary will not be able to identify any participant with probability more than 50%.

ID	QI			SA	Pr
	GID	Age	Zip	I2	
t1	1	[21-25]	[12k-25k]	d	50
n1	1	[21-25]	[12k-25k]	f	50
c1	2	[23-24]	[21k-22k]	c	50
t4	2	[23-24]	[21k-22k]	g	50
t7	3	[37-38]	[33k-34k]	d	50
c3	3	[37-38]	[33k-34k]	m	50
c2	4	[43-44]	[26k-27k]	c	50
t9	4	[43-44]	[26k-27k]	g	50
t5	5	[41-56]	[31k-34k]	f	50
n3	5	[54-56]	[31k-34k]	d	50
n2	6	[41-56]	[31k-34k]	g	50
t11	6	[54-56]	[31k-34k]	c	50
n4	7	[60-65]	[36k-44k]	g	50
n5	7	[60-65]	[36k-44k]	f	50

i. Second Release (R2) (2-different)

Table 41 : 2-different anonymized data at R2

Hilbert curve (see 5.5.1) allow us to map Quasi-identifiers attributes to one dimension. Figure 5 represent the mapping to T1 and T2 datasets. Hilbert curve preserves the data points' locality, which allows us to join closed tuples together.

5. MSA DIVERSITY ALGORITHM

5.1 Adversary Model and Privacy Standard

We assume that the private data is in the form of a table T (see Table 4). Each tuple in T is related to an individual and the table contains a set of attributes called quasi-identifiers (QI) and another set of attributes called sensitive attributes. QI attributes are attributes that, when used together, can be used to identify individual users. SA attributes contain individual's opinions in our case.

We also assume there is no unique identifiers (UI) in T such as SSN, name. We assume the following adversary model in the case of public opinions scenario:

- 1) The adversary has access to an external dataset P (see Table 3) that contains UI and QI attributes.
- 2) The adversary may know non-membership knowledge on some individuals.

Non-membership knowledge is defined as follows:

Definition 2: (non-membership information) For a group G of individuals, we say an adversary has one bit of non-membership information if the adversary knows that an individual u in G does not vote for opinion o on some issue i for exactly one u , o , and i . The adversary can have many bits of non-membership information on the same individual or on up to i different individuals.

Our adversary model is realistic for public opinion datasets for the following reasons: It has been shown in [13] that the external tables as in table P are available in the form of public datasets (such as voters datasets in US). Besides, the QI information for a specific individual can easily be known by an adversary that has close relations with the individual (such as friends and family). The non-membership information mentioned in 1.1 can be gained from two sources. As the public opinion datasets are open to public, the adversary herself might be one of the voters or might collude with some other voters to learn their opinions. As we shall see shortly, due to the nature of anonymization, this creates non-membership knowledge on some groups of people. The non-membership information might also come from close relationships. Note that, given only background knowledge, the data holder cannot release T as it is even though T does not contain any UI attributes. Otherwise, an adversary knowing P and seeing T , can join the two tables to discover that, say Bob, votes

for a and w on issues I_1 and I_2 . Thus, the released data has to ensure that such disclosure is limited probabilistically. The following definition highlights one of our privacy requirements:

Definition 3: (ℓ -pdiversity) We say a released table T^* is ℓ -pdiverse if and only if given T^* and P , the probability that any individual t in T can be mapped to any opinion o on any issue i is bounded by $1/\ell$. The following definition makes it easy to check if a given anonymization is ℓ -pdiversity.

Definition 4: (Equivalence Class) The set of all tuples in a table T^* containing identical values of QI.

For example, in Table 42, t1, t2, and t3 form an equivalence class as they have the same age and zip code. Similarly, t6, t8, and t9 form another equivalence class.

Tuple ID	(QI)		(SA)	
	Age	Zip code	I_1	I_2
t1	[20-25]	[3000-4000]	a	w
t2	[20-25]	[3000-4000]	b	z
t3	[20-25]	[3000-4000]	d	x
t6	[40-55]	[5000-6500]	a	y
t8	[40-55]	[5000-6500]	b	x
t9	[40-55]	[5000-6500]	c	w
t4	[30-45]	[4500-6500]	a	x
t5	[30-45]	[4500-6500]	b	y
t7	[30-45]	[4500-6500]	c	z

Table 42 : MSA-diversity released data T_2^*

Theorem 1: An anonymization T^* of T is ℓ -pdiversity if and only if for every equivalence set in T^* and for every issue; $\frac{\text{the number of the most frequent opinion}}{\text{the size of the equivalence class}} < \frac{1}{\ell}$.

Note that in Table 42, T_2^* is 3-pdiverse anonymization of T given P . For example, an adversary knowing that Linda (t6) is in the anonymization can map Linda to the second

equivalence class. The probability that Linda votes for any opinion on any issue is bounded by $1/3$. In Table 22, T_1^* is 2-pdiverse; the probability that Linda votes for c is $1/2$.

We now formally define MSA-diversity for non-membership attacks:

Definition 5: (MSA-diversity) We say a released table T^* is ℓ -mdiverse if and only if T^* is $(\ell-i)$ -pdiverse under i bits of non-membership knowledge for $i \geq 0$.

Surely, MSA diversity is a harder problem, especially when one faces multiple issues in the dataset. As an example, in Table 2; even if the dataset is 3-pdiverse, it violates 3-mdiversity. If the adversary knows that Amy does not vote for c for issue 1, the anonymization would still satisfy 2-pdiversity with respect to the same issue. However, the anonymization would violate 2-pdiversity with respect to issue 2. On the other hand, in Table 42, Table T_2^* is 3-mdiverse anonymization of T . Even if the adversary knows that Linda does not vote for c , the probability that she votes for any opinion on any issue is still bounded by $1/2$.

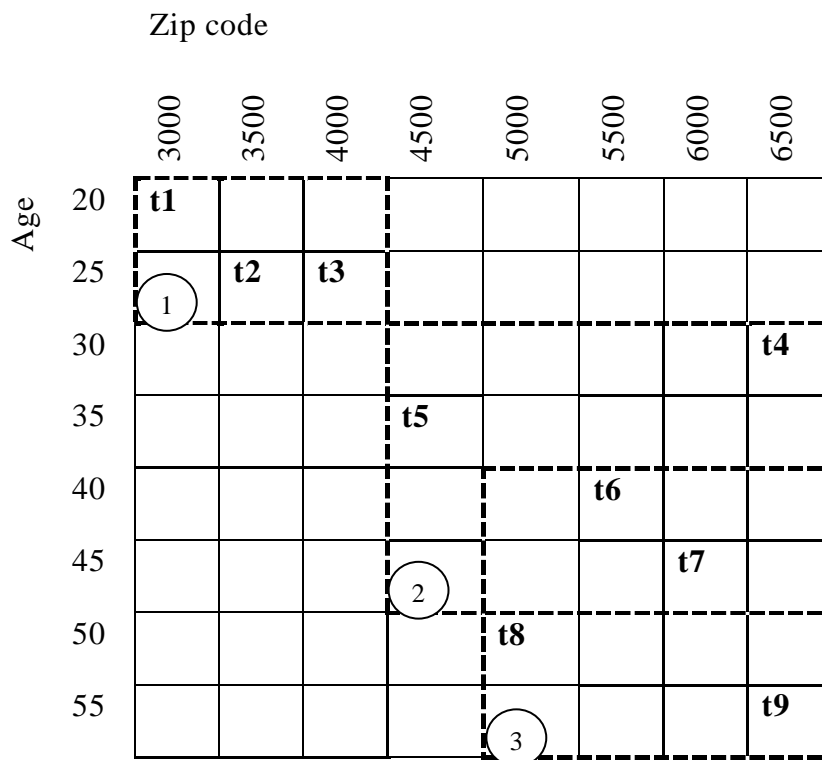


Figure 6 : 3-diversity groups using our model

As an example, consider Table 4 and Table 42 . Figure 6 shows a 2D representation of table T, where the x-dimension is Zipcode And y-dimension is Age and tuples are placed according to their Age and zip-codes. As we shall see in chapter 5, our approach creates the anonymization T_2^* . In Figure 6, dashed rectangles represent the grouping in T_2^* . The groups are enclosed by three rectangles and the corresponding anonymization is more utilized with respect to the metrics mentioned in chapter 2.6.

5.2 Problem Formulation

Given a private table T with quasi-identifiers and sensitive issues, and a privacy parameter ℓ , find an anonymization T^* of T, such that

- T^* is ℓ -mdiverse with respect to the adversary given in this section.
- T^* minimizes the information cost metric.

5.3 Data preprocessing

Data preprocessing is an important step to the dataset to make it more suitable for data mining and getting more efficient results. Datasets may be noisy, incomplete, and inconsistent due to their huge size. Bad or low data quality will lead to poor results. Many data preprocessing techniques can be used like: data cleaning, data transformation, attribute construction, data reduction, data discretization and/or data linkage. In our experiments we have used data cleaning to remove noisy data specially outliers tuples and data reduction to reduce the data size.

5.4 Checking for MSA Diversity

Checking if a given group of tuples (e.g., equivalence class) satisfies MSA Diversity is a sub-problem in our algorithm. Unfortunately, checking an arbitrary group for MSA diversity is not a trivial task. Instead, we aim to create a subfamily of groups that are proven to be MSA diverse. Such groups satisfy the 'SA-distinct' property which is defined as follows:

Definition 6: (SA-distinct Group) A given group of tuples G is SA-distinct group if and only if for any pair of tuples $t_1 \in G$ and $t_2 \in G$ and for any issue i , $t_1[i] \neq t_2[i]$. For a given group of tuples, assuming we have two issues i_1 and i_2 , one can plot the distribution of sensitive opinions in a group G over a matrix. The dimension i of the matrix represents the issue i . Any tuple $t \in G$ is drawn on the matrix index (q, r) if $t[i_1] = q$ and $t[i_2] = r$. The group G is SA-distinct if every row and column of the matrix contains at most one tuple. As an example, in

Figure 10(a), we place the first three tuples t_1 , t_2 , and t_3 on the matrix. This group is SA-distinct as no row or column contains more than one tuple. However the group t_5 , t_6 , t_8 and t_9 is not SA-distinct as can be seen in Figure 10(b).

Theorem 2: SA-distinct group of at least ℓ elements satisfies ℓ -mdiversity.

5.5 Generalization Algorithm

We now propose an efficient heuristic algorithm for the MSA diversity problem. Our algorithm has two phases:

- Dataset grouping: T is partitioned into a set of disjoint groups such that each group is SA-distinct, thus ℓ -mdiverse.
- Generalization: The groups are generalized or anatomized.

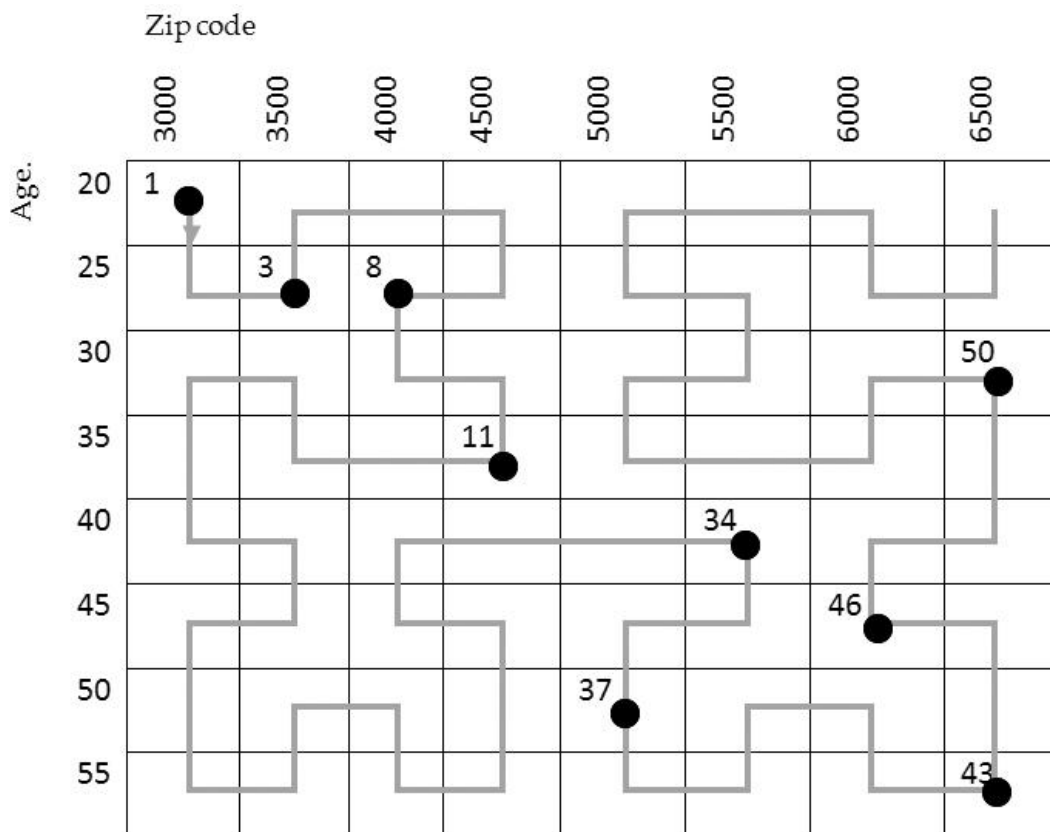


Figure 7 : Hilbert curve mapping

5.5.1 Mapping multi-dimensional QI to one-dimension

MSA- diversity model works in a top-down manner. QI attributes needed to be arranged in a way that similar tuples with respect to QI stay closed to each other. Constructing equivalence classes from similar QI and diverse SA will produce a high-quality released data.

QI have many attributes therefore a mechanism needed to map multidimensional attributes to one dimension, moreover this mechanism should preserving locality of the data points. Let represent tuples as data points in a 2D space (for Example Figure 7). One way of mapping the multi-dimensional QI to one-dimension is space-filling curve. It works like a single path passes over all data points (tuples). Many types of space-filling curves have been discussed in literature. The main difference between it is the way of representing the one-dimension space. Figure 8 shows three different kinds of space-filling curves.

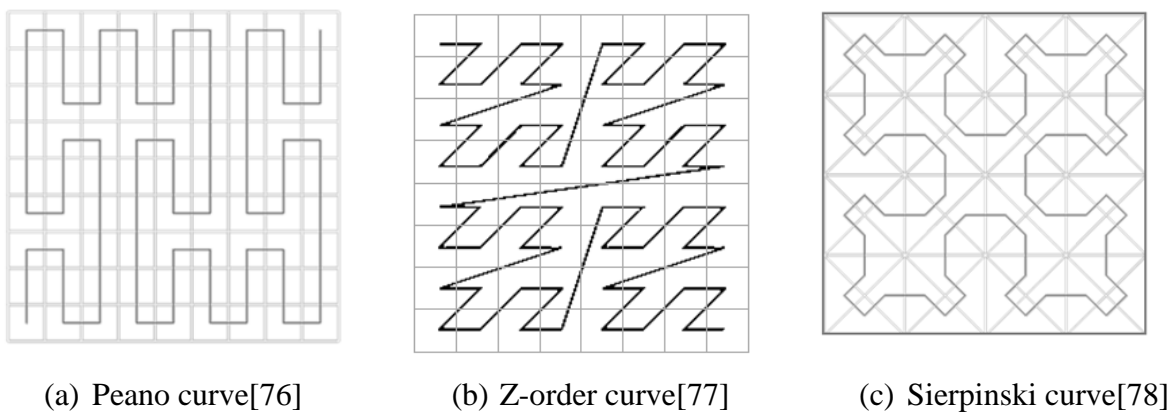


Figure 8 : Different types of space-filling curves

The Peano curve is the first space-filling curve technique [76]. The mapping is based on the ternary subdivision. Its structure is shown in Figure 8(a). The Z-order curve technique [77] maps quadrants recursively. The resulting order is similar to results from depth-first traversal of a quad-tree. Therefore the Z-ordering can be used to construct high dimensional data structures. Its structure is shown in Figure 8(b). The Sierpinski curve is based on a triangular subdivision. Its geometric construction is shown in Figure 8(c). The Hilbert curve is a space filling curve that visits every data point in a square grid. It makes better use of the harmony of neighboring data points. [79] and [80] show that Hilbert curve outperforms other techniques by minimizing the number of clusters for 2×2 range queries. It provides the minimum number of clusters. Moreover it preserves the data points' locality. Therefore we deployed Hilbert curve in our work.

The pseudo code for our heuristic algorithm is given in Figure 13. The algorithm accepts two parameters: the microdata T and the privacy parameter ℓ , obtains QI groups for publication. The algorithm first maps the multi-dimensional QI attributes to a single dimension using Hilbert space filling curve as shown in Figure 7, and sorts the records according to their QI value as shown in Table 43 (lines 1-3). Then in each iteration, the algorithm constructs SA-distinct groups of at least ℓ tuples heuristically. Starting with the first ℓ tuples; it checks if it's possible to construct an SA-distinct group. If it's possible; proceeds to the next QI group. Else if there are extra tuples which cannot be added to any group, the algorithm borrows some tuples from the next tuple in order to construct a new partition. The remaining ungrouped tuples are grouped together.

Tuple ID	(QI)	(SA)	
	1D	(I ₁)	(I ₂)
t1	1	a	w
t2	3	b	z
t3	8	d	x
t5	11	b	y
t6	34	a	y
t8	37	b	x
t9	43	c	w
t7	46	c	z
t4	50	a	x

Table 43 : Microdata with one dimension QI

From Table 43 we have two issues (I1 and I2) and for each issue there are 4 distinct opinions. Figure 9 depicts the 3D scatter plot for Table 43. For the first three tuples, we can easily construct a SA-distinct group which by definition 2 satisfies 3-ndiversity. This can be seen from the matrix in Figure 10(a), as t1, t2 and t3 are in different rows and columns, thus $G1 = \langle t1, t2, t3 \rangle$ becomes the first group. As shown in Figure 10(b) for t5, t6, and t8; we can't construct a SA-distinct group as $t5[I_2] = t6[I_2]$ therefore we remove t5 and borrow t9 in order to construct another SA-distinct group $G2 = \langle t6, t8, t9 \rangle$. Moreover, t5 still does not belong to any group. By grouping the remaining tuples we construct the last SA-distinct group as $G3 = \langle t5, t7, t4 \rangle$. The three resulting groups G1, G2, and G3.

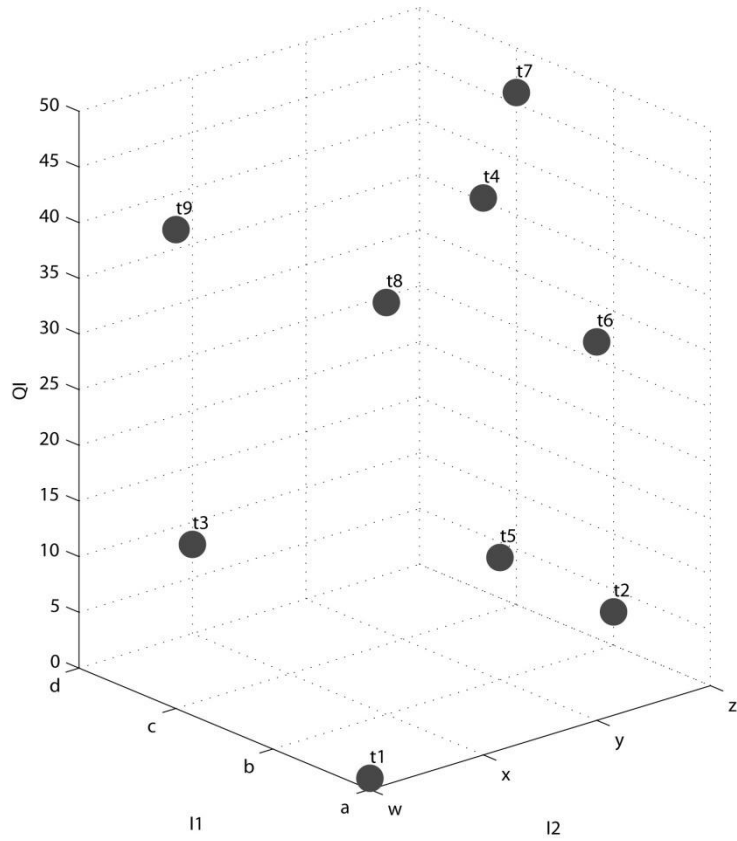


Figure 9 : 3D scatter plot for Table 43

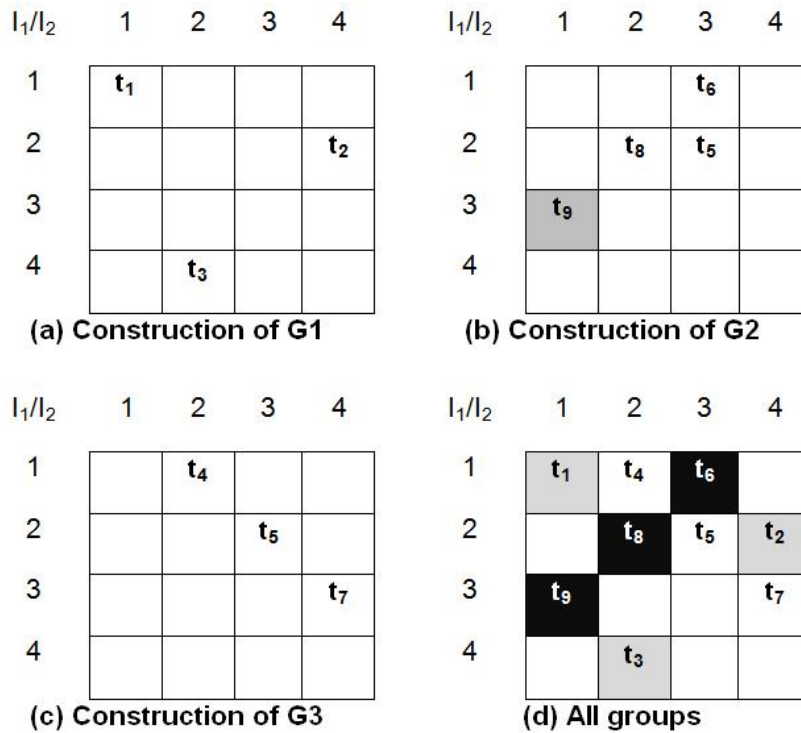


Figure 10 : Groups construction process

Figure 10 represents the groups' construction process. To construct an equivalence class which respect ℓ -pdiversity, we need to represent the SA values in a matrix like Figure 10(a). Next MSA-diversity algorithm checks the ℓ -pdiversity condition.

A permutation matrix is a square binary matrix. It has a property that all its entries are 0's and 1's, where each row and each column has a single 1.

1	0	0
0	1	0
0	0	1

1	0	0
0	0	1
0	1	0

0	1	0
1	0	0
0	0	1

0	1	0
0	0	1
1	0	0

0	0	1
1	0	0
0	1	0

0	0	1
0	1	0
1	0	0

Figure 11 : Permutation matrices for 3 elements

Figure 11 depicts 6 different matrices for 3 elements. The number of permutations matrices of n distinct elements is equals the n factorial. Therefore in Figure 11 example there are six different matrices.

*		
	*	
		*

Figure 12 : One matrix of Costas arrays for 3 elements

Costas array [81-83] is a special permutation matrix. For n elements it can constructs n! matrices having same property of permutation matrix. Costas arrays arise in sonar, radar and cryptography applications. In our model, we have a set of sensitive attribute values and each row/column represent a distinct SA value. The key question is to find maximum number of Costas arrays.

As shown in dataset of Table 4 and its generalization data in Table 42, there are 3 different groups have been constructed heuristically based on the general idea of Costas array. Figure 10 depicts the construction process.

5.5.2 The MSA-diversity Heuristic Algorithm

Algorithm 1 Heuristic algorithm

Require: microdata T , value of ℓ ;

Ensure: output T^* satisfies probabilistic ℓ -m diversity.

- 1: **Map** the MD QI to 1D using Hilbert space filling curve
- 2: **Sort** the records according to their 1D QI value
- 3: $P =$ empty set, $F =$ empty set

Traverse(

- 4: **for** ($i=0, i < k, i++$) { /* where k is the number of distinct QI values */
- 5: $G_i = QI_i$ tuples
- 6: $j = 0$
- 7: **Group:**
- 8: **Grouping**(G_i, P_i, ℓ, j)
- 9: **if** ($|F_{ij}| = 0$) { /* no groups in P_i */
- 10: $G_i = G_i + G_{i+1}$ /*merge the current QI values with the next one */
- 11: $i = i + 1$
- 12: **Go to Group**
- 13: }
- 14: }

Grouping(G, P, ℓ, j)

- 15: **while** (!Satisfy(P, ℓ) and $|G| > 0$) {
- 16: **maxitems**(x): Select value x which has max number of items in its row (r) and column (c).
- 17: $G' = G \setminus \langle r, c \rangle$
- 18: **Add** x to P
- 19: **Grouping**(G', P, ℓ, i)
- 20: }
- 21: **if** (Satisfy(P, ℓ)) {
- 22: $F_{ij} = P$ /* to save the final groups for each QI */
- 23: $j = j + 1$
- 24: $G = G \setminus P$
- 25: $P = \emptyset$
- 26: **if** (Satisfy(G, ℓ)){
- 27: **Grouping**(G, P, ℓ, i)
- 28: }
- 29: }

Satisfy(P, ℓ)

- 30: **if** $|P| \geq \ell$
 - 31: **return true**
 - 32: **else**
 - 33: **return false**
-

Figure 13 : Pseudo code for our heuristic algorithm

6. EXPERIMENTAL RESULTS

In this chapter, we compare our model with the existing state-of-the-art. The Algorithm is implemented in C# and the experiments were run on a Dell 2.4GHz machine with 2GB of memory, running windows 7.

We used MovieLens dataset obtained from the GroupLens research lab¹. It contains 10000054 ratings applied to 10681 movies by 71567 users. Ratings for each movie vary from 1 to 7. We used three quasi-identifier attributes Age, Gender and Zip code. We picked seven movies from the dataset that are the most frequently rated among all movies. We marked the movie ratings as sensitive (movies can be thought as issues and actual ratings as opinions). In our experiment we chose a sub set contains users will 7 ratings for all of them, therefore our data set become 684 users. We used mainly discernibility metric (DM), loss metric (LM) and *average relative error* (AvRE) as information loss metric. We evaluate data accuracy using aggregate query answering as follows. First, we compute the corresponding generalized groups [24, 58, 84]. Second, we process a workload of 684 queries one query for each tuple on the resulting tables. The effectiveness of generalization is computed by the average relative error.

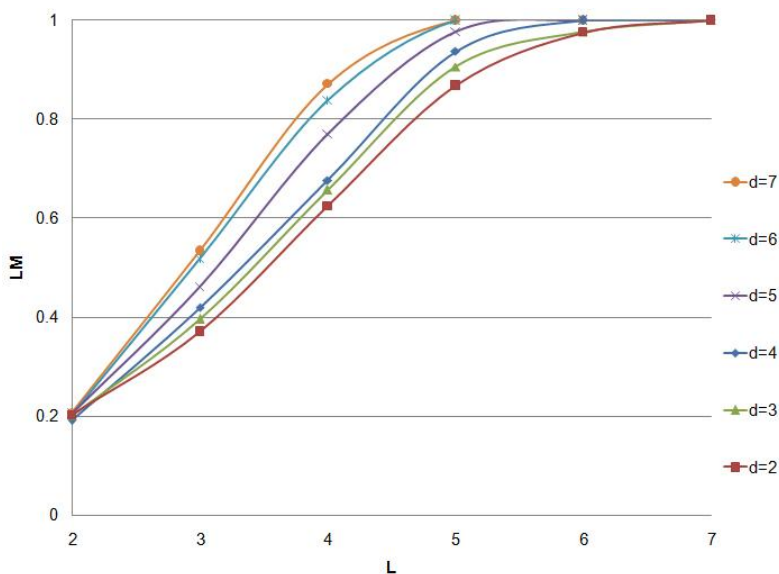


Figure 14 : LM, Information loss with varying ℓ and d

¹ URL: <http://www.grouplens.org>

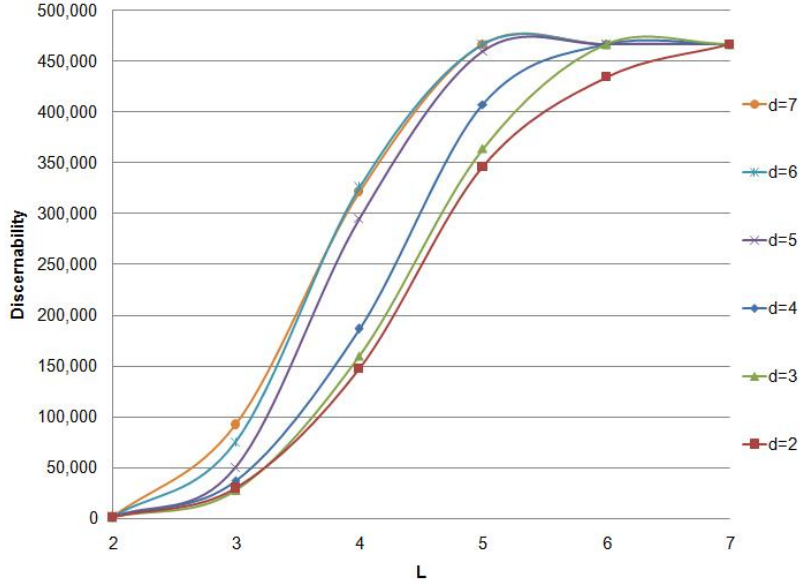


Figure 15 : DM, Information loss with varying ℓ and d

6.1 Utility - varying ℓ and d

Figure 14 depicts the LM for various ℓ and d (number of ratings). Recall that a 0 value of LM means no information loss. Figure 15 reports DM. As can be seen from the figures, an increase in the privacy parameter ℓ results in more information loss due to the privacy/utility tradeoff. Similarly higher numbers of ratings have similar effect due to curse of dimensionality. Compared to the parameter d , utility is more sensitive to the changes in ℓ . For very small and very large ℓ , the number of ratings has little effect of the utility.

6.2 Comparison with Previous Work

We now compare our approach with the state-of-the-art anonymization algorithm for multiple sensitive attributes by Gal's et al. For Gal's et al model, we assume $k=\ell$. We would like to emphasize that in terms of probability of disclosure, the ℓ -diversity definition adapted by Gal et al is weaker than ℓ -mdiversity proposed in 4.2.5. At the same privacy level ℓ and no non-membership information, ℓ -diversity by Gal et al ensures the number of distinct sensitive values should be smaller than ℓ , thus does not guarantee a bound on the probability of disclosure. ℓ -mdiversity, on the other hand, bounds the probability by $1/\ell$. Moreover, MSA diversity algorithm ensures both privacy metrics. Thus, in our domain, MSA-diversity algorithm offers higher levels of privacy for all ℓ . We experimentally demonstrate probabilities of disclosure for both algorithms in Section 6.2.3.

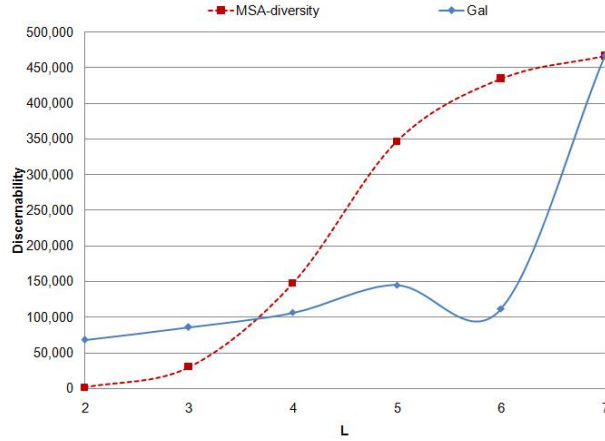


Figure 16 : DM comparison, varying ℓ and $d=2$

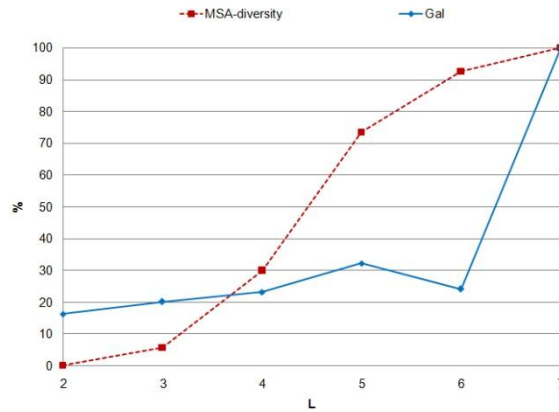


Figure 17 : Query accuracy with varying ℓ and $d=2$

6.2.1 Utility comparison - varying ℓ

For $d=2$ and $d=5$, we depict utility metric results in Figure 16 and Figure 17. Query accuracy results given in Figure 18 and Figure 19 show a similar behavior. For low ℓ values (which we believe to be the most practical privacy parameters as utility drops fast with high ℓ values), MSA-diversity method results in less information loss. As mentioned in `retpahC5`, Hilbert curve-based generalizations are more flexible than partition-based approaches and can achieve higher levels of utility. The reason why we do not have the same relative performance for larger ℓ is because the ℓ -diversity definition adapted by Gal et al is less restrictive.

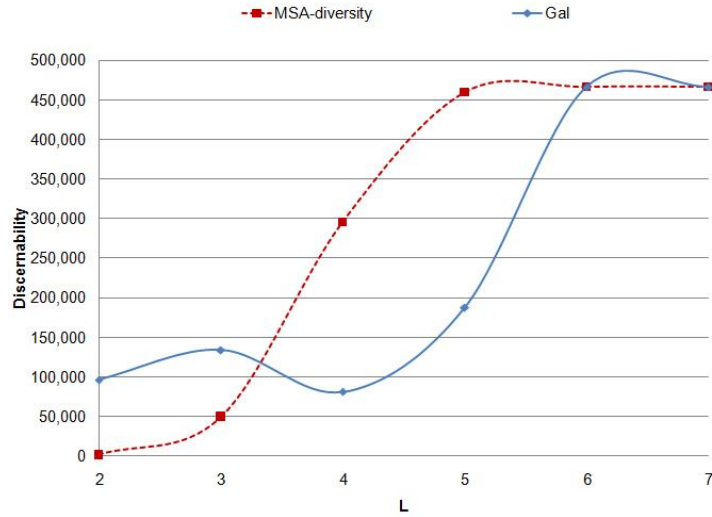


Figure 18 : DM comparison, varying ℓ and $d=5$

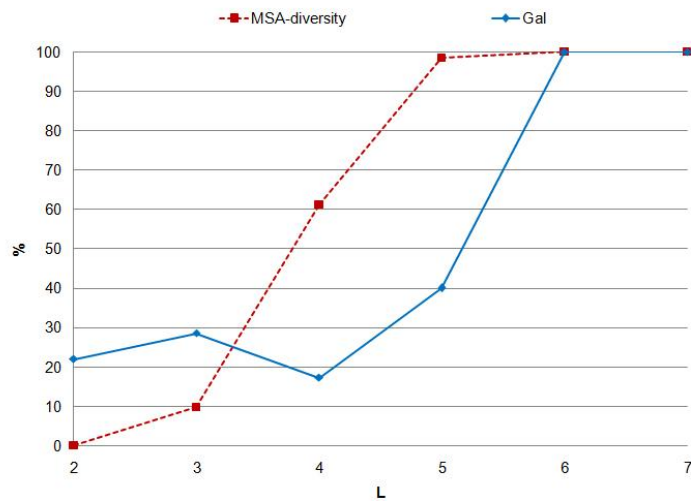


Figure 19 : Query accuracy with varying ℓ and $d=5$

6.2.2 Utility comparison - varying d

Figure 20 and Figure 21 depict the utility comparison of two approaches. As also shown earlier, The MSA-diversity algorithm, when compared to Gal et al, creates better utilized anonymizations for $\ell = 2$ but performs worse for $\ell = 5$. We also observe that the utility performance of both algorithms is not very sensitive to the number of sensitive attributes. Figure 22 and Figure 23 showing query accuracy results support these results.

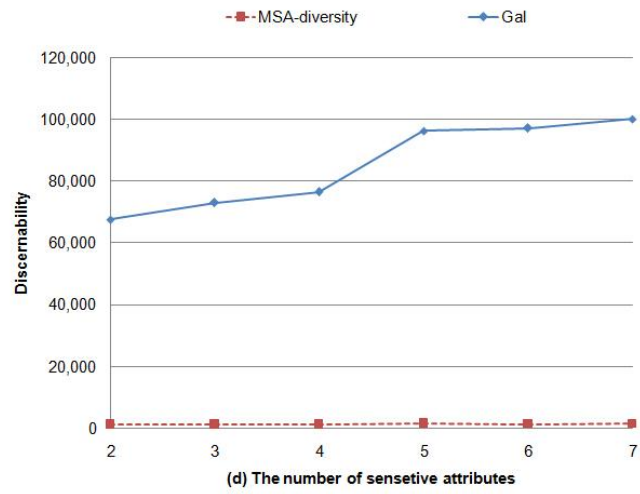


Figure 20 : DM, varying number of sensitive attributes and $\ell=2$

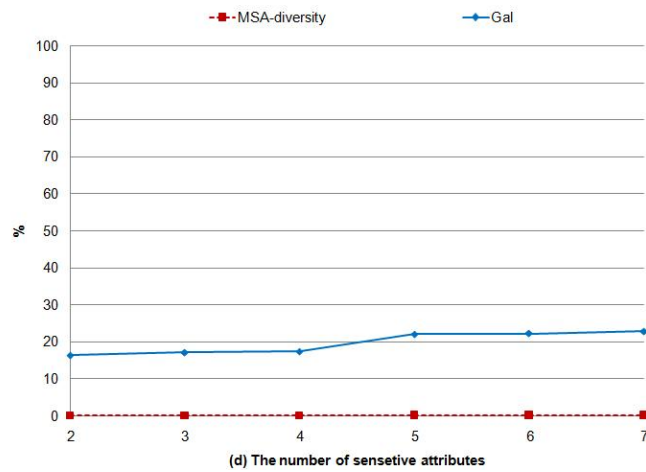


Figure 21 : Query accuracy with varying number of sensitive attributes and $\ell = 2$

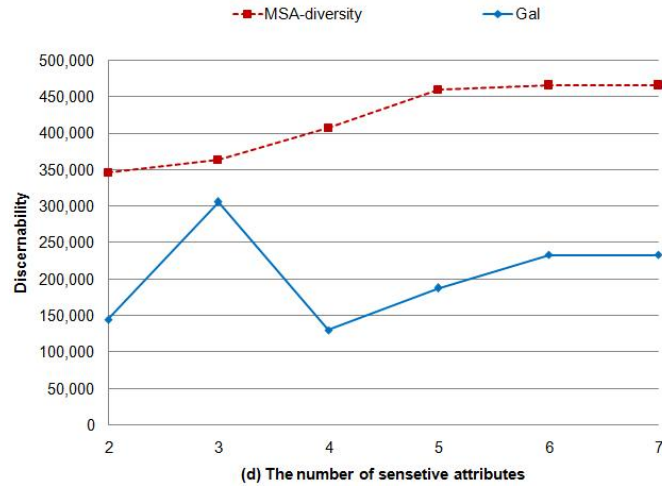


Figure 22 : DM, varying number of sensitive attributes and $\ell = 5$

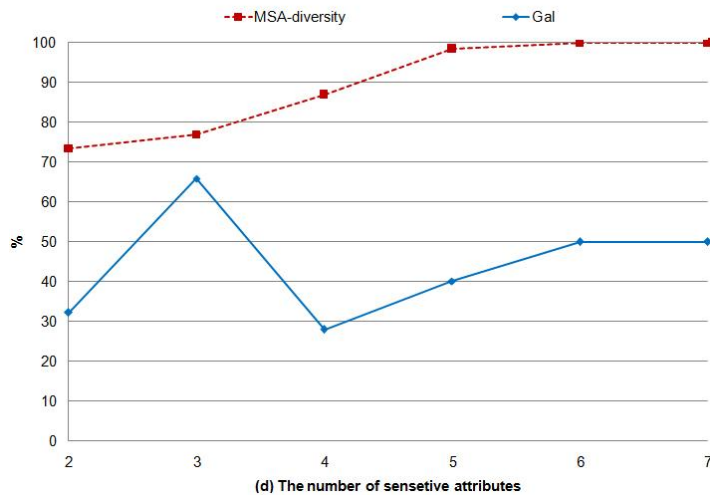


Figure 23 : Query accuracy with varying number of sensitive attributes and $\ell = 5$

6.2.3 Probability of disclosure comparison

As mentioned earlier, work by Gal et al does not guarantee a bound on the probability of disclosure (e.g., the probability that an individual will be associated with an opinion). Figure 24 and Figure 25 show the disclosure probabilities for all tuples after applying MSA-diversity model and Gal's et al model, where the x-dimension represents the probability of disclosure and y-dimension represent the number of tuples. Figure 24 depicts the results for $d = 2$ and $\ell = 5$. MSA-diversity algorithm ensures a maximum of 0.2 probability for all non-suppressed tuples, while Gal's et al anonymization algorithm results in disclosures with probabilities that can be as high as 0.6. Similarly, in Figure 25, we show disclosure

probabilities for $d = 5$ and $\ell = 2$. The MSA-diversity algorithm ensures a maximum of 0.5 probability for all non-suppressed tuples. However Gals et al algorithm results in disclosures with probabilities that can be as high as 0.86. This clearly shows that MSA-diversity algorithm achieving ℓ -mdiversity, provides better privacy by protecting sensitive information against probabilistic adversaries.

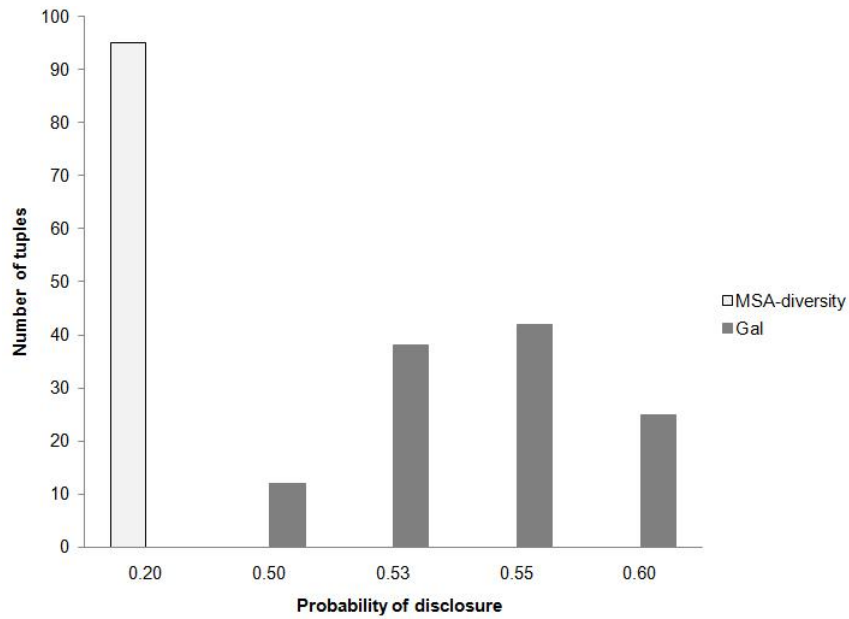


Figure 24 : Probability of disclosure for each tuple, $d=2$ and $\ell = 5$

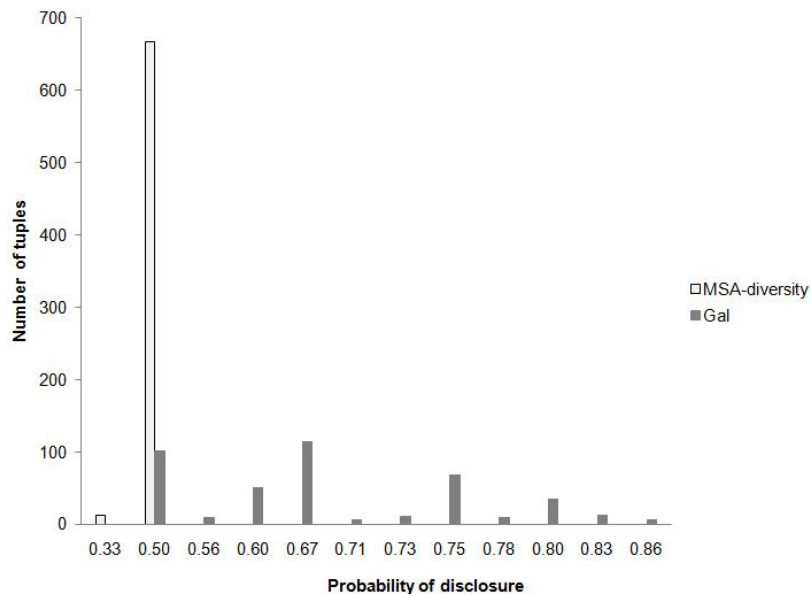


Figure 25 : Probability of disclosure for each tuple, $d=5$ and $\ell=2$

7. CONCLUSIONS

In this thesis, new anonymization models are proposed. ρ -different is a probabilistic model for dynamic release data. MSA-privacy is a probabilistic definition for releasing data with multiple sensitive attributes. It's an alternative model that allows accurate anonymization. Most well known anonymization models are designed for data with single sensitive attributes; these models are not applicable for data with multiple sensitive attributes. Adversaries can use a new attack like non-membership attack to breach individual privacy. Some of recent research concerns this type of attacks and provides a remedy for it. However they are fallen in other types of attacks such as membership attack and probabilistic attack.

ρ -different preserve privacy for dynamic data with insertion, deletion and update operations. What's more it considers all correlations between sensitive values.

The important advantage of the MSA-diversity model is the simplicity of deploying the algorithm. Data holder needs only to feeds the algorithm with the data set and the privacy level (ℓ value).

ρ -different and MSA-diversity models are applicable to publish anonymized tabular data for any other domain. Public opinion polls problem is an example.

For future work, ρ -different and MSA-diversity may used to solve sequential data release problem with multiple sensitive attributes. Where a released data has been published and data holder needs to publish another modified copy.

APPENDIX A: LIST OF ACRONYMS

T	Table
G	Group
<i>o</i>	Opinion
MSA	Multiple Sensitive Attributes
SSA	Single Sensitive Attribute
PPDP	Privacy-Preserving Data Publishing
QI	Quasi-Identifiers
EU	Explicit Identifiers
EC	Equivalence Class
UI	Unique Identifiers
DM	Discernibility Metric
LM	Information Loss Metric
AvRE	Average Relative Error
EMD	Earth Mover's Distance
CODIP	Complete Disjoint Projections
PAC	Political Action Committees

REFERENCES

1. Sweeney, L.S.a.L., *Achieving K-Anonymity Privacy Protection Using Generalization and Suppression*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002. **10**: p. 2002.
2. LeFevre, K., D.J. DeWitt, and R. Ramakrishnan, *Mondrian Multidimensional K-Anonymity*, in *Proceedings of the 22nd International Conference on Data Engineering*2006, IEEE Computer Society. p. 25.
3. Machanavajjhala, A., J. Gehrke, and D. Kifer, *l-diversity: Privacy beyond k-anonymity*2007.
4. Ninghui, L., L. Tiancheng, and S. Venkatasubramanian. *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*. in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. 2007.
5. Donald R.Kinder, D.O.S., *Public Opinions and Political Action*, in *The Handbook of Social Psychology*1985: New York. p. 659-741.
6. Lippmann, W., *Public opinion*1997: Transaction Publishers.
7. Page, B.I. and R.Y. Shapiro, *Effects of Public Opinion on Policy*. The American Political Science Review, 1983. **77**(1): p. 175-190.
8. Horowitz, J.M., *WORLD PUBLICS WELCOME GLOBAL TRADE – BUT NOT IMMIGRATION*2007.
9. European_commission. *public opinion*. 2008; Available from: http://ec.europa.eu/public_opinion/index_en.htm.
10. Naurath, N. *Perceived Acceptance of Homosexuals Differs Around Globe*. 2007; Available from: <http://www.gallup.com/poll/102478/perceived-acceptance-homosexuals-differs-around-globe.aspx/>.
11. Gal, T.S., Z. Chen, and A. Gangopadhyay, *A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes*. International Journal of Information Security and Privacy, 2008: p. 28-44.
12. YANG Xiao-Chun, L.X.-Y., WANG Bin, YU Ge *Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing*. Chinese Journal of Computers, 2008.
13. Sweeney, L., *k-anonymity: a model for protecting privacy*. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 2002. **10**(5): p. 557-570.

14. Zhong, S., Z. Yang, and T. Chen, *k-Anonymous data collection*. Information Sciences, 2009. **179**(17): p. 2948-2963.
15. Shuchi Chawla, C.D., Frank Mcsherry, Kunal Talwar. *On the Utility of Privacy-Preserving Histograms*. in *In 21st Conference on Uncertainty in Artificial Intelligence* 2005. AUAI Press.
16. Aggarwal, C.C., *On k-anonymity and the curse of dimensionality*, in *Proceedings of the 31st international conference on Very large data bases*2005, VLDB Endowment: Trondheim, Norway. p. 901-909.
17. Truta, T.M. and B. Vinay. *Privacy Protection: p-Sensitive k-Anonymity Property*. in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. 2006.
18. Byun, J.-W., et al., *Efficient k-Anonymization Using Clustering Techniques Advances in Databases: Concepts, Systems and Applications*, R. Kotagiri, et al., Editors. 2007, Springer Berlin / Heidelberg. p. 188-200.
19. Nergiz, M.E., C. Clifton, and A.E. Nergiz. *MultiRelational k-Anonymity*. in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. 2007.
20. Matatov, N., L. Rokach, and O. Maimon, *Privacy-preserving data mining: A feature set partitioning approach*. Information Sciences, 2010. **180**(14): p. 2696-2720.
21. Ahn, L.v., A. Bortz, and N.J. Hopper, *k-anonymous message transmission*, in *Proceedings of the 10th ACM conference on Computer and communications security*2003, ACM: Washington D.C., USA. p. 122-130.
22. Chen, X., et al., *New receipt-free voting scheme using double-trapdoor commitment*. Information Sciences, 2011. **181**(8): p. 1493-1502.
23. Bayardo, R.J. and R. Agrawal, *Data Privacy through Optimal k-Anonymization*, in *Proceedings of the 21st International Conference on Data Engineering*2005, IEEE Computer Society. p. 217-228.
24. LeFevre, K., D.J. DeWitt, and R. Ramakrishnan, *Workload-aware anonymization*, in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*2006, ACM: Philadelphia, PA, USA. p. 277-286.
25. Li, Z. and X. Ye, *Privacy protection on multiple sensitive attributes*, in *Proceedings of the 9th international conference on Information and communications security*2007, Springer-Verlag: Zhengzhou, China. p. 141-152.

26. Ghinita, G., et al., *Fast data anonymization with low information loss*, in *Proceedings of the 33rd international conference on Very large data bases*2007, VLDB Endowment: Vienna, Austria. p. 758-769.
27. Chung, K.-L., Y.-L. Huang, and Y.-W. Liu, *Efficient algorithms for coding Hilbert curve of arbitrary-sized image and application to window query*. *Information Sciences*, 2007. **177**(10): p. 2130-2151.
28. A. Trevor Thrall, J.C., *Why Did the United States Invade Iraq?*, 2008, University of Michigan – Dearborn.
29. HAFNER, K. *And if You Liked the Movie, a Netflix Contest May Reward You Handsomely* October 2, 2006; Available from: <http://www.nytimes.com/2006/10/02/technology/02netflix.html>.
30. Chen, B.-C., et al., *Privacy-Preserving Data Publishing*. *Found. Trends databases*, 2009. **2**(1;2): p. 1-167.
31. Sweeney, L., *Simple Demographics Often Identify People Uniquely*. , 2000, Carnegie Mellon University, Data Privacy Working Paper 3: Pittsburgh
32. Agrawal, R. and R. Srikant, *Privacy-preserving data mining*. *SIGMOD Rec.*, 2000. **29**(2): p. 439-450.
33. Agrawal, D. and C.C. Aggarwal, *On the design and quantification of privacy preserving data mining algorithms*, in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*2001, ACM: Santa Barbara, California, United States. p. 247-255.
34. Aggarwal, C.C. *On Randomization, Public Information and the Curse of Dimensionality*. in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. 2007.
35. Dalenius, T. and S.P. Reiss, *Data-swapping: A technique for disclosure control*. *Journal of Statistical Planning and Inference*, 1982. **6**(1): p. 73-85.
36. Li, Y. and H. Shen, *Equi-Width Data Swapping for Private Data Publication*, in *Proceedings of the 2009 International Conference on Parallel and Distributed Computing, Applications and Technologies*2009, IEEE Computer Society. p. 231-238.
37. Qing, Z., et al. *Aggregate Query Answering on Anonymized Tables*. in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. 2007.
38. Rubner, Y., C. Tomasi, and L.J. Guibas. *A metric for distributions with applications to image databases*. in *Computer Vision, 1998. Sixth International Conference on*. 1998.

39. Guibas, S.D.C.a.L.J., *The earth mover's distance: Lower bounds and invariance under translation*, 1997.
40. Li, J., 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 data2008*, ACM: Vancouver, Canada. p. 473-486.
41. Fung, B.C.M., et al., *Privacy-preserving data publishing: A survey of recent developments*. ACM Comput. Surv., 2010. **42**(4): p. 1-53.
42. Nergiz, M.E., M. Atzori, and C. Clifton, *Hiding the presence of individuals from shared databases*, in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data2007*, ACM: Beijing, China. p. 665-676.
43. Xiao, X. and Y. Tao, *Personalized privacy preservation*, in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data2006*, ACM: Chicago, IL, USA. p. 229-240.
44. Lodha, S. and D. Thomas, *Probabilistic Anonymity*
Privacy, Security, and Trust in KDD, F. Bonchi, et al., Editors. 2008, Springer Berlin / Heidelberg. p. 56-79.
45. Chawla, S., et al., *Toward Privacy in Public Databases*, in *Theory of Cryptography*, J. Kilian, Editor 2005, Springer Berlin / Heidelberg. p. 363-385.
46. Cynthia Dwork, F.M., Kobbi Nissim, Adam Smith. *Calibrating noise to sensitivity in private data analysis*. in *In Proceedings of the 3rd Theory of Cryptography Conference*. 2006. Springer.
47. Rastogi, V., 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 bases2007*, VLDB Endowment: Vienna, Austria. p. 531-542.
48. Blum, A., K. Ligett, and A. Roth, *A learning theory approach to non-interactive database privacy*, in *Proceedings of the 40th annual ACM symposium on Theory of computing2008*, ACM: Victoria, British Columbia, Canada. p. 609-618.
49. Machanavajjhala, A., et al. *Privacy: Theory meets Practice on the Map*. in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. 2008.
50. Meyerson, A. and R. Williams, *On the complexity of optimal K-anonymity*, in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems2004*, ACM: Paris, France. p. 223-228.

51. Park, H. and K. Shim, *Approximate algorithms for K-anonymity*, in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data2007*, ACM: Beijing, China. p. 67-78.
52. LeFevre, K., D.J. DeWitt, and R. Ramakrishnan, *Incognito: efficient full-domain K-anonymity*, in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data2005*, ACM: Baltimore, Maryland. p. 49-60.
53. TheNewYorkTimes. *A Face Is Exposed for AOL Searcher No. 4417749*. 2006; Available from: <http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all>.
54. Li, T., et al., *Slicing: A New Approach for Privacy Preserving Data Publishing*. IEEE Trans. on Knowl. and Data Eng., 2012. **24**(3): p. 561-574.
55. Bu, Y., et al., *Privacy preserving serial data publishing by role composition*. Proc. VLDB Endow., 2008. **1**(1): p. 845-856.
56. Ji-won Byun, Y.S., Elisa Bertino, Ninghui Li. *Secure anonymization for incremental datasets*. in *in the Third VLDB Workshop on Secure Data Management (SDM)*. 2006.
57. Fung, B.C.M., et al., *Anonymity for continuous data publishing*, in *Proceedings of the 11th international conference on Extending database technology: Advances in database technology2008*, ACM: Nantes, France. p. 264-275.
58. Xiao, X. and Y. Tao, *M-invariance: towards privacy preserving re-publication of dynamic datasets*, in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data2007*, ACM: Beijing, China. p. 689-700.
59. Wong, R.C.-W., et al., *Minimality attack in privacy preserving data publishing*, in *Proceedings of the 33rd international conference on Very large data bases2007*, VLDB Endowment: Vienna, Austria. p. 543-554.
60. Khaled El Emam, E.J., Scott Sams, Emilio Neri, Angelica Neisa, Tianshan Gao, Sadrul Chowdhury,, *Pan-Canadian De-Identification Guidelines for Personal Health Information*, 2007, the Office of the Privacy Commissioner of Canada.
61. Li, C., H. Shirani-Mehr, and X. Yang, *Protecting Individual Information Against Inference Attacks in Data Publishing Advances in Databases: Concepts, Systems and Applications*, R. Kotagiri, et al., Editors. 2007, Springer Berlin / Heidelberg. p. 422-433.
62. Hinke, T.H., H.S. Delugach, and R.P. Wolf, *Protecting databases from inference attacks*. Computers & Security, 1997. **16**(8): p. 687-708.

63. Raymond Heatherly, M.K., Bhavani Thuraisingham, *Preventing Private Information Inference Attacks on Social Networks*, 2009.
64. Kifer, D., *Attacks on privacy and deFinetti's theorem*, in *Proceedings of the 35th SIGMOD international conference on Management of data*2009, ACM: Providence, Rhode Island, USA. p. 127-138.
65. Cormode, G., *Personal privacy vs population privacy: learning to attack anonymization*, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*2011, ACM: San Diego, California, USA. p. 1253-1261.
66. Adam Grove, J.H.a.D.K., *Random Worlds and Maximum Entropy*. In Proc. 7th IEEE Symp. on Logic in Computer Science, 1994. **2**: p. 22-33.
67. Iyengar, V.S., *Transforming data to satisfy privacy constraints*, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*2002, ACM: Edmonton, Alberta, Canada. p. 279-288.
68. Proietti, G. and C. Faloutsos, *Analysis of range queries and self-spatial join queries on real region datasets stored using an R-tree*. Knowledge and Data Engineering, IEEE Transactions on, 2000. **12**(5): p. 751-762.
69. Ye, Y., et al., *Decomposition: Privacy Preservation for Multiple Sensitive Attributes*, in *Proceedings of the 14th International Conference on Database Systems for Advanced Applications*2009, Springer-Verlag: Brisbane, Australia. p. 486-490.
70. Fang, Y., M.Z. Ashrafi, and S.K. Ng, *Privacy beyond single sensitive attribute*, in *Proceedings of the 22nd international conference on Database and expert systems applications - Volume Part I*2011, Springer-Verlag: Toulouse, France. p. 187-201.
71. Wang, K. and B.C.M. Fung, *Anonymizing sequential releases*, in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*2006, ACM: Philadelphia, PA, USA. p. 414-423.
72. Yeye, H., S. Barman, and J.F. Naughton. *Preventing equivalence attacks in updated, anonymized data*. in *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. 2011.
73. Wong, R.C.W., et al. *Global privacy guarantee in serial data publishing*. in *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. 2010.
74. Byun, J.-W., et al., *Privacy-preserving incremental data dissemination*. J. Comput. Secur., 2009. **17**(1): p. 43-68.

75. Shmueli, E., et al., *Limiting disclosure of sensitive data in sequential releases of databases*. Inf. Sci., 2012. **191**: p. 98-127.
76. Peano, G., *Sur une courbe, qui remplit toute une aire plane*. Mathematische Annalen, 1890. **36**: p. 157-160.
77. Orenstein, J.A. and T.H. Merrett, *A class of data structures for associative searching*, in *Proceedings of the 3rd ACM SIGACT-SIGMOD symposium on Principles of database systems*1984, ACM: Waterloo, Ontario, Canada. p. 181-190.
78. OVERSTEEGEN, J.M.A.A.L.G., *The dynamics of the Sierpinski curve*. proceedings of the american mathematical society, 1994. **120**.
79. Jagadish, H.V., *Linear clustering of objects with multiple attributes*, in *Proceedings of the 1990 ACM SIGMOD international conference on Management of data*1990, ACM: Atlantic City, New Jersey, United States. p. 332-342.
80. Moon, B., et al., *Analysis of the clustering properties of the Hilbert space-filling curve*. Knowledge and Data Engineering, IEEE Transactions on, 2001. **13**(1): p. 124-141.
81. Taylor, K., S. Rickard, and K. Drakakis, *Costas Arrays: Survey, Standardization, and MATLAB Toolbox*. ACM Trans. Math. Softw., 2011. **37**(4): p. 1-31.
82. Golomb, S.W. and H. Taylor, *Constructions and properties of Costas arrays*. Proceedings of the IEEE, 1984. **72**(9): p. 1143-1163.
83. Silverman, J., V.E. Vickers, and J.M. Mooney, *On the number of Costas arrays as a function of array size*. Proceedings of the IEEE, 1988. **76**(7): p. 851-853.
84. Xiao, X. and Y. Tao, *Anatomy: simple and effective privacy preservation*, in *Proceedings of the 32nd international conference on Very large data bases*2006, VLDB Endowment: Seoul, Korea. p. 139-150.

VITA

- 2007-2012 PhD student in computer science and engineering, Sabanci University, Istanbul, Turkey.
Supervisor: Prof. Dr. Yucel Saygin
- 2006 MA degree in computer science from VUB, Brussels, Belgium. Title of Master's thesis: *Techniques for Personalization in E-learning*.
Supervisor: Prof. Dr. Ir. Geert-Jan Houben
- 2001 BS degree in computer engineering from IUG, Palestine. Title of Bachelor's project: *distributed banking system*.
Supervisor: Prof. Dr. Hatem HAMAD
- 1996 Secondary school: Palestine secondary school, Gaza, Palestine.
- 1978 Born on 27th of July, 1978 in Gaza, Palestine.