Study of Data Correlation's impact on Privacy Leakage in Privacy Preserving Models

Submitted in partial fulfillment of the requirements

for the award of the degree of

DOCTOR OF PHILOSOPHY

Submitted by

Hemkumar D

(Roll No. 716043)

Under the guidance of

Dr. S. Ravichandra

and

Prof. D.V.L.N. Somayajulu

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL
TELANGANA - 506004, INDIA

October 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL
TELANGANA - 506004, INDIA

THESIS APPROVAL FOR Ph.D.

This is to certify that the thesis entitled, Study of Data Correlation's impact on Privacy Leakage in Privacy Preserving Models, submitted by Mr. Hemkumar D [Roll No. 716043] is approved for the degree of DOCTOR OF PHILOSOPHY at National Institute of Technology Warangal.

Examiner

Research Supervisor
Dr. S. Ravichandra
Associate Professor
Dept. of Computer Science and Engg.
NIT Warangal &
India

Research Supervisor
Prof. D.V.L.N. Somayajulu
Professor
Dept. of Computer Science and Engg.
NIT Warangal
India

Chairman
Prof. P. Radha Krishna
Professor
Head, Dept. of Computer Science and Engg.
NIT Warangal
India

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL
TELANGANA - 506004, INDIA

CERTIFICATE

This is to certify that the thesis entitled, Study of Data Correlation's impact on Privacy Leakage in Privacy Preserving Models, submitted in partial fulfillment of requirement for the award of degree of DOCTOR OF PHILOSOPHY to National Institute of Technology Warangal, is a bonafide research work done by Mr. Hemkumar D [Roll No. 716043] under our supervision. The contents of the thesis have not been submitted elsewhere for the award of any degree.

Research Supervisor
Dr. S. Ravichandra
Associate Professor
Dept. of Computer Science and Engg.
NIT Warangal &
India

Warangal
Date: 25-10-2021

Research Supervisor
Prof. D.V.L.N. Somayajulu
Professor
Dept. of Computer Science and Engg.
NIT Warangal
India

Warangal
Date: 25-10-2021

DECLARATION

This is to certify that the work presented in the thesis entitled "*Study of Data Correlation's impact on Privacy Leakage in Privacy Preserving Models*" is a bonafide work done by me under the supervision of Dr. S.Ravichandra and Prof. D.V.L.N. Somayajulu and was not submitted elsewhere for the award of any degree.

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/date/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Hemkumar D

(Roll No. 716043)

Date: 25-10-2021

ACKNOWLEDGMENTS

"Difficulties in your life do not come to destroy you but to help you realise your hidden potential and power. Let difficulties know that you too are difficult" - A. P. J Abdul Kalam

Every day during my Ph.D. has been a great opportunity for learning. This thesis work is the result whereby I have been supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

First and foremost, I would like to thank the Lord almighty for glorifying me with all the strength and health to carry out my research work. It is with great pleasure that I acknowledge my sincere thanks and deep sense of gratitude to my supervisors Dr.S. Ravichandra and Prof. D.V.L.N. Somayajulu for their valuable guidance throughout the course. Their technical perception, profound knowledge, sincere effort in guiding a student and devotion to work have greatly charmed and constantly motivated me to work towards the goal. They always gave me ample time for discussions and suggesting requisite corrections.

I extend my gratitude to the Doctoral Scrutiny Committee (DSC) members comprising of Prof. BB Amberker, Dr. Rashmi Ranjan Rout, Prof. D. Srinivasacharya for their insightful comments and suggestions during oral presentations. I am lucky to attend lectures by Prof. B. B. Amberker during my tenure. I am immensely thankful to Dr. Ch. Sudhakar, Prof. R. B. V. Subramanyam and Prof. P. Radha Krishna Heads of Dept. of CSE and chairmans of DSC, during my tenure for providing adequate facilities in the department to carry out the oral presentations.

I wish to express my sincere thanks to Prof. N.V. Ramana Rao, Director, NIT Warangal for providing the infrastructure and facilities to carry out the research. I am also very much grateful to the faculty members of Computer Science and Engineering Department for their moral support throughout my research work.

On the personal level, I would also like to thank my scholar friends in NIT

Warangal for their valuable suggestions and for extending selfless cooperation. Lastly, my gratitude to my family for their unconditional love, support and prayers for my success in achieving the goal.

Hemkumar D

Dedicated to

My Family & Teachers

ABSTRACT

The rapid growth in the usage of location-based services has resulted in extensive research on users' trajectory data publishing or on continuous publications of location statistics. The users' trajectory information or location statistics are provided valuable knowledge that supports various social benefits such as smart healthcare, real-time traffic monitoring, online advertisement, etc. Even publishing or sharing such users' trajectory or location statistics without preserving users' privacy. Then the participated users in a published dataset may presume that a malicious adversary can breach participated users' privacy because it contains users' private information like disease, habits, etc. In literature, there exist privacy preservation models to provide a privacy guarantee to users against an efficient adversary, namely Data anonymization, ϵ-Differential Privacy (ϵ-DP), and ϵ-Local Differential Privacy (ϵ-LDP). Data anonymization model protects users' sensitive information from record/identity linkage, attribute linkage, and table linkage attack, whereas ϵ-Differential Privacy, and ϵ-Local Differential Privacy address probabilistic attack.

The above privacy preserving models preserve users privacy by assuming that data records are independent from each other. However, in reality data records are not independent (or correlated) which leads to achieve less privacy guarantee as compared to traditional privacy preserving models. In other words, if an adversary has additional knowledge about the correlated records, then these privacy models does not prevent all participated users privacy. Specifically, single Data Anonymization approach is not addressed the correlated-records linkage attack along with three common linkage attacks, namely identity linkage, attribute linkage, and similarity attack. Further, the traditional DP and LDP mechanisms are not provide sufficient privacy guarantee especially the the data records are in correlated in nature.

To address the correlation behavior challenges in all three privacy preservation models, firstly we proposed a data anonymization approach to protect users' privacy against a correlated-records linkage attack along with three common linkage attacks, namely identity linkage, attribute linkage, and similarity attack. The proposed method consists of two phases, namely virtualization and suppression. The virtualization method works as a replacement mechanism for the sensitive attribute, and the suppression method works as an anonymization mechanism for users' trajectories, in order to anonymize the trajectory datasets for preserving users' privacy from the above four linkage attacks. Secondly, we presented a reformulated Differential Privacy definition to quantify the impact of temporal correlation on privacy leakage. We also introduced a privacy budget allocation method for allocating an adequate amount of privacy budget to each successive timestamps under the protection of differential privacy. Thirdly, the ϵ-Local Differential Privacy also suffers more privacy leakage when the dataset involves temporal correlation. We propose a privacy budget allocation method to allocate a sufficient amount of privacy budget at each timestamp and prove that our proposed method achieves ϵ-LDP over an infinite length of a user stream. Fourthly, we quantify the impact of data correlation on privacy leakage in a combined (LDP+DP) approach. Since this combined (LDP+DP) approach having many possible combinations by considering with or without temporal correlation, it is necessary to study the impact of data correlation on privacy leakage in all possible combinations of the combined (LDP+DP) approach.

In this thesis, we addressed the correlation challenge in privacy preserving models and proposed possible privacy preserving methods under Data anonymization, ϵ-DP, and ϵ-LDP models against the correlation issues. Finally, we evaluate the data utility of all proposed methods by presenting experimental results for real and synthetic data sets.

Contents

ACKNOWLEDGMENTS i

ABSTRACT iv

List of Figures xi

List of Tables xvii

List of Algorithms xviii

List of Notations xx

List of Abbreviations xxiii

1 Introduction 1

1.1 Overview of the Contribution of this Thesis 5

1.1.1 Quantify the Impact of Prior Knowledge on Privacy Leakage in Trajectory Data Publishing 5

1.1.2 Quantify the Impact of Data Correlation on Privacy Budget Allocation in ϵ-Differential Privacy 7

1.1.3 Quantify the Impact of Data Correlation on Privacy Leakage in ϵ-Local Differential Privacy 8

1.1.4 Compare the Impacts of Data Correlation on Privacy Leakage in a Combined Privacy Preserving Approaches 10
1.2 Organization of the Thesis 11

2 Preliminaries **13**
2.1 *Notations* 13
2.2 *Definitions* 15
2.2.1 *Data anonymization* 15
2.2.2 *Differential privacy* 16

3 Literature Survey **18**
3.1 Privacy Preserving Models 19
3.1.1 Data Anonymization Approach 19
3.1.1.1 Clustering based anonymization methods 20
3.1.1.2 Quasi identifiers based anonymization methods . 21
3.1.1.3 Location Privacy 23
3.1.2 ϵ-Differential privacy 24
3.1.2.1 Trajectory or Location Privacy without correlation 25
3.1.2.2 Trajectory or Location Privacy with correlation . 26
3.1.3 ϵ-Local differential privacy 30
3.1.3.1 ϵ-LDP without correlation 30
3.1.3.2 ϵ-LDP with correlation 31
3.2 Budget Allocation Methods 33

4 Quantify the Impact of Prior Knowledge on Privacy Leakage in Trajectory Data Publishing **35**
4.1 System Framework 38

4.1.1 *Adversary model* . . . 38
4.1.2 *Privacy Requirement* . . . 40
4.1.2.1 *Virtualization* . . . 41
4.1.2.2 *Suppression* . . . 42
4.1.3 *Utility Metrics* . . . 42
4.2 Proposed Method . . . 43
4.2.1 Correctness Proof of Algorithms . . . 49
4.3 Experimental Results . . . 50
4.3.1 Metrics of Evaluation . . . 51
4.3.2 Comparison . . . 55
4.3.3 Complexity Analysis . . . 57
4.4 Summary . . . 58

5 Quantify the Impact of Data Correlation on Privacy Budget Allocation in ϵ-Differential Privacy **59**

5.1 System Framework . . . 64
5.1.1 Differential Privacy under continual observation . . . 64
5.2 Temporal correlation (TC) privacy leakage analysis . . . 65
5.2.1 Adversary's knowledge . . . 65
5.2.2 TC privacy leakage . . . 66
5.3 Proposed Method . . . 69
5.3.1 Privacy Analysis . . . 73
5.3.2 Utility Analysis . . . 75
5.4 Experimental Results . . . 76
5.4.1 Metrics of Utility Evaluation . . . 77
5.4.2 Result Analysis . . . 78

5.4.3 Compare with Baseline approaches 80
5.4.3.1 Analysis and Evaluation 82
5.5 Summary 85

6 Quantify the Impact of Data Correlation on Privacy Leakage in ϵ-Local Differential Privacy 87
6.1 System Framework 91
6.1.1 Local differential privacy under continual observation 91
6.2 Threat model: privacy leakage analysis 93
6.2.1 Temporal correlation (TC) privacy leakage analysis 93
6.3 The Proposed Algorithm 99
6.3.1 Privacy Analysis 104
6.4 Experimental Results 106
6.4.1 Impact of correlation on privacy leakage 106
6.4.1.1 Correlation Behavior 107
6.4.1.2 Privacy leakage under different degrees of correlation 108
6.4.2 Utility Evaluation 109
6.5 Summary 115

7 Compare the Impacts of Data Correlation on Privacy Leakage in a Combined Privacy Preserving Approaches 116
7.1 System Framework 119
7.1.1 Differential Privacy under continual observation 120
7.1.2 Local Differential Privacy under continual observation 120
7.1.3 Privacy leakage analysis for non-correlated dataset of DP and LDP 121

7.1.4 Privacy leakage analysis for temporally correlated dataset of DP and LDP . . . 121
7.2 Comparative analysis . . . 125
7.2.1 LDP and DP . . . 126
7.2.2 LDP($\mathcal{TC}$) and DP . . . 126
7.2.3 LDP and DP($\mathcal{TC}$) . . . 127
7.2.4 LDP($\mathcal{TC}$) and DP($\mathcal{TC}$) . . . 127
7.3 Experimental Results . . . 128
7.4 Summary . . . 140

8 Conclusion and Future Scope **142**
8.1 Conclusion of the thesis . . . 142
8.2 Future Scope . . . 144

Bibliography **146**

List of Publications **157**

List of Figures

4.1 A Taxonomy Tree of Human diseases 40

4.2 The average privacy risk of users with respect to A's prior knowledge of various lengths while fix $\sigma = 0.5$. a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset.. 53

4.3 The average trajectory information loss in T' with various K threshold values a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset. 53

4.4 The average query-answer error rate with various A's prior knowledge a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset. 54

4.5 Effect of Privacy-Height Γ threshold on users trajectory-information loss for different K-anonymity values a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset. 54

4.6 The average frequency of PoIs visited between the original dataset and the anonymized datasets published from the proposed method, KCL-PPTD, KCL-local and KCL-Global. a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset. 56

4.7 The average trajectory information loss in T' with various K threshold values a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset. 57

4.8 The run-time performance between the proposed privacy approach with KCL-PPTD, KCL-local and KCL-Global a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset. 58

5.1 Illustration of example 1 (a) Collection of users location data-points in continuous timestamps (b) Statistics for event-level or user-level privacy (c) Statistics for w-event privacy while set w = 3 61

5.2 Distribution of privacy budget over timestamps(or event) within the sliding window of size $w = 3$. 62

5.3 Distribution of privacy budget over timestamps(or event) within the sliding window of size $w = 3$. 73

5.4 Analysis of privacy risk (or leakage) of 1-DP under different types of temporal correlation 79

5.5 Privacy leakage versus different degrees of correlation while set $\epsilon = 1$ (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 79

5.6 MAE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 80

5.7 MSE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 81

5.8 MAE vs. ϵ while fixing w=40 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 81

5.9 MSE vs. ϵ while fixing w=40 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 82

5.10 MAE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 83

5.11 MSE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 83
5.12 MAE vs. ϵ while fixing w=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 84
5.13 MSE vs. ϵ while fixing w=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets. 84

6.1 Distribution of privacy budget over timestamps(or event) when the user stream's data-points are correlated within the sliding window of size $w = 3$. 89
6.2 Transition probabilities between the data-points 94
6.3 Distribution of privacy budget over timestamps(or event) within the sliding window of size $w = 3$. 103
6.4 Privacy leakage vs Timestamps under different types of correlation (a) Strong correlation (b) Moderate correlation (c) No correlation 107
6.5 Privacy leakage vs different degrees of correlation while set $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 108
6.6 Privacy leakage vs different degrees of correlation while set $\epsilon = 0.1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 108
6.7 MAE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 110
6.8 MSE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 110
6.9 MAE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 111
6.10 MSE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 112

6.11 MAE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 113
6.12 MSE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 113
6.13 MAE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 114
6.14 MSE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 114

7.1 Different combinations of combined (LDP+DP) approach assuming with or without temporal correlation 125
7.2 MAE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 130
7.3 MSE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 131
7.4 MAE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 131
7.5 MSE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 132
7.6 MAE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 133
7.7 MSE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 133

7.8 MAE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 134
7.9 MSE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 134
7.10 MAE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 135
7.11 MSE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 136
7.12 MAE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 136
7.13 MSE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 137
7.14 MAE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 137
7.15 MSE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 138

7.16 MAE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 139
7.17 MSE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets. 139

List of Tables

3.1 List of privacy preservation approaches which prevents different types of linkage attacks. 24

4.1 A Hospital trajectory dataset 36
4.2 Virtual-Trajectory dataset 41
4.3 Anonymized Trajectory dataset $\mathcal{T}'$ 44

5.1 Transition probability matrix and sample dataset 66
5.2 The privacy guarantee of various privacy budget allocation methods on temporally correlated data-points of length T 85

7.1 The privacy guarantee of combined approaches by adopting various methods of LDP and DP under temporal correlation 140

List of Algorithms

4.1 Pseudocode of SaV() 45

4.2 Pseudocode of EVsV() 46

4.3 Pseudocode of ATdb() 48

5.1 Pseudocode of PBA mechanism at k^{th} timestamp ($\mathcal{M}_k$) 71

6.1 Pseudocode of $\mathcal{M}_k$ in PBA 101

List of Notations

$\mathcal{T}$	Trajectory dataset
T_i	Trajectory of user i
x_i^n	n^{th} position data-point of user i
$(T_i)^m$	Sequence of data-points upto m^{th} position of user i
$\|T_i\|$	Total number of data-points in user $i's$ trajectory
τ_i	Sub-trajectory of user i
$T_i(s)$ or s_i	Sensitive value of user i
$\mathcal{S}$	Sensitive attribute domain
A	Adversary
η	Represented as Taxonomy tree
ν	Set of nodes
ℓ	Set of node's level
$\ell(\nu_i)$	Returns a level of sensitive value of user i
$\hbar(\nu_i)$	Family of ancestor of node ν_i
$\hbar^k(\nu_i)$	k^{th} ancestor of node ν_i
Γ	Privacy Height threshold
∂	Adversary's prior knowledge
ρ	Adversary knows the maximum ρ number of moving points
s^v	Virtual sensitive value
$\mathcal{T}^v$	Virtual trajectory dataset

$\mathcal{T}'$	Anonymized Trajectory dataset
$(\mathcal{T})^s$	Set of data records having same sensitive value s
$\Upsilon(x^n)$, $\Upsilon(\tau_i)$	Suppression score of a data-point and a sub-trajectory respectively
σ	Privacy breach threshold
$\beta(T_i)$	Privacy breach of a data record T_i
$IL(T_i)$	Information loss of a data record T_i
q	Counting Query
μ	Query Answering mechanism
n^ρ	Set of discrete sub-trajectory of length ρ
χ	Universe of all possible location data-points
ϵ	Privacy Budget
Δq	Global sensitivity of query q
D_t	Dataset at timestamp t
S	Stream prefix of location data-points
S_t	Stream prefix S up to t timestamp
$S[i]$	Dataset at timestamp i
a_i	True dataset output at time t
ω_i	Noisy dataset output at time t
$\mathcal{M}$	Privacy Mechanism
w	Size of sliding window
L	Data-point vector size
$\widehat{x}_i^t$	Noisy data-point at time t
$\mathcal{M}$	Privacy Mechanism
w	Size of sliding window
L	Data-point vector size
$\widehat{x}_i^t$	Noisy data-point at time t

Abbreviations

QID	Quasi Identifiers
NWA	Newer Walk Alone
PPTD	Preserving Privacy in Trajectory Data publishing
LKC	Length of trajectory, K-anonymity threshold, Confidence bounding
KCL/$(K, C)_L$	K-anonymity threshold, Confidence bounding, Length of trajectory
DP	Differential Privacy
DDP	Dependent Differential Privacy
PTCP	Private Trajectories Calibration and Publication System
BA	Budget Absorption
BD	Budget Distribution
ESA	Encode, Shuffle and Analyze
MC	Markov Chain process
TC	Temporal Correlation
PBA	Privacy Budget Allocation
PIM	Planar Isotropic Mechanism
MAE	Mean of Absolute Error
MSE	Mean of Square Error

W-RR	W-Randomized Response
K-RR	K-Randomized Response
LRM	Local Randomize Matrix
GRR	Generalized Randomized Response
LDP	Local Differential Privacy

Chapter 1

Introduction

The advancements in the usage of location-aware devices such as GPS mobile phones, RFID tags, has facilitated the easy collection of the user's spatio-temporal data-points. The sequence of spatio-temporal data-points of a user known as user's trajectory, can be used in various applications such as real-time traffic monitoring [1], smart healthcare [2], online advertisement [3] etc. Further, numerous applications require continuous publication of location statistics for providing various social benefits to users or to support various decision-making purpose. However, sharing such user's trajectory or location statistics without preserving users' privacy may lead to serious mistrust between the users and the data published organization [4]. This is because, the published data contains users' sensitive information like disease, habit's etc. For instance, a hospital wants to share a collected radio frequency identification (RFID) data of patients who suffer from a specific disease with the researchers for the analysis of disease and disorder outbreaks. Due to the inadvertent sharing of patients' data, the researchers can identify and disclose the type of disease a targeted patient suffers from, which leads to compromising patients' privacy. Therefore, the privacy mechanism is required to prevent the users privacy.

However, the privacy mechanisms provide privacy to participated users with a key concern, which is *"Access to the published dataset should not allow the adversary to learn anything extra about any target victim compared to no access to the dataset, even with the presence of any adversary's background knowledge obtained from other sources"*. There exist few privacy models for preserving users privacy namely Data Anonymization, ϵ-Differential Privacy and ϵ-Local Differential Privacy. Data Anonymization is the process of modifying the relation in such a way that minimal user's private data may be inferred. It protects the user's private data by erasing or removing identifiers that connect a user to stored data. Even though removed user identity, an efficient adversary can infer a user's record (or targeted victim) with high probability by using his/her prior (or background) knowledge. There are many techniques to achieve a privacy guarantee in data anonymization, such as generalization [5], suppression [6], perturbation [7] and clustering [8]. There are many privacy preservation approaches in the literature to provide privacy against either single or a combination of linkage attacks such as identity linkage attack, attribute linkage attack, and similarity linkage attack.

The ϵ-Differential Privacy (ϵ-DP) [9] is a novel privacy mechanism for query answering. It is proved that DP provides strong privacy guarantees to users against an adversary with unbounded knowledge. It ensures that any user's privacy leakage is to be strictly bounded by at most a ϵ value, where ϵ is a user parameter. If the value of ϵ is small, it achieves a strong privacy guarantee and vice versa. The ϵ-DP releases a noisy output instead of true output for hiding user's sensitive information. This noisy output is computed by adding a random noise (derived from the Laplace distribution with scale λ) to the true output. On the other hand, Local differential privacy (LDP) [10] is a variant of standard differential privacy. The LDP mechanism address that most of the existing approaches under ϵ-DP assume that the service provider collects the user's sensitive information, adds noise to the

output for a particular query, and releases the perturbed output. However, it is unfeasible to assume that all service providers are trusted, which leads to untrusted service providers misuse the collected information for other purposes [11, 12]. To overcome this, the LDP mechanism follows a scheme which is, users make their location data-point private before sending it to the service provider. Hence, this privacy model promises a privacy guarantee to users even though the service provider is not trusted. Under the protection of LDP, the service provider can still compute the correct statistical results even though not collecting users' private location data-points. According to the ϵ-LDP, the adversary cannot infer users' private or sensitive data with high confidence (controlled by ϵ). Here, the ϵ is a privacy budget that controls the level of privacy guarantee.

Further, The above these three privacy preserving models also been applied in continuous data publishing settings. Assume that the data records of individual are not independent (or correlated) in continuous data publishing settings. Then, the data anonymization models run out to maintain the trade-off between utility and privacy because of correlation and the DP and LDP mechanisms are provide less privacy guarantee, especially when the user's records are not independent (i.e., correlated) or user's data-points are not independent between consecutive timestamps (i.e., temporally correlated). This challenge is motivated us to study and propose possible solutions for the above privacy models.

The major contribution in this thesis are as follows,

- **Quantify the impact of prior knowledge on privacy leakage in trajectory data publishing:** This work present an Data Anonymization method that prevents users sensitive information from the four linkage attacks, namely correlated-records linkage attack, Identity linkage attack, Attribute linkage attack and Similarity attack. In addition, the Virtualization method is introduced in proposed method for sensitive

attributes to achieve no sensitive attribute information loss in the published dataset.

- **Assess the impact of data correlation on privacy budget allocation in continuous publication of location statistics:** This work explores the effect of data correlation on privacy leakage in ϵ-Differential Privacy, showing that if a temporal correlation occurs between the data-points of a user stream, the privacy guarantee of the ϵ-Differential Privacy mechanism can be violated. It also presents a Privacy Budget Allocation strategy to allocate privacy budgets to the correlated location data-points to achieve ϵ-DP in continuous location data publishing.

- **Quantify the impact of data correlation on privacy leakage in Local differential privacy for Continuous Data Release Settings** This work analyzes the impact of data correlation on privacy leakage in ϵ-Local Differential Privacy and shows that ϵ-Local Differential Privacy mechanism could be degrading the privacy guarantee if the temporal correlation exists between the data-points of a user stream. A Privacy Budget Allocation (PBA) mechanism is proposed in local settings. It allows to compute and allocate the quantity of privacy budget to each publication in continuous data release settings. Further, theoretically prove that the proposed mechanism achieves ϵ-LDP.

- **Compare the impacts of data correlation on privacy leakage in a combined privacy preserving approach:** This work compares the privacy leakage of a combined traditional (LDP+DP) approach with a combined temporally correlated (LDP+DP) approach. And evaluate the data utility of all four different types of combined (LDP+DP) approaches.

1.1 Overview of the Contribution of this Thesis

In this section, an overview of chapter-wise contributions of this thesis has been presented. Each subsection presents summary of contributions of the corresponding chapter.

1.1.1 Quantify the Impact of Prior Knowledge on Privacy Leakage in Trajectory Data Publishing

The various linkage attacks are possible in trajectory data publishing such as Identity linkage attack, Attribute linkage attack and Similarity attack. Therefore, privacy is very important to preserve users privacy from the efficient adversary.

The adversary model for breaching user's privacy is characterized by two important specification (1) Adversary's prior (or background) knowledge, (2) Adversary's goal. We assume that adversary can gain prior knowledge about the target victim from various sources and his goal is to breach the sensitive information of target victim by linking his prior knowledge to the published trajectory dataset. Let A be the adversary, that could be a data analyst or data collector itself and his aim is to find the record or sensitive value of the target victim in published dataset. Based on the A's prior knowledge about the targeted victim, A could perform following linkage attacks, to breach victim's record or his sensitive value.

Identity linkage attack: If the targeted victim's record is unique in the published dataset, then an adversary can identify a victim's data-record by using his prior knowledge and consequently he can find victim's sensitive information.

Attribute linkage attack: This attack happens only when the sensitive value of a targeted victim is occurring more frequently. Then there is a possibility that an adversary could

breach the sensitive information even though the adversary does not have unique trajectory information of a targeted victim.

Similarity attack: It happens only when an adversary succeeds to identify all possible sensitive values of the targeted victim by using his prior knowledge. If the identified sensitive values are semantically similar, then the adversary can breach a generalized sensitive value of the targeted victim.

Correlated-records linkage attack: In the real dataset, a targeted victim may have multiple data-records in the published dataset. If an adversary has additional knowledge about the correlated records, then an adversary can predict the victim's sensitive value with high confidence.

The existing mechanism under data anonymization model in the literature to provide privacy against three common linkage attacks either single or combination of linkage attacks such as identity linkage attack, an attribute linkage attack and similarity linkage attack. However, the correlated-records linkage attack has not been addressed in existing data anonymization approaches. Although, there is no privacy preservation approach to address all the above four linkage attacks. Further, ϵ-Differential Privacy (ϵ-DP) is a new privacy notion and is defined as a property of a query answering mechanism[13]. However, publishing trajectory dataset with DP may not able to provide data truthfulness in published trajectory dataset[14, 15]. This is due to the resulted output (or trajectory dataset) is untruthful because of uncertainty (eg. Laplace noise [13]) introduced for achieving ϵ-DP. Therefore, it is required to develop a data anonymization method to prevent four different types of linkage attacks against efficient adversary. The contributions of this work as follows.

- We propose a privacy preservation anonymization method that preserves users' privacy from four linkage attacks by enabling the virtualization method on the user's

sensitive values and suppression method on the user's trajectory.

- We introduced a privacy threshold called privacy-height in our method, and it is used to fix the upperbound to the privacy-risk.
- Finally, we compare the rate of information loss between previous approaches with the propose anonymization approach by conducting an experiment on the synthetic and real-time datasets.

1.1.2 Quantify the Impact of Data Correlation on Privacy Budget Allocation in ϵ-Differential Privacy

ϵ-Differential Privacy (ϵ-DP) [9] is a query-answering privacy mechanism. It is proved that DP provides strong privacy guarantees to users against an adversary with unbounded knowledge. It ensures that any user's privacy leakage is to be strictly bounded by at most a ϵ value, where ϵ is a user parameter. Although, it has been applied in settings of continuous data publishing [16, 17, 18].

In the literature, there exist a few privacy approaches under the protection of ϵ-DP for continuous data publishing., such as Event-level privacy[19], User-level privacy[20] and w-event privacy[17]. The Event-level privacy and User-level privacy has limited applicability in most of the real-world applications. Recently, w-event privacy mechanism has been proposed to address the limited use of event-level privacy and user-level privacy. This mechanism offers a strong privacy guarantee to any user stream within a window of w timestamps and achieves ϵ-DP on infinite length of an user stream. However, the w-event privacy mechanism provides less privacy guarantee than traditional ϵ-DP, especially when the user's data-points are not independent (i.e., temporally correlated) between consecutive timestamps. It happens due to the allotted privacy budget at timestamps within

a window of size w is not adequate, especially where the data-points of users' stream involve temporal correlation. The proposed allocation scheme involves two phases and these two phases operate sequentially by using half of the total privacy budget. In the first phase, calculates a dissimilarity value between the true statistic and last release private statistic. Then, the obtained dissimilarity value is forwarded into the second phase. The second phase divides a privacy budget into two parts: publication privacy budget and absorption privacy budget. The second phase decides whether to publish a true publication with noise or null publication (last release private output). The contributions of this work are as follows.

1. We present a reformulated differential privacy definition for continuous data publication and prove that it can achieve ϵ-DP. Then we quantify the impact of temporal correlation on privacy leakage in reformulated ϵ-DP and analyze the privacy leakage in ϵ-DP with a numerical example.

2. We introduce a Privacy Budget Allocation method for allocating an adequate amount of privacy budget to each successive timestamps under the protection of ϵ-Differential privacy.

3. Finally, we evaluate the data utility of our method by computing the average error per timestamps through conducting a series of experiments on real and synthetic datasets.

1.1.3 Quantify the Impact of Data Correlation on Privacy Leakage in ϵ-Local Differential Privacy

The Local-Differential Privacy mechanism address the privacy issues at user side of an real-time application. To achieve this, users make their location data-point private before

sending it to the service provider. Hence, this privacy model promises a privacy guarantee to users even though the service provider is not trusted. A few basic privacy approaches are existed in literature for continuous data publishing, namely Event level, User-level, and w-event privacy. The w-event privacy under the protection of ϵ-LDP offers a privacy to the user's stream of length *infinity* by using a sliding window methodology. However, w-event privacy achieves comparatively less privacy guarantee than that traditional ϵ-LDP especially when a user stream's location data-points are correlated. This is because the privacy budget allots to each timestamp's data-point are not adequate due to the presence of correlated location data points within the user stream. Therefore, it is necessary to design a privacy budget allocation scheme under the protection ϵ-LDP for allocating privacy budgets to the correlated location data-points within the user's stream. The proposed allocation strategy within the sliding window involves two phases. These two phases operate sequentially by using independent randomness. At the first phase of any timestamp, compute dissimilarity between current location data-point and last release noisy data-point. This dissimilarity value is made private by using the allotted privacy budget, and this private dissimilarity value is forwarded into phase 2. The second phase decides whether to publish the current location data-point or not. If the second phase decides not to publish the current location data-point, then the allotted privacy budget becomes free and can be used for future publication if necessary. On the other hand, the second phase decides to publish the current location data-point, then it absorbs one extra privacy budget that became available from previous skipped publication if and only if the correlation exists between the current location data-point and last published location data-point. Otherwise, it uses only the allotted privacy budget for publication. The contribution of this work is as follows.

1. We present a definition ϵ-Local Differential Privacy for continuous data publication

and prove that it can achieve ϵ-LDP. We quantify the privacy degradation when correlation exists in continuous data publication and analyze the privacy leakage with a numerical example.

2. We propose a Privacy Budget Allocation method on ϵ-Local Differential Privacy for distributing an adequate amount of privacy budgets to each timestamp's data under the protection of ϵ-LDP.

3. Finally, we demonstrate the effectiveness of our proposed method in terms of data utility with existing allocation methods by considering real and synthetic datasets.

1.1.4 Compare the Impacts of Data Correlation on Privacy Leakage in a Combined Privacy Preserving Approaches

In the era of digitization, some applications require users data from both data collection and data sharing phases to provide better social benefits to users such as smart health monitoring system, smart traffic control systems etc. The service provider collects the user's private data and provides the services to users or shares that collected private (or sensitive) data with other service providers for providing better social benefits to users. However, if it combined collected users' data and shared users' data may compromise the user's privacy, leading to disclosing the user's sensitive information. Many privacy preservation methods have been proposed for providing a privacy guarantee either at the time of data collection or data sharing phases. ϵ-Differential Privacy provides a strong privacy guarantee with the assumption that service providers are trustworthy and it is used in data sharing phase. Since it is difficult to presume that all service providers are trusted, a variant of standard Differential privacy for local settings has been proposed, named as ϵ-Local Differential Privacy and it is used in data collection phase. There are few real-

time applications require either a privacy guarantee by the data collector (i.e., DP) or a privacy guarantee by the data provider itself (i.e., LDP) or a privacy guarantee by both data provider and data collector (i.e., LDP and DP).

There is limited works to study the impact of data correlation on privacy leakage either only DP involved in the application or only LDP involved in the application, but not both mechanisms involved in the application. So, it is necessary to study the impact of data correlation on privacy leakage of a combined approach. There are four different combinations of a combined approach, such as either it requires (traditional LDP + traditional DP) or (LDP with temporal correlation(TC)+ traditional DP) or (traditional LDP + DP with temporal correlation) or (LDP with temporal correlation + DP with temporal correlation). Depends on the type of query requirement, the curator chooses one amoung the four different combinations of a combined approach and release the statistics. The contributions of this work are as follows.

1. We quantify the impact of data correlation on privacy leakage of all cases of a combined approach in continuous data release settings.

2. We performed a series of experiments with real and synthetic datasets to determine the average error rate per timestamp for evaluating the data utility of a combined approach (LDP with TC+DP with TC) with other states of the art methods.

1.2 Organization of the Thesis

The main focus of this thesis is to analyze the impact of adversary's prior knowledge (i.e., Data correlation) on privacy leakage in various privacy-preserving models. The proposed algorithms achieve a strict privacy guarantee in location data publishing. The rest of the thesis has been organized into six chapters.

Chapter 2: In this chapter, Data Anonymization, ϵ-Differential Privacy and ϵ-Local Differential Privacy enabled privacy preserving approaches have been discussed.

Chapter 3: A correlated-records linkage attack has been discussed and proposed a privacy reserving anonymization approach to achieve a privacy guarantee against correlated-records linkage attack along with three common linkage attacks, namely Identity, Attribute and Similarity.

Chapter 4: In this chapter, quantify the impact of data correlation on privacy leakage in ϵ-Differential Privacy is presented. Design a privacy budget allocation scheme for allocating a privacy budget over the correlated data-points of successive timestamps. In addition, proved that our proposed method satisfies ϵ-Differential Privacy.

Chapter 5: ϵ-Local Differential Privacy definition for continuous data publication is presented in this chapter and also, quantifies the impact of data correlation on privacy leakage in ϵ-Local Differential Privacy in continuous data release setting A Privacy Budget Allocation scheme under the protection of ϵ-Local Differential Privacy for distributing an adequate amount of privacy budgets to each timestamp's data is presented.

Chapter 6: This chapter presents the comparison of the privacy leakages of a combined traditional (LDP+DP) approach with a combined temporally correlated (LDP+DP) approach. And an evaluation of the data utility of all four different types of combined (LDP+DP) approaches is presented.

Chapter 7: This chapter summarizes the outcomes of the contributions and future scope for expanding the work.

Chapter 2

Preliminaries

In this section, we are discussing a set of mathematical notations that helps to understand the following chapters, the definitions of linkage attacks, differential privacy under continual observation and local differential privacy under continual observation. Then, we also discussed a privacy challenges in privacy preserving models.

2.1 *Notations*

A trajectory dataset consists of users' data-records and each data-record allows an identifier, a user trajectory and a set of sensitive values. The trajectory dataset $\mathcal{T}$ is represented as,

$$\mathcal{T} = \{(id_1, T_1, s_1), (id_2, T_2, s_2) \ldots\ldots\ldots (id_i, T_i, s_i)\}$$

Where id is a keyword which identifies user's record uniquely in $\mathcal{T}$, s_i or $T_i(s)$ is a sensitive value of a user i derived from corresponding sensitive-attribute domain and T_i is a

sequence of moving points (location and time) of user i and is represented as,

$$T_i = \{(loc_1, time_1)^i, (loc_2, time_2)^i, (loc_3, time_3)^i \ldots\ldots (loc_n, time_n)^i\}$$

Let $x_i^n = (loc_n, time_n)^i$ is a n^{th} position moving point of a user i and $(T_i)^m$ be a sequence of moving points upto m^{th} position of user i. The total number of moving points in a user $i's$ trajectory is denoted as $|T_i|$.

Let a trajectory $T_i = \{x_i^1, x_i^2, \ldots\ldots, x_i^m\}$, then the trajectory $T_j = \{x_j^1, x_j^2, \ldots\ldots, x_j^k\}$ is said to be a sub-trajectory of user i iff there exist an integer $k \leq m$ and $(T_j)^k \equiv (T_i)^k$ and is denoted as $\tau_j \subseteq T_i$. Further, any two trajectory $T_i = \{x_i^1, x_i^2, \ldots, x_i^k\}$ and $T_j = \{x_j^1, x_j^2, \ldots, x_j^k, \ldots, x_j^m\}$ can merge by using union operation and is represented as,

$$T_i \cup T_j = \{x_i^1, x_i^2, \ldots, x_i^k, x_j^{k+1}, \ldots, x_j^m\} \; if \; (T_i)^k \equiv (T_j)^k$$

Let χ be the universe of possible input location data-points. In our setting, we assume that each location data-point is a vector of size L, where L is a total number of all possible locations. Suppose, if a user is at location $l_i \in L$, then the corresponding bit (or column) i in the vector is set to be 1, and all remaining bits are set to zero. At every timestamp t, curator collects a dataset D_t with k rows, denoted as the set of indices $[k] = \{1, 2, ...k\}$. Let S be the stream prefixes of location data-points and we represent the stream prefix S up to t timestamp is $S_t = (D_1, D_2, \ldots, D_t)$. Let $q : D \rightarrow \mathbb{R}^L$ be the counting query function where D is the set of all datasets with L columns. The curator aims to publish a result (or statistics) for a counting query at each timestamp. a_i is the output of the dataset $S[i] = D_i$.

Each location data-point x_i^t is a vector of size L, $L \in \chi$. Each location data-point of a user i has a value $x_i^t = l_i^t$. This value l_i^t is private to user i. In addition, user i has visited

location information (say V_i). This V_i is private to user i, but it is correlated with current location data point of user i i.e., l_i^t. This correlation is modeled as follows. If V_i be the set of location data-points visited by the user i, then $l_i^t \sim \Theta_i$, where Θ_i is a distribution of transition probabilities over all possible location data points visited by the user i that is known to all users and mechanism. At every timestamp t, curator collects a dataset D_t with k rows.

2.2 *Definitions*

Definition 2.1 A trajectory dataset T preserves privacy for any sequence or trajectory T_i if and only if T_i satisfies $0 < |T_i| \leq \rho$ and $|T_i(\tau)| \geq K$ for any integer ρ and K , where $|T_i|$ is the total number of moving points of user $i's$ trajectory and $|T_i(\tau)|$ is the total number of data records which contains a sub-trajectory τ. In other words, an adversary (with bounded prior knowledge) cannot identify any user's trajectory T_i or sensitive value s_i with high confidence.

2.2.1 *Data anonymization*

Definition 2.2 *(Identity linkage attack:)* If the targeted victim's record is unique in the published dataset, then an adversary can identify a victim's data-record by using his prior knowledge and consequently he can find victim's sensitive information.

Definition 2.3*(Attribute linkage attack:)* This attack happens only when the sensitive value of a targeted victim is occurring more frequently. Then there is a possibility that an adversary could breach the sensitive information even though the adversary does not have unique trajectory information of a targeted victim.

Definition 2.4*(Similarity attack:)* It happens only when an adversary succeeds to iden-

tify all possible sensitive values of the targeted victim by using his prior knowledge. If the identified sensitive values are semantically similar, then the adversary can breach a generalized sensitive value of the targeted victim.

Definition 2.5*(Correlated-records linkage attack:)* In the real dataset, a targeted victim may have multiple data-records in the published dataset. If an adversary has additional knowledge about the correlated records, then an adversary can predict the victim's sensitive value with high confidence.

2.2.2 *Differential privacy*

Definition 2.6 *(Laplace mechanism:)* It is one of the most common methods for achieving ϵ-DP. Given a counting query q, the Laplace mechanism generates random noise (say x) derived from Laplace distribution with scale $Lap(\lambda)$ and is added to the true answer of a query q i.e., $\omega_t = q(D_t) + x$. The probability density function of Laplace distribution is

$$P(x) = \frac{1}{2\lambda} exp(-|x|/\lambda) \tag{2.1}$$

Where $\lambda = \Delta q/\epsilon$, Δq is a global-sensitivity of a query q, which is a maximum difference between the outputs over the adjacent stream prefixes S_t and $S_t^{'}$ i.e., $\Delta q = \max_{S_t, S_t^{'}} ||q(S_t) - q(S_t^{'})||$. The parameters Δq and ϵ plays a significant role in calibrating the noise to the query outputs.

Definition 2.7 (Adj$(S_t, S_t^{'})$:) Let S_t and $S_t^{'}$ are the stream prefixes of location data-points drawn from the χ. The S_t is adjacent to $S_t^{'}$ if and only if they are differing in one or more data-points of any one user stream prefixes. More formally, Adj$(S_t, S_t^{'})$ iff $\exists m, n \in \chi$ and $\exists k \subseteq [|S_t|]$ such that $S_t|_{k:m \to n} = S_t^{'}$. Here, k is a set of indices in the stream prefix S_t and $S_t|_{K:m \to n}$ is the result of modifying all the occurrences of m at these indices

with n.

Chapter 3

Literature Survey

The recent advancements in technologies boost many real-world applications to serve various social benefits to users. These technologies collect and analyze users' sensitive information to gain rich knowledge, helps to play various activities in many applications. However, by collecting and analyzing users' sensitive information, there may be a chance to compromise users' privacy. So, Privacy is an essential phenomenon in various applications that involve data publishing [21][22], data mining [23] [24] [25], etc. For instance, since privacy is not incorporated in the medical dataset the state massachusetts, the medical information of the governor of massachusetts was disclosed by linking his medical data record with the voter registration list [26]. Various linkage attacks are possible in published datasets in order to disclose users' senstive information. To provide the privacy to users against various linkage attacks, it is necessary to build privacy preservation methods for protecting users sensitive information.

3.1 Privacy Preserving Models

3.1.1 Data Anonymization Approach

Data anonymiztion is one of the privacy preserving method, is a process of modifying the relation in such a way that minimal user's private data may be inferred. To protect the users private data from adversary, it is necessary to remove users unique identifiers from the published dataset. Even though removed user identifiers, an efficient adversary can infer a user's record (or targeted victim) with high probability via various linkage attacks. To adress the linkage attack, Sweeney proposed a k-anonymity protection privacy model [26]. This k-anonymity made the released dataset private, where at least the k number of records are unique in each QID group in the released table. Later, various anonymization methods have been proposed to protect users' sensitive information from the other types of linkage attacks, namely, attribute linkage attack, table linkage attack, similarity attack, etc. Further, the privacy preserving in location data release settings or trajectory dataset publishing is an emerging research topic in the recent decade. In literature, various data anonymization methods have been proposed for protecting users' sensitive information against various linkage attacks in setting location data release settings or trajectory data publishing. These privacy methods protect users' sensitive information from either one or a combination of linkage attacks from the adversary with bounded background knowledge. However, the data anonymization methods are broadly classified into two categories: Clustering based anonymization methods and Quasi identifiers based anonymization methods. These two categories are protected users' sensitive information from the published dataset by using different methodologies.

3.1.1.1 Clustering based anonymization methods

The following methods protect users' sensitive information from various linkage attacks by using clustering methods. *Abdul et al.* [27] proposed a (k,δ)-anonymity model to preserve users location information. The trajectory of the moving object is presented in a cylindrical volume instead of three-dimensional space. As they know that the moving object is presented in the cylindrical volume, but they don't know where it is placed. The other moving objects are indistinguishable from each other in cylindrical volume. This proportion leads to the definition of (k,δ) -anonymity. It takes advantage of location uncertainty to limit the amount of distortion required to anonymize trajectory data. Although they present a NWA method for achieving (k,δ)-anonymity. The NWA method starts with partitioning the trajectory dataset into equivalent classes, then generates a set of clusters, then transforms it into a space translation cluster to minimize translation distortions. *Monreale et al.*[28] felt that a new privacy concern is required for mobility data published. The de-identification of high precision and resolution trajectories is the only weak protection and also it is possible to re-identify user mobility by observing routine mobility information. They present a method of combining the notion of location generalization and K-anonymity. It ensures that all generalized trajectories satisfy K-anonymity. Although, this method ensures that data is protected and fixes the upper bound to the probability of re-identification. *Nergiz et al.* [29] adrress the privacy issues while publishing the trajectory dataset. They proposed a randomization-based reconstruction algorithm by adopting the standard privacy notion k-anonymity to anonymize individuals' trajectories in published datasets. Further, it shows that how the underlying techniques can be adapted to other anonymity standards. *Domingo-Ferrer et al.* [30] proposed Swap location and Reach location data anonymization methods. Both methods preserve location user privacy. These methods anonymize the trajectory dataset, which has no perturb or generalize the loca-

tions and also it guarantees the anonymized dataset satisfies k-anonmity. Mahdavifar et al. [8] claims that many privacy preservation approaches provide the same privacy level to all participated users. They present a greedy clustering method in which it anonymizes the trajectory dataset based on users' location data-point privacy requirements. Hence this method provides users personalized privacy in published trajectory datasets. This method allows to assign a privacy level to user trajectory based on their privacy requirement, then partition the trajectories into fixed radius cluster.

3.1.1.2 Quasi identifiers based anonymization methods

The following methods protect users' sensitive information from various linkage attacks by using Quasi identifiers based anonymization methods. *Fung et al.* [5] analyze the privacy leakage threats caused by RFID data publishing. They claim that even though removing quasi identifier values from the published dataset, the attacker can identify users' sensitive information by using his prior knowledge of users' visited location information. If they applied the traditional k-anonymity method to RFID data, it suffers the curse of high dimensionality, leading to affect the data truthfulness. Therefore, they define a privacy model called LKC privacy. It ensures that every sub-trajectory with at most L length is shared at least K number of data-records in the trajectory dataset and confidence of any users' sensitive value is not greater than threshold C. *Mohammed et al.*[31] applied the traditional QID-based anonymization method and its variants to the RFID dataset, conclude that these methods are not suitable because of the curse of high dimensionality, which leads to inadequate data usefulness. To over come this, they propose an efficient anonymization algorithm to address the particular challenges of anonymizing high dimensional, sparse and sequential RFID datasets. *Terrovitis and Mamoulis* [32] studies how to transfer a trajectory dataset into a format that controls adversaries infer a user's missing

location information with high certainty. They propose a privacy preservation method using the suppression technique. It iteratively suppresses the location information in order to achieve the privacy requirement for secure data publishing. *chen et al.* [6] study the privacy issue in trajectory dataset publishing and analyze the effectiveness in data mining activities. They claim that the traditional privacy models are ineffective while dealing with data mining activities. To address this challenge, they introduce the local suppression method to make the secure trajectory dataset publishing. This method adopts the existing $(K, C)_L$-privacy model for trajectory data anonymization and also allows various utility metrics for different data mining tasks. This method supports both local and global suppression, intending to increase the data utility of different data mining tasks. *Hussaeni et al.*[33] propose a new algorithm incremental trajectory stream anonymizer, which anonymizes users' trajectory stream using a sliding window concept. This sliding window updated continuously while joining and leaving user trajectory. An efficient data structure is used for updating the sliding window when massive data is collected. *Terrovitis et al.*[34] propose four anonymization methods that prevent linkage attacks from adversaries whose having partial knowledge of users' trajectory information. These anonymization methods adopt the l-diversity privacy notion to protect users' location information. They proposed four anonymization methods namely suppression of locations, splitting of trajectories or both, to handle large trajectory datasets. These methods employ both suppression and splitting of users trajectories to anonymize the data in order to reduce the location information loss. *Liu et al.*[35] proposed a new privacy protection framework named SLAT-a Sub-trajectory Linkage Attack Tolerance. Also introduced a (α, K, L) privacy model, which adopts both generalization and suppression techniques to preserve users' identity and sensitive location information in published trajectory dataset. This privacy model is an extended version of the LKC privacy model, which preserves privacy in the joint release of trajectory and users-sensitive attributes. This model not only preserve user identity and

sensitive value linkage attacks, but also sensitive location linkage attack. *Ghasemzadeh et al.*[36] propose a hybrid approach that aims to preserve both spatio-temporal data privacy and information quality of passenger flows. This approach utilizes both local and global suppression to achieve a fair tradeoff between privacy and information quality and also comparing two probabilistic flow graphs to assess the information quality before and after data anonymization. *Nergiz et al.* [29] propose a randomization based reconstruction algorithm for publishing anonymized trajectory data and also presented how the underlying techniques can be adapted to other anonymity standards. *Elahe Ghasemi Komishani et al.*[15] present a novel approach PPTD(preserving privacy in trajectory data publishing) for preserving personalized privacy in trajectory data publishing. It involves two major steps: sensitive attribute generalization and trajectory local suppression for generalize the sensitive values and suppress the trajectories respectively. *Lin Yao et al.*[37] propose a (l, α, β) privacy model to resist three types of linkage attacks. It is enhanced version of l-diversity mechanism.

3.1.1.3 Location Privacy

Liu et al. [38] proposed a distributed dummy user generation system based on game-theoretic approach and is the first mechanism using game theoretic approach to achieve k-anonymity in trajectory data publishing. It ensures that users control their privacy themselves, called personalized privacy. They formulate two bayesian game models to analyze the non-cooperative behaviors of users and propose a selection algorithm to gain optimized payoffs. *Cicek et al.* [39] address the released data consists of nodes in the map rather than users trajectories. Therefore, they proposed a ρ-confidentiality technique that addresses the diversification problem in k-anonymity. It ensures that bound to the probability of a user who visits a sensitive location with ρ input parameter to achieve location

diversity.

Table 3.1: List of privacy preservation approaches which prevents different types of linkage attacks.

Approaches	Identity linkage attack	Attribute linkage attack	Similarity linkage attack	Correlated-records linkage attack
Monreale[28]	✓	–	–	–
Ghasemzadeh[36]	✓	–	–	–
Nergiz[29]	✓	–	–	–
Fung[5]	✓	✓	–	–
Mohammed[31]	✓	✓	–	–
Terrovitis[32]	✓	✓	–	–
Chen[6]	✓	✓	–	–
Hussaeni[33]	✓	✓	–	–
Terrovitis[34]	✓	✓	–	–
Liu[35]	✓	✓	–	–
komishani[15]	✓	✓	✓	–
Yao[37]	✓	✓	✓	–
Proposed approach	✓	✓	✓	✓

The above three categorize anonymization methods provide a privacy guarantee either from single or combination three linkage attacks namely identity, attribute and similarity. The Table 3.1 describes a list of privacy preservation approaches that prevent different types of linkage attacks. To the best of our knowledge, the correlated-records linkage attack has not been studied in the previous privacy preservation approaches, and also there is no privacy preservation approach to address all the above four linkage attacks.

3.1.2 ϵ-Differential privacy

As we aware that a huge development in wireless sensor systems and crowdsourced information systems, a large amount of information is collected and analyzed to build rich useful information about a single user or group of users, which can provide various social benefits to everyone. The above data anonymization methods have limited applicability

in real-world applications (ex: crowdsourced systems etc.) because more information loss occurs when dealing with the crowdsourced system. Although, there may be a chance that more new types of linkage attacks are possible in the future. To overcome this, recently, ϵ-Differential privacy (DP) [40] has been introduced, in which it protects users' sensitive information from an adversary with unbounded background knowledge. This mechanism provides a strict privacy guarantee only by removing or modifying a single data record from the dataset. It is defined as a property of a query answering mechanism. It ensures that the resulted answers are not affected by adding or removing of any data-record in the trajectory dataset. Recently, few methods adopt this idea of DP on trajectory dataset [41][42]. The goal of DP method is to publish noisy aggregate information that are effective for some specific data mining task[43] such as count query answering and frequent itemset mining.

3.1.2.1 Trajectory or Location Privacy without correlation

The novel ϵ-DP privacy mechanism can also been applied in settings of continuous data publishing. There are few mechanism which adopts DP privacy notion for continuous data publishing. *Dwork et al.* [44, 19] first initiated and proposed two privacy approaches such as event-level and user-level privacy for studying differential privacy under continual observation or trajectory data publishing. There are privacy preserving approaches under the protection of differential privacy that address only trajectory data publishing. *Li et al.* [45] claim that the related works on trajectory data publishing cannot achieve strict differential privacy due to adding unbounded noise, which affects to leak more privacy leakage and the utility of information also less. Moreover, existing merging methods for trajectory data publishing remove some users' trajectory data from the input dataset. To ovecome these limitations, they proposed a bounded noise generation algorithm under the

protection differential privacy and also presented a trajectory merging algorithm. *Quan et al.* [46] reformulate the standard differential privacy definition for the data, which belongs to an arbitrary domain of secrets. Based on this reformulated definition, they can adjust the noise based on privacy parameters. They present a trajectory or location obfuscation mechanism in which it adds the noise using a polar laplacian method. If the settings require a strong privacy requirement, then the privacy parameter is set to small (eps=0.1), whereas the noise would be adding more if the setting requires less privacy requirement. In the worst case, release the dataset without noise if and only if the setting does not require any privacy requirement. *Cao et al.* [47] claim that if the user's data points are infinite, then the traditional e-differential privacy is not protected every data-points of the user's trajectory under the protection of ϵ-DP. Further, every user has not required the same privacy level. So, they introduced an l-trajectory privacy model for preference length of user trajectory under the protection of ϵ-differential privacy. Also, they presented an algorithmic framework to publish l length user trajectory.

3.1.2.2 Trajectory or Location Privacy with correlation

In literature, a few privacy approaches exist under the protection of ϵ-DP that address the correlation issue in trajectory data publishing. There are two types of correlation in trajectory datasets: the correlation between the user records and the correlation between the data points of the same user. *Kifer et al.* [48] first address the privacy issue in correlated kinds of datasets. To overcome this kind of privacy issue, they proposed a *pufferfish* framework, which requires three components that need to be explicitly specified: potential secrets, discriminative pairs, and data generation. However, there are some challenges to applying *pufferfish* framework to address the correlated issue in trajectory data publishing, which is the lack of a suitable privacy mechanism. *Song et al.* [49] claim that there is a challenge by

using the Pufferfish mechanism, which is lack of suitable mechanism. So, they proposed a novel privacy mechanism, named as Wasserstein Mechanism. This mechanism is adopted in any general pufferfish instantiation. Since the privacy mechanism is insufficient especially when the dataset involves correlation, they proposed a Markov Quilt mechanism. This mechanism exploits the properties of the Bayesian network in order to reduce the computational complexity. *Yang et al.* [50] claim that the correlation involved in the dataset and adversary's background knowledge affect privacy. To overcome this privacy issue, they proposed a new definition for the Pufferfish mechanism, Bayesian differential privacy. It provides a strict privacy guarantee, especially when the dataset involves in the published dataset. Also, they proposed a Gaussian correlation model for complex data correlation. *Zhu et al.* [51] present an effective correlated differential privacy mechanism by defining the correlated sensitivity. This sensitivity significantly decreases the noise compared with traditional global sensitivity. *Liu et al.* [52] proposed an extended version of the differential privacy for correlated data, called dependent differential privacy (DDP). It uses a dependence coefficient to find accurate query sensitivity for dependent data, leading to better data-utility at the same privacy level. This mechanism ensures that no sensitive information leaks even though the dataset involves data correlation. *Wu et al.* [53] address a few privacy challenges in dataset publishing. The challenges are how to represent the correlated relationship between the two different datasets in terms of privacy, how to measure the utility of published dataset that involves correlation, and how to evaluate the data owners' value of privacy. To address these challenges, they proposed a game-based definition of correlated differential privacy to evaluate the privacy level of a single user's record influenced by the other user. And analyze the above game model to evaluate the efficiency or utility of adopted pure Nash equilibrium. *Chen et al.* [54] identifies two vulnerabilities in the setting of sequential data release: to balance the trade-off between the information of the underlying dataset preserve and the extent of noise added. To ad-

dress these, they proposed the n-gram model under the protection of differential privacy for sequential data publishing. Also, they introduced a set of novel Markov techniques that include budget allocation adaptively, the best choice of the threshold value, and the other consistency constraints. *He et al.* [55] claim that raw users trajectories contain highly detailed information about user sensitive information, and it is not easy to preserve privacy in order to maintain the original behavior of users. Based on this challenge, they present a trajectory synthesis method named DPT system to synthesize the users' trajectories based on original users' trajectories; simultaneously, it provides strict privacy guarantees by achieving a differential privacy mechanism. The DPT system involves a set of algorithms for capturing the user movements at various speeds by using a hierarchical reference system, using an adoptive method to choose a small set of a reference system, and build prefix tree counts privately and finally, to improve the utility of the system they use direction-weighted sampling. *Wang et al.* [56] introduced a Private Trajectories Calibration and Publication System (PTCP) to publish large-scale user trajectories. This system provides a strict privacy guarantee with high utility. It builds a noisy enhanced prefix tree to calibrate the noise for large-scale user trajectories. It adopts a post-processing sampling method to increase the scale of data utility in the published dataset. This system supports privacy budget distribution adaptively so that it saves privacy budget, leads to control the noise. The above privacy methods deal with the correlation between the user's records in the dataset, i.e., user-user correlation, whereas, in our settings, we consider the correlation among single user's data at different timestamps, i.e., temporal-correlation.

On the other hand, to the best of our knowledge, there is very limited research on the analysis of privacy risk of differential privacy under temporal correlation in continuous location data release settings. *Wang et al.* [57, 58] present a RescueDP method under the protection of differential privacy for publishing real-time Spatio-temporal crowd-sourced data. This

method groups the regions based on similarity of data change and add the calibrated noise to each group instead of each regions so that it avoids the error for small statistics. This method allows the privacy budget allocation by using an adaptive sampling approach in order to allocate a sufficient amount of privacy budget to each group. Since the presence of temporal correlation in Spatio-temporal data, they use Kalman Filter to improve the accuracy of released crowd-sourced data. *Xiao et al.* [59] addess the temporal correlation privacy issue while sharing users' location information to location-based application host. They present a δ-location set method with differential privacy guarantee to protect users' true location information from adversaries. The main idea of this solution is to hide users' true location information in the δ-location set so that the adversary cannot distinguish the location of a particular user. Also, they present a planar isotropic mechanism (PIM) for perturbing the users' location efficiently. *Cao et al.* [18] define the adversary's background knowledge about the temporal correlation in the dataset using the Markov model. The temporal correlation in user trajectory can be defined as either backward and forward correlations. Next, they analyze the privacy leakage when the dataset involves temporal correlation using the optimal solution, i.e., linear-fractional programming problem. The analysis report that the temporal privacy leakage increases over time in continuous data publishing. Therefore, they proposed α-differential privacy, especially for temporal correlation in user steam or trajectory against the adversary. However, in continuous data publishing settings, the w-event privacy mechanism [17] provides less privacy guarantee than traditional ϵ-DP, especially when the user's data-points are not independent (i.e., temporally correlated) between consecutive timestamps. It happens due to the allotted privacy budget at timestamps within a window of size w is not adequate, especially where the data-points of users' stream involve temporal correlation. Therefore, the privacy budget distribution strategies in w-event privacy such as Budget Distribution (BD) and Budget Absorption (BA) are not suitable in the presence of correlated datasets within a window.

Therefore, it is necessary to design a privacy budget distribution method for allocating a sufficient privacy budget to all timestamps within the sliding window of size w.

3.1.3 ϵ-Local differential privacy

Moreover, ϵ-Differential Privacy provides a strict privacy guarantee against an adversary with unbounded knowledge. However, DP mechanism assumes that the system servers or curators are trusted, but it is unfeasible to believe that all system servers or curators are trusted, which leads to disclose users' sensitive information before aggregation. To this end, ϵ-Local Differential Privacy has been introduced to provide a strict privacy guarantee even though the system servers or curators are not trusted. This mechanism allows users to send their data private before sending it to an untrusted server or curator.

3.1.3.1 ϵ-LDP without correlation

The differential privacy in local settings was first proposed to investigate learning algorithms in local settings [60]. Later, the local differential privacy has become the most promising privacy technique and has been applied in various applications such as Google's Chrome browser [61], Apple's IOS [62, 63], and collecting telemetry data in Microsoft [64]. A randomized response $(W - RR)$ method is used to achieve LDP, which is first proposed by warner in 1965, and it applies only to binary attributes [65]. Due to this limitation, *Kariouz et al.* [11] studies the trade-off between the local differential privacy and information utility function. This study's intuition is to maximize the utility of released statistics from the released dataset. To achieve this, they present a set of extremal privatization mechanisms named staircase mechanisms. These mechanisms show that it maximizes the information-theoretic utility function solved by linear programming. Also, they introduce a K-randomized response mechanism for multiple attributes. Later, *Wang*

et al. [66] propose a generalized randomized response method; it is a generalization of $W - RR$ and $K - RR$. *Wang et al.* [67] propose a framework for the frequency estimation problem, and it enables us to analyze, compare, generalize, and choose an optimal mechanism based on LDP based application's requirement. In addition, there exist a few LDP mechanisms to study on set-valued data [66, 12] and high dimensional data [68, 69].

3.1.3.2 ϵ-LDP with correlation

There is very limited research on local differential privacy in continuous location data publishing. *Bittau et al.* [70] claim that the data collection, process, and privacy concerns are very important in software engineering practice. Nevertheless, unfortunately, these concerns are not addressed and are required to be addressed to give better data utility and simultaneously provide a strong privacy guarantee. So, they present a system architecture that includes Encode, Shuffle, and Analyze (ESA) architecture and its PROCHLO implementation for large-scale monitoring of users' data, and it provides high utility while preserving user privacy. In addition, PROCHLO introduce a new type of cryptographic primitives and an algorithm to balance privacy and utility. *Joseph et al.* [71] claim that a single-purpose algorithm does not provide a strict privacy guarantee over large-scale deployments because these deployments periodically recollect the users' data and recompute the statistics by using an algorithm that is made for single use. Therefore, they present a new LDP-adopted technique for maintaining up-to-date statistics over time. This method helps track and identify small changes in underlying distribution, and it analyzes theoretically but not practical verification. *Erlingsson et al.* [72] finds the privacy gap between central differential privacy and local differential privacy in the setting of continuous data release setting. To overcome this challenge, they proposed a combination of local differential privacy and anonymity concept to protect the users' data points from the published

datasets. Also, they prove that the random shuffling of data points ensures that any local differential privacy protocol satisfies central differential privacy. In addition, they propose a real-time monitoring protocol that protects the longitudinal privacy of users over timestamps, irrespective of whether it involves correlated or independent. *Ding et al.* [64] address that the traditional LDP mechanism provides a strong privacy guarantee while collecting a single round of telemetry data. However, collect the telemetry data periodically then degrades the privacy guarantee. Hence, they claim that the naive memoization method [61] fails to provide a differential privacy guarantee when a user's data-point is constant over timestamps due to the presence of temporal correlation in a user stream. To overcome this, they present a rounding-based discretization and memoization method to solve the problem. Also, they analyze two basic tasks, mean estimation and histogram estimation, and proved that these two tasks provide the same level of privacy as in the traditional LDP data collection mechanism. *Chen et al.* [73] identified few challenges to address the issues in local differential privacy. The challenges are obtaining each individual's privacy requirement in local settings, building a suitable mechanism that computes statistics efficiently, and designing a unified approach for an untrusted server by considering each individual's privacy requirement and learning accurate statistics. To address these challenges, they present a unified private spatial data aggregation framework to achieve personalized local differential privacy and maximize data utility by taking full advantage of users' personalized privacy requirements. This framework adopts a local randomizer based on a randomized matrix (abbreviated as LRM) to protect a user location data-point at each timestamp in local settings. *Xiong et al.* [74] studies the real-time problem while using local differential privacy in continuous data release settings. To address this problem, they presented a new privacy notion (ϵ, δ)-LDP, provides the privacy against individual personalized privacy requirement and temporal correlation in the real-time dataset. Also, they build a privacy mechanism that adopts the generalized randomized response

(GRR) concept, and it satisfies (ϵ, δ)-LDP. The above privacy-preserving methods under spatio-temporal correlation can provide comparatively less privacy than the privacy gain in traditional ϵ-LDP. Also, there are few approaches of ϵ-LDP to address privacy budget allocation under temporal correlation in continuous data release settings. Therefore, it is necessary to design a privacy budget distribution method for allocating a sufficient privacy budget to all timestamps under the protection of ϵ-LDP.

3.2 Budget Allocation Methods

There are a few baseline methods for allocating privacy budgets to each timestamp in literature, such as Uniform [44], Sampling[75, 20], and Budget Absorption (BA) [17]. Firstly, the Uniform method is a basic idea; it divides the privacy budget uniformly and allocates to each timestamp. In other words, each timestamp acquires (ϵ / N) privacy budget if the length of a user stream is N. This method achieves ϵ-LDP if combine all privacy budgets of N length user stream. Second, a Sampling method allocates a privacy budget at given sample interval data-points (say I) of the user stream. That is, the given sample interval data-point acquires ($\epsilon*$ I / N) privacy budgets while assuming an N length user stream. Finally, Budget Absorption (BA) is one of the allocation methods in w-event privacy. It allocates the privacy budget at any timestamp when the privacy mechanism decides to publish outputs, or otherwise, the respective timestamp's budget has become free for future publication. The available privacy budget from the previous skipped publications is used in the next immediate successive publication. The above three baseline methods violate ϵ-LDP in continuous location data publication settings if the correlation exists between the location data-points at the different timestamp. To the best of our knowledge, the privacy budget allocation scheme has not been studied in existing DP and LDP adopted approaches in continuous location data release setting. In this thesis, we proposed a privacy budget al-

location scheme under the protection of ϵ-DP and ϵ-LDP for temporally correlated location data-points release setting and achieves better utility while preserving user's privacy.

This chapter summarizes that the above privacy models such as data anonymization, ϵ-differential privacy, and ϵ-local differential privacy have been applied in continuous data publishing settings. However, these privacy models provide less privacy guarantee, especially when the user's records are not independent (i.e., correlated) or the user's data points are not independent between consecutive timestamps (i.e., temporally correlated). It happens due to an adversary has additional knowledge about the correlated records in the data anonymization approach, or the allotted privacy budget at each timestamp in a user stream is not adequate in ϵ-DP or in ϵ-LDP, especially where the data-points of a users' stream involve temporal correlation.

Chapter 4

Quantify the Impact of Prior Knowledge on Privacy Leakage in Trajectory Data Publishing

In the era of digitization it is very important to preserve users privacy from the efficient adversary. Many privacy-preserving methods have been proposed to protect users' privacy and these methods follow a basic privacy protection principle like the removal of user identity from the trajectory dataset before bringing (or publishing) into the public domain to achieve users' privacy in the published dataset. Moreover, an efficient adversary can infer a user's record (or targeted victim) with high probability by using his/her prior (or background) knowledge [14]. The prior knowledge about the targeted victim may gain from various resources [5], or in most of the scenarios, it is readily available in public [76]. Further, an adversary can perform various linkage attacks on a published (or private) dataset to infer users' sensitive information with high probability [77, 78, 79]. The following example illustrates how users' privacy can breach via performing various linkage

attacks.

Consider a hospital X that maintains a database contain patients' details in the form of ID, patient's trajectory and their medical data as shown in Table 4.1. A patient's trajectory is a sequence of locations visited by a patient with respect to time and is represented as a pair of $(loc, time)$. For instance, a patient ID 6 visited locations are k,n and m at timestamps 6,7 and 8 respectively and the corresponding sensitive value is *High blood sugar*. The hospital wants to release (or publish) a dataset to the data miners for research purposes. The patients may expect that the malicious data miners (or adversary) can misuse the published information in relates to disclosing patient's sensitive information by performing the four linkage attacks such as Identity, Attribute, Similarity and Correlated-records linkage attacks.

Table 4.1: A Hospital trajectory dataset

Id	Trajectory	Sensitive attribute
1	$a1 \rightarrow d2 \rightarrow p3 \rightarrow m4 \rightarrow k6 \rightarrow m8$	Dengue
2	$d2 \rightarrow n5 \rightarrow k6 \rightarrow n7$	High blood sugar
1	$a1 \rightarrow p3 \rightarrow n7 \rightarrow m8$	Virus infection
3	$p3 \rightarrow m4 \rightarrow k6 \rightarrow m8$	High blood sugar
4	$a1 \rightarrow d2 \rightarrow n5 \rightarrow k6 \rightarrow m9$	Dengue
5	$n5 \rightarrow k6 \rightarrow m9$	Lung infection
6	$k6 \rightarrow n7 \rightarrow m8$	High blood sugar
7	$a1 \rightarrow d2 \rightarrow k6 \rightarrow n7 \rightarrow m9$	Cholera
8	$d2 \rightarrow n5 \rightarrow n7 \rightarrow m9$	Typhoid

Moreover, the privacy models such as K-anonymity [80], l-diversity [81] and confidence bounding [81] are not suitable for anonymous the trajectory dataset due to these models are failing to address the following challenges such as high dimensionality [82], sparseness [5] and sequentiality [6]. However, few methods under data anonymization focus either one or more above challenges and present solutions for protecting users privacy against linkage

attacks. The data anonymization methods are broadly classified into two categories: Clustering based anonymization methods and Quasi identifiers based anonymization methods. In literature, [28, 36, 29] present privacy preserving methods based on Quasi identifiers and prevent only one linage attack (identity linkage attack) from the adversary with limited background knowledge. [34, 35, 33] proposed methods based on clustering technique and [5, 31, 33] based on Quasi identifiers and these two categories based approaches prevents only two linkage attacks (identity and attribute linkage attacks) from the efficient adversary. Finally, the works of [15, 37] proposed privacy solutions against three linkage attacks (identity, attribute and similarity). The above approaches provide users privacy against either single or combination of linkage attacks such as identity linkage attack, an attribute linkage attack and similarity linkage attack. On the other side, Differential Privacy (DP) is a new privacy notion and is defined as a property of a query answering mechanism [40]. It ensures that the resulted answers are not affected by adding or removing of any data-record in the trajectory dataset. However, publishing trajectory dataset with DP may not able to provide data truthfulness in published trajectory dataset [14, 15]. This is due to the resulted output (or trajectory dataset) is untruthful because of uncertainty (eg. Laplace noise [13]) introduced for achieving differential privacy. To the best of our knowledge, the correlated-records linkage attack has not been studied in the previous privacy preservation approaches, and also there is no privacy preservation approach to address all the above four linkage attacks. In this work, we used a virtualization and suppression methods to anonymize the trajectory datasets for preserving users' privacy from the above four linkage attacks. The virtualization method works as a replacement mechanism for the sensitive attribute by using privacy height threshold and the suppression method works as anonymization mechanism for users data trajectories by using K-anonymity threshold.

The contributions of this work as follows. We propose a privacy preservation anonymiza-

tion method that preserves users' privacy from four linkage attacks by enabling the virtualization method on the user's sensitive values and suppression method on the user's trajectory. We are introducing a privacy threshold called privacy-height in our method, and it is used to fix the upper-bound to the privacy-risk. Finally, we are comparing the rate of information loss between previous approaches with the propose anonymization approach by conducting an experiment on the synthetic and real-time datasets.

The rest of this work is organized as follows. Section 4.1 introduces the basic notations, privacy attacks, privacy requirements and a utility metric definition. In section 4.2, we propose a privacy preservation anonymization method for providing users privacy from four linkage attacks. The results that experimented with synthetic and real-time datasets are presented in section 4.3. Finally, the summary of this work is presenting in section 4.4.

4.1 System Framework

4.1.1 *Adversary model*

The adversary model for breaching user's privacy is characterized by two important specification 1) Adversary's prior (or background) knowledge, 2) Adversary's goal. We assume that adversary can gain prior knowledge about the target victim from various sources and his goal is to breach the sensitive information of target victim by linking his prior knowledge to the published trajectory dataset. Let A be the adversary, that could be a data analyst or data collector itself and their aim is to find the record or sensitive value of the target victim in T. For instance, Bob is a user and his details are stored in the given Table 4.1. Based on the $A's$ prior knowledge about the Bob's record, A could perform the four types of linkage attacks such as identity, attribute, similarity and correlated-records linkage attack, to breach victim's record or his sensitive value.

Identity linkage attack: Let an adversary A knows Bob's visited locations d and m at timestamps 2 and 4 respectively, then A can claim that a record T_1 is belongs to Bob. This is because, T_1 is the only one record containing sub-trajectory $\{d2, m4\}$ and subsequently A finds Bob's sensitive value is $Dengue$ with 100% confidence.

Attribute linkage attack: Let A knows Bob's visited locations with labeled a and k at timestamps 1 and 6 respectively, then A can identify three records that contains a sub-trajectory $\{a1, k6\}$ such as T_1, T_4 and T_7. The adversary A could predict that bob has $Dengue$ disease with 67% confidence because of two out of the three records having the same sensitive value.

Similarity attack: Let $A's$ knowledge about the Bob's record is $\{n7, m9\}$, then A can find two records T_7 and T_8 and the corresponding sensitive values are $Cholera$ and $Typhoid$ respectively. The adversary A could predict that bob has $Bacterial\ infection$ with 100% confidence, because the sensitive values of T_7 and T_8 are the different types of bacterial infection.

Correlated-records linkage attack: Assume a user (Bob) having multiple records in a given dataset. Let A knows the number of data-records that bob has in the dataset and also knowledge about bob's visited location $\{a1, m8\}$. Then the adversary A could predict that bob has $Dengue$ plus $virus\ infection$ with 100% confidence. Because the number of data-records identified by A is matched with $A's$ prior correlated knowledge.

To prevent the above the linkage attacks, it is required to provide privacy to each user who participated in the published trajectory datasets. To achieve privacy against from the adversary, it is necessitate to adopt few privacy requirements in our proposed method.

4.1.2 *Privacy Requirement*

We used a taxonomy tree of sensitive attribute [83] in our propose method for calculating the level of sensitive values. In general, the taxonomy tree is defined as a set of tuple $\eta = (\nu, \ell)$, where ν is a set of nodes and ℓ is a set of node's level. The taxonomy tree of a sensitive attribute is a hierarchy of various nodes (or sensitive-values) and it establish a relationships among the various nodes of different levels. Each node in the taxonomy tree has its own level, called privacy level and the level number starts from root node to leaf nodes. Let a function $\ell(\nu_i)$ is defined to return a level of sensitive value of user i. A node ν_j is a parent of a node ν_i iff $\ell(\nu_i) > \ell(\nu_j)$ and $\nu_j \in \hbar(\nu_i)$, where $\hbar(\nu_i)$ is a family of ancestors of node ν_i and $\hbar^k(\nu_i)$ is a k^{th} ancestor of node ν_i. For example, $\hbar^1(H1N1) = virus\ infection$ as shown in the Figure 4.1.

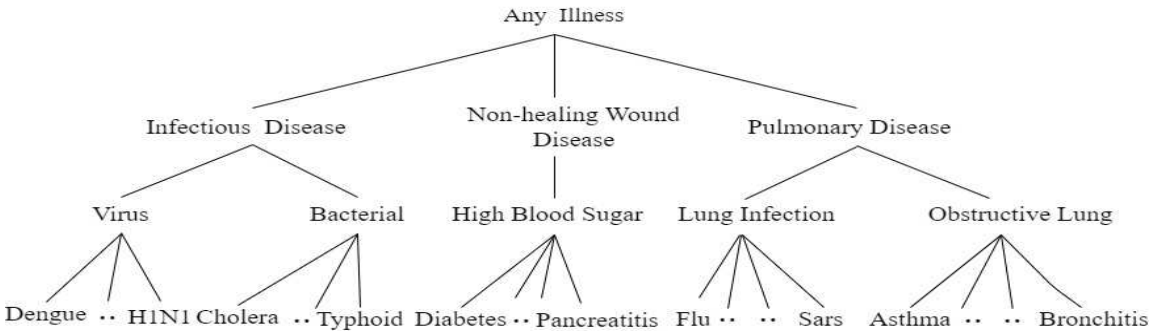

Figure 4.1: A Taxonomy Tree of Human diseases

Next, we introduced a threshold called privacy-height in our method and is denoted as Γ. It is used to fix the upper-bound to the privacy-risk. In other words, the proposed method offers a privacy guarantee to the users in which the level of sensitive-values is greater than the privacy-height threshold. For instance, the record T_4 violates privacy-height threshold when Γ is set to 2, because the level of sensitive value of T_4 ($Dengue$) is 3. Also, we adopt traditional thresholds[6] such as the trajectory of a user is appearing

at least K times and limit the adversary's prior knowledge (represented as ∂), i.e., an adversary knows the maximum ρ number of moving points of any targeted user.

4.1.2.1 *Virtualization*

A method virtualization is used to virtualize(or pseudonym) the sensitive-value of users. If a sensitive-value of any data-record violates given privacy-height threshold Γ (called *critical data-record*), then the virtualization method finds an appropriate virtual-sensitive value from the taxomy tree which satisfies given privacy-height threshold and replace it with original sensitive value. A sensitive value $s^v \in \mathcal{S}$ is said to be a virtual-sensitive value for a critical data record $T_i(s)$ iff $\ell(s^v) \leq \Gamma$ and $s^v \in \hbar(s)$. For instance, assume $\Gamma = 2$, the record T_4 violates Γ threshold (see Table 4.1). Then the virtualization method determines a virtual-sensitive value for T_4 i.e., $Virus\ infection$ and it is replaced with original sensitive value of record T_4 as shown in the Table 4.2. The modified trajectory dataset (Table 4.2) is named as virtual-trajectory dataset and is denoted as $\mathcal{T}^v$.

Table 4.2: Virtual-Trajectory dataset

Id	Trajectory	Virtual sensitive attribute
1	$a1 \rightarrow d2 \rightarrow p3 \rightarrow m4 \rightarrow k6 \rightarrow m8$	Virus infection
2	$d2 \rightarrow n5 \rightarrow k6 \rightarrow n7$	High blood sugar
1	$a1 \rightarrow p3 \rightarrow n7 \rightarrow m8$	Virus infection
3	$p3 \rightarrow m4 \rightarrow k6 \rightarrow m8$	High blood sugar
4	$a1 \rightarrow d2 \rightarrow n5 \rightarrow k6 \rightarrow m9$	Virus infection
5	$n5 \rightarrow k6 \rightarrow m9$	Lung infection
6	$k6 \rightarrow n7 \rightarrow m8$	High blood sugar
7	$a1 \rightarrow d2 \rightarrow k6 \rightarrow n7 \rightarrow m9$	Bacterial infection
8	$d2 \rightarrow n5 \rightarrow n7 \rightarrow m9$	Bacterial infection

4.1.2.2 *Suppression*

Suppression is a method to suppress (or eliminate) violate moving-points from the user's trajectory. A set of moving-points are said to be violate (or critical sub-trajectory) iff $|T_j(\tau_i)|_{T_j \in (\mathcal{T})^{s'}} \leq K-1$ for all $s' = \mathcal{S}-s$, where $|T_j(\tau_i)|$ be the total number of data records that contains a sub-trajectory τ_i. The suppression method can apply in two settings. 1) It eliminates a moving-point only from the corresponding trajectory of the dataset is called local-suppression. 2) It eliminates a moving-point from all trajectories of the dataset is called global suppression. In our anonymization method, we adopt a local-suppression setting to achieve high data utility in the published trajectory dataset. For instance, assume a threshold $K = 2$, a sub-trajectory $\{a1, m4\}$ of record T_1 not satisfy K value. Because $\{a1, m4\}$ is not appeared in any trajectories of $\mathcal{T}^v$ with respect to sensitive value of record T_1 as shown in Table 4.2. To suppress a moving-point of critical sub-trajectory is depends on the suppression metric, which is discussed in the next following sub-section.

4.1.3 *Utility Metrics*

It is important to balance the trade-off between the users' privacy and utility of anonymized trajectory dataset $\mathcal{T}'$. We define a suppression metric to measure the suppression score of each moving-point in the critical sub-trajectory. It helps to decide which moving-point has to eliminate from the critical sub-trajectory, so that the anonymized trajectory datset gains more utility. The suppression score is defined as follows.

Let $(\mathcal{T})^s$ be set of data records having same sensitive value $s \in \mathcal{S}$ and $T_i \in (\mathcal{T})^s$ be a data record of user i. Assume that $\tau_i \in T_i$ is a critical sub-trajectory of user i. The suppression score of a moving point $x^n \in \tau_i$ with respect to $(\mathcal{T})^s$, is denoted by $\Upsilon(x^n, (\mathcal{T})^s)$ and

calculated as follows.

$$\Upsilon(x^n, (\mathcal{T})^s) = \frac{|(\mathcal{T}(x^n))^s|}{|(\mathcal{T})^s|} \tag{4.1}$$

Where $(\mathcal{T}(x^n))^s$ is a set of trajectories of sensitive value s contains a moving point x^n. The suppression metric of critical sub-trajectory τ_i with respect to $(\mathcal{T})^s$, is denoted by $\Upsilon(\tau_i, (\mathcal{T})^s)$ and calculated as follows.

$$\Upsilon(\tau_i, (\mathcal{T})^s) = \underset{x^n \in \tau_i}{max} \Upsilon(x^n, (\mathcal{T})^s) \tag{4.2}$$

Moreover, the local-suppression method always eliminates a moving-point whose suppression score lesser in the critical sub-trajectory. Once the method is to suppress a moving-point of all critical sub-trajectories of the dataset $\mathcal{T}^v$, then replace all original sensitive values in place of virtual sensitive values of the corresponding data-records in $\mathcal{T}^v$, as shown in the Table 4.3. To the best of our knowledge, none of the previous anonymization approaches use virtualization method in order to preserve the privacy of users in the trajectory data publishing scenario. And also notice that there is no sensitive-attribute information loss in our proposed method. Hence, the proposed method provides better data utility and also ensures a better privacy guarantee against four linkage attacks.

4.2 Proposed Method

In this section, we present an anonymization method that prevents users sensitive information from the four linkage attacks. The proposed method consists of two phases, 1) the sensitive attribute virtualization - works as a replacement mechanism for sensitive attribute by using privacy height threshold, 2) the trajectory suppression - works as anonymization

Table 4.3: Anonymized Trajectory dataset $\mathcal{T}'$

Id	Trajectory	Sensitive attribute
1	$p3 \rightarrow m4 \rightarrow k6 \rightarrow m8$	Dengue
2	$d2 \rightarrow n5 \rightarrow k6 \rightarrow n7$	High blood sugar
1	$n7 \rightarrow m8$	Virus infection
3	$p3 \rightarrow m4 \rightarrow k6 \rightarrow m8$	High blood sugar
4	$d2 \rightarrow n5 \rightarrow k6 \rightarrow m9$	Dengue
5	$n5 \rightarrow k6 \rightarrow m9$	Lung infection
6	$k6 \rightarrow n7 \rightarrow m8$	High blood sugar
7	$d2 \rightarrow k6 \rightarrow m9$	Cholera
8	$d2 \rightarrow n5 \rightarrow m9$	Typhoid

mechanism for users' trajectories by using K-anonymity threshold. The detailed procedure for above two phases is follows.

1) Sensitive-attribute virtualization: The sensitive-attribute virtualization method aims to identify all critical data records from the given trajectory dataset and replace their sensitive value with virtual-sensitive value by using a taxonomy tree of the corresponding sensitive-attribute domain. The algorithm 4.1 shows the pseudo-code of Sensitive-attribute Virtualization $SaV()$ method and is follows.

The algorithm SaV() takes a raw trajectory dataset, privacy-height threshold and a taxonomy tree as inputs and yields a virtual trajectory dataset as an output. The $SaV()$ starts with finding records that belong to the same user by comparing each record with successive records of $\mathcal{T}$. If records are found, then verify whether both the records are satisfied privacy-height threshold. If it satisfies, then check whether the sensitive values of both records are equal. If equal, then it is not necessary to replace the virtual-sensitive value. Otherwise, $SaV()$ selects a record with the maximum level's sensitive value among the two records. Set maximum level's sensitive value as a virtual-sensitive value and replaced with other record's sensitive value (Line $3 - 10$). Further, there is a possibility that one

Algorithm 4.1 Pseudocode of SaV()

INPUT: Trajectory dataset $\mathcal{T}$, Taxonomy of sensitive-attribute, Privacy-Height Γ.
OUTPUT: Virtual-Trajectory dataset $\mathcal{T}^v$.

Scan Trajectory Dataset $\mathcal{T}$
for each T_i compare with successive T_{i+1} **do**
    if (T_i, T_{i+1}) are belongs to an individual **then**
        if $(\ell(T_i(s)) \leq \Gamma \wedge \ell(T_{i+1}(s)) \leq \Gamma)$ **then**
            if $(T_i(s) \equiv T_{i+1}(s))$ **then**
                set: the sensitive values as virtual values
            else
                $v^s \leftarrow \max(\ell(T_i(s)), \ell(T_{i+1}(s)))$,
                set: $T_i(s), T_{i+1}(s) \leftarrow v^s$
            end if
        else if one of T_i, T_{i+1} hold Γ threshold **then**
            set: $v^s \leftarrow T_i(s)$, assume that $\ell(T_i(s)) \leq \Gamma$,
            set: $T_{i+1}(s) \leftarrow v^s$
        else
            call EVsV$(T_i(s))$,
            set $T_i(s), T_{i+1}(s) \leftarrow v^s$
        end if
    else
        if $(\ell(T_i(s)) \leq \Gamma \wedge \ell(T_{i+1}(s)) \leq \Gamma)$ **then**
            set: the sensitive values as virtual values
        else if one of T_i, T_{i+1} hold Γ threshold **then**
            call EVsV$(T_i(s))$, assume $\ell(T_{i+1}(s)) \leq \Gamma$,
            set: $T_i(s) \leftarrow v^s$
        else
            call EVsV() for both T_i, T_{i+1}
            set: $T_i(s) \leftarrow v^s$
            set: $T_{i+1}(s) \leftarrow v^s$
        end if
    end if
end for

of the records does not satisfy privacy threshold Γ, then the $SaV()$ set a sensitive value of other records as a virtual-sensitive value and it's replaced with the sensitive value of violated record (Line $11-13$). In the worst case, none of the records satisfy privacy threshold Γ, then the $SaV()$ call $EVsV()$ algorithm to find virtual-sensitive value for both records and is replaced with a corresponding sensitive value of violated records(line $14-16$). In case both records belong to the different users, then repeat a similar procedure as explained before $(line19-20)$. But in case one of the records does not satisfy the privacy-height threshold, then the $SaV()$ call $EVsV()$ algorithm to find the virtual-sensitive value for violated record and it's replaced with the sensitive value of violated data-record. In the worst case, none of the data-records are satisfied Γ threshold, then the $SaV()$ call $EVsV()$ algorithm to find a virtual-sensitive value and it's replaced with the sensitive value of both records $(line21-28)$.

Algorithm 4.2 Pseudocode of EVsV()

INPUT: Sensitive value of record $T_i(s)$, Taxonomy of sensitive attribute, Privacy-Height Γ.
OUTPUT: Virtual-sensitive value v^s.

```
1:  Initialize v^s ← ∅, j ← 1
2:  do
3:      set v^s ← ħ^j(T_i(s))
4:      if (ℓ(v^s) ≡ Γ) then
5:          t ← false
6:      else
7:          J++
8:          T_i(s) ← v^s
9:          t ← true
10:     end if
11: while (t)
12: Return v^s
```

The algorithm 4.2 $EVsV()$ takes a sensitive value of the violated record, taxonomy tree of the sensitive-attribute domain and a privacy-height threshold are as input and it gives a virtual sensitive-value as an output. The algorithm $EVsV()$ starts with identifying

a parent of given sensitive value by using a taxonomy tree (line 3) and verify whether the identified parent (or sensitive) value satisfies a privacy-height threshold (line 4). If yes, then return a parent value as a virtual-sensitive value. Otherwise, followed the same procedure for updated value or parent value.

Now, the given trajectory dataset is transformed into the virtual-trajectory dataset $\mathcal{T}^v$ by using $SaV()$ and $EVsV()$ algorithms. Further, several critical moving points are present in the trajectories of the database $\mathcal{T}^v$, which leads to breach sensitive information of users. To overcome this problem, the suppression method is applied to the dataset $\mathcal{T}^v$ in order to protect sensitive information of users.

2) Trajectory suppression The trajectory suppression method is used to remove all critical moving points from the trajectories of the dataset $\mathcal{T}^v$ and produced anonymized trajectory dataset. The algorithm 4.3 shows the pseudo-code of the trajectory suppression method $ATdb()$ and as follows.

The algorithm $ATdb()$ takes a virtual-trajectory dataset, adversary's prior knowledge ∂ and K threshold are as inputs and returns an anonymized trajectory dataset $\mathcal{T}'$ as an output. It starts with grouping a set of all distinct sensitive values, named as set S (line 1). For each sensitive value s from $\mathcal{S}$ do the following. Split the dataset $\mathcal{T}^v$ into two different datasets, say a set A consists of data-records with sensitive value s and the remaining data-records into the set B (line $5-6$). Then, finds all sub-trajectory of length 1 from the records of the set A and examine whether every sub-trajectory of set A is appearing at least $K-1$ times in the data-records of set B. If yes, keep the sub-trajectory as it is in the set A. Otherwise, eliminate a moving-point of the sub-trajectory from the corresponding data-record of the set A (line $7-15$). Next, it finds all sub-trajectory of length 2 by using union operation. And examine whether any sub-trajectory of length 2 is not satisfied K privacy threshold. If found, then eliminate a moving-point (computed by suppression score metric) from the

Algorithm 4.3 Pseudocode of ATdb()

INPUT: Virtual-Trajectory dataset $\mathcal{T}^v$, $A's$ prior knowledge ∂ with maximum length ρ, K threshold

OUTPUT: Anonymized Trajectory dataset $\mathcal{T}'$.

Scan Trajectory dataset $\mathcal{T}^v$
let S = {set of all distinct sensitive values}
for each $s \in S$ **do**
    $i = 1, \mathtt{C}_r = \emptyset, \mathtt{D}_i = \emptyset$
    $\mathtt{A} = \{T_r \mid T_r \in \mathcal{T}^v \wedge T_r(s) = s\}$
    $\mathtt{B} = \mathcal{T}^v - \{\mathtt{A}\}$
    for each $T_r \in \mathtt{A}$ **do**
        $\mathtt{C}_r = \{\tau_r \mid \tau_r \subseteq T_r \wedge \mid \tau_r \mid = 1\}$
        for each $\tau_r \in \mathtt{C}_r$ **do**
            if $(\mid \tau_r \in T_r \mid_{\forall T_r \in \mathtt{B}} \geq \mathtt{K})$ **then**
                $\mathtt{D}_i = \mathtt{D}_i \cup \tau_r$
            else
                remove τ_r from $T_r \in \mathtt{A}$
            end if
        end for
    end for
    while $(i + 1 \leq \rho)$ **do**
        for each $\tau_r \in \mathtt{D}_i$ join with successive τ_{r+i} in $\mathtt{D}_i$ **do**
            if $(\mid \tau_r \cup \tau_{r+i} \in T_r \mid_{\forall T_r \in \mathtt{B}} \geq \mathtt{k})$ **then**
                $\mathtt{D}_{i+1} = \mathtt{D}_{i+1} \cup \{\tau_r \cup \tau_{r+i}\}$
            else
                $\mathtt{x} = \tau_r \cup \tau_{r+i}$
                remove $\Upsilon(x)$ from $T_r \in \mathtt{A}$
            end if
        end for
        $i = i + 1$
    end while
end for
Replace all original sensitive values

sub-trajectory, and the remaining moving points of the sub-trajectory are kept as it is in the corresponding data-record (line $18 - 25$). Repeat the steps until the length of the sub-trajectory is equal to the adversary's prior knowledge length ∂. Repeat the same procedure for all S values. Then, the original sensitive values of all records are copied into the corresponding virtual sensitive value of records in $\mathcal{T}'$, as shown in Table 4.3. Hence, there is no sensitive-attribute information loss in our proposed method.

4.2.1 Correctness Proof of Algorithms

In this section, we validate our proposed algorithms resist all four linkage attacks, namely identity, attribute, Similarity and correlated-records linkage attacks. The proposed anonymization method produces an anonymized trajectory dataset $\mathcal{T}'$ (Table 4.3) and it ensures that protect users' privacy against four linkage attacks. To the best of our knowledge, none of the previous anonymization approaches use the virtualization method to preserve the privacy of users in the trajectory data publishing. Now we show that the anonymized trajectory dataset $\mathcal{T}'$ is resistant to all four linkage attacks.

Consider Table 4.3 is a anonymized trajectory dataset and it satisfies the privacy requirements such as $\Gamma = 2$, $K = 2$ and $\rho = 2$. Let an adversary A with prior knowledge $\partial = 2$, for example $\{d2, m4\}, \{a1, k6\}, \{n7, m9\}$ and $\{a1, m8\}$. Then, A can perform all four linkage attacks on Table 4.1 that discussed in subsection 2.2. While Table 4.3 is resistant to all four linkage attacks against $A's$ prior knowledge $\partial = 2$. For example, $A's$ prior knowledge about Bob is $\{d2, m4\}$, the adversary is not able to identify any data record which matches the prior knowledge $\{d2, m4\}$ in Table 4.3 and there are four data-record which matches the prior knowledge $\{d2\}$, but the inference of Bob's sensitive value is lesser than or equal to minimum confidence 50%. Further, adversary cannot perform these four linkage attacks with other prior knowledge of length $\partial = 2$. For example, given prior

knowledge is $\{k6, m9\}$, the adversary can infer Bob's sensitive value is $Dengu$ with 33% confidence, that is lesser than or equal to minimum confidence 50%. Hence, Table 4.3 is resistant to all four linkage attacks against $A's$ prior knowledge $\partial = 2$.

4.3 Experimental Results

We conduct an experiment for evaluating the performance of our proposed algorithm in terms of information-loss in anonymized trajectory dataset $\mathcal{T}'$. Generally, the information-loss occurs in any anonymized trajectory dataset due to either eliminate a set of moving points from the users' trajectory or distortion (or generalization) of the sensitive value of users or both. In our method, the sensitive value of all users in anonymized trajectory dataset $(\mathcal{T}')$ is not distorted because the original sensitive values of all records are copied into the corresponding virtual sensitive value of records in $\mathcal{T}'$ (see Table 4.3). Thus, the information loss in our method only from the users' trajectory, not from the sensitive attribute. We considered four trajectory datasets for conducting experiments such as Geolife[84], T-Drive[85], Metro100K[31] and private Wi-Fi datasets. A Geolife dataset is a real-time GPS trajectory dataset that collects from 182 users in a span of three years. The average frequency of data (location value) collection from the users is every 1 to 60 second. The $T-Drive$ dataset contains the routes of 10357 taxis in a span of one year and the average frequency of data (location value) collection from the users is every 3 minutes. A private Wi-Fi dataset is a real-time dataset that involves around 12500 users trajectories obtained from the 175 Wi-Fi points in 24 hours and the frequency of data collection from the users in every 1 to 5 minute. Finally, a Metro100K dataset is a synthetic dataset that contains 100000 users trajectories in a metropolitan area with 26 cities in 24 hours. In all datasets, each trajectory corresponds to the routes of one person and a randomly assigned sensitive value with one of six possible values to each trajectory data record.

4.3.1 Metrics of Evaluation

We used various measures (or metrics) to evaluate the strength of our proposed method in terms of privacy leakage and information loss.

- **Privacy Breach (or leakage):**

  The privacy breach (denoted as σ) of a data-record T_i with respect to $s \in \mathcal{S}$ is denoted as $\beta(T_i)^s$ and is computed as follows.

$$\beta(T_i)^s = (P_b(T_i(s) \mid T_i(\tau)) > \sigma) \tag{4.3}$$

  Where τ is a sub-trajectory of length at most ρ number of moving-points.

- **Trajectory Information Loss:**

  Given an anonymized trajectory dataset $\mathcal{T}'$ and its original trajectory dataset $\mathcal{T}$. Let $\Phi : \mathcal{T}' \longrightarrow \mathcal{T}$ be a function that maps a data record of $\mathcal{T}'$ to its corresponding data record of $\mathcal{T}$. Then, the trajectory information loss of the data record T_i' is.

$$IL(T_i') = \frac{\mid \Phi(T_i') \mid - \mid T_i' \mid}{\mid \Phi(T_i') \mid} \tag{4.4}$$

  Where $|\Phi(T_i')|$ be the total number of moving points in T_i of the dataset $\mathcal{T}$ and the $|T_i'|$ be the total number of moving points in the corresponding data record T_i' of the dataset $\mathcal{T}'$. Then, the total trajectory information loss of the dataset $\mathcal{T}'$ is computed as $IL(\mathcal{T}') = \Sigma_{i=1}^{|\mathcal{T}'|} IL(T_i')$.

- **Query Answering mechanism**

  Let μ be the mechanism that reads a query q and trajectory datasets as inputs and it returns the error rate of anonymized trajectory dataset. The mechanism μ is com-

puted as

$$\mu = \frac{\mid \mu_q^{\mathcal{T}} \mid - \mid \mu_q^{\mathcal{T}'} \mid}{\mid \mu_q^{\mathcal{T}} \mid} \tag{4.5}$$

Where $|\mu_q^{\mathcal{T}}|$ and $|\mu_q^{\mathcal{T}'}|$ are the total number of data-records satisfies the query q in $\mathcal{T}$ and $\mathcal{T}'$ datasets respectively.

Initially, we start by analyzing the number of user's privacy that can breach when publishing a raw trajectory dataset into the public sector. To analyze, we adopt a privacy breach threshold and it helps to fix the upper-bound to the privacy-risk. In other words, if the probability of privacy-breach on the user's sensitive value is greater than σ value, then the user's privacy is at risk from various linkage attacks. The adversary's prior knowledge ∂ and K values increase, then the average users' privacy breach $\beta(T_i)^s$ in $\mathcal{T}$ also increases because more number of users' trajectories not satisfy K threshold. The effect of the privacy breach threshold (fix $\sigma = 0.5$) on users' privacy risk with various ∂ and K values as shown in Figure 4.2. The result shows that the length of ∂ and K value increased while the average users' privacy breach also increases.

It is essential to calculate the number of user's trajectory information loss in the published trajectory dataset $\mathcal{T}'$. Otherwise, the result of data analysis may give incorrect output. The trajectory information loss occurs due to the elimination of critical moving points from the user's trajectory for satisfying the privacy threshold values. Figure 4.3 shows the average trajectory information loss in four trajectory datasets with various K threshold values. The result suggests that, with increasing K values, then the average trajectory information loss in $\mathcal{T}'$ is also increasing because the number of critical moving points is increased due to not satisfying the privacy threshold values.

Further, we applied a query answering mechanism [15] on both the original trajectory

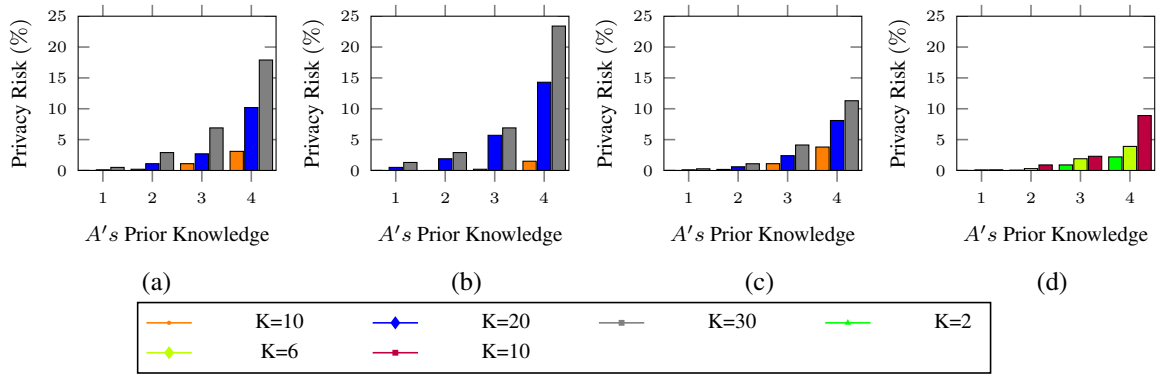

Figure 4.2: The average privacy risk of users with respect to A's prior knowledge of various lengths while fix $\sigma = 0.5$. a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset..

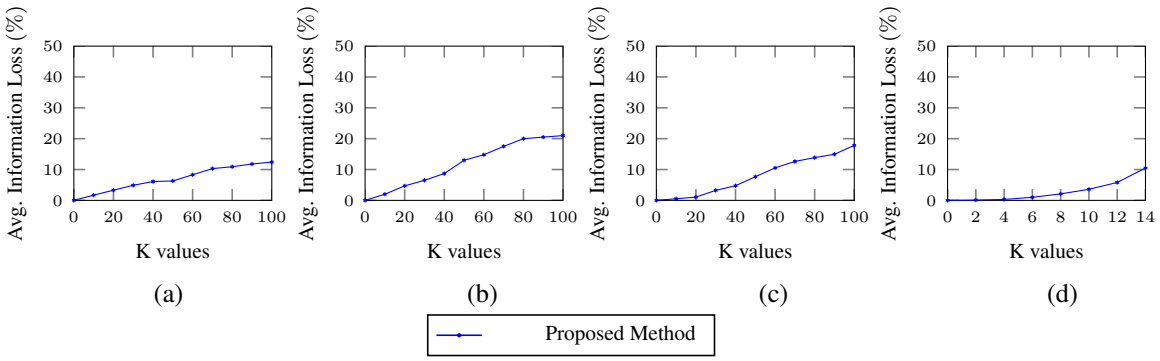

Figure 4.3: The average trajectory information loss in T' with various K threshold values a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset.

dataset and anonymized trajectory dataset to calculate the query-answer error rate (or to test the utility of published trajectory dataset). Let q be the count query (ex: count the number of data-records in $\mathcal{T}$ which contain a sub-trajectory τ'). For experiment, we randomly choose 500 sub-trajectories of different sizes from the four trajectory datasets and calculate the average error-rate of count query. Figure 4.4 shows the result of the average query-answer error rate while fixing various $A's$ prior knowledge ∂. The result shows that the average query-answer error rate increases while increasing $A's$ prior knowledge due

to more number of moving points are eliminated from the critical sub-trajectory as ∂ increases. The effectiveness of privacy thresholds on the utility of the anonymized trajectory dataset is as follows.

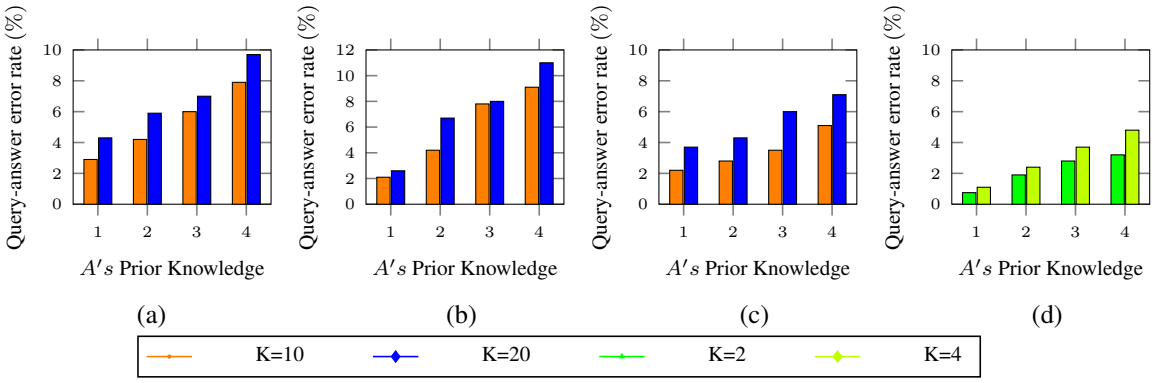

Figure 4.4: The average query-answer error rate with various A's prior knowledge a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset.

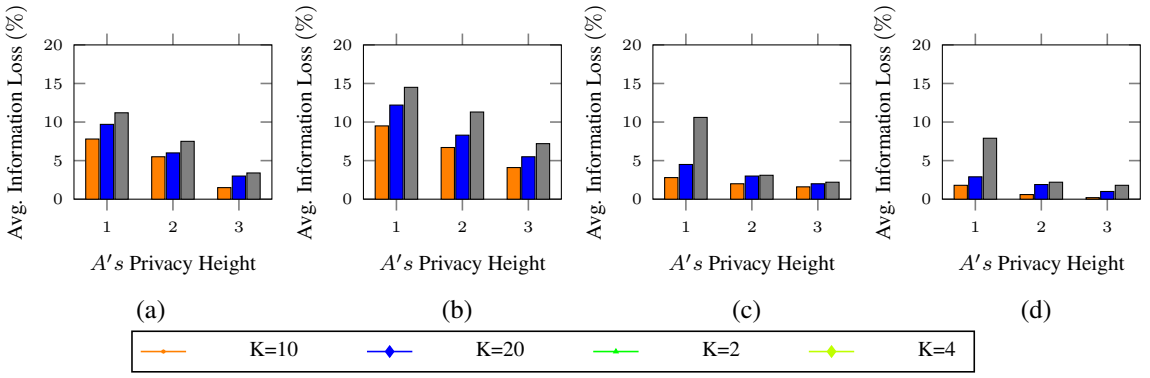

Figure 4.5: Effect of Privacy-Height Γ threshold on users trajectory-information loss for different K-anonymity values a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset.

Effect of Γ *threshold:* Figure 4.5 shows the effect of the privacy-height threshold on users trajectory in $\mathcal{T}'$ of various K values, while fixing $\rho = 2$. The result shows that the average trajectory information loss is diminished when the value of the privacy-height is increased.

Because a small number of moving points are eliminated from the dataset $\mathcal{T}$ due to a small number of sensitive values are generalized. Notice that the root node cannot be the value of the privacy-height threshold. Because of the sensitive value of all users become a unique sensitive value, that leads to an increase in more sensitive-attribute information loss.

Effect of K threshold: The effect of K privacy threshold on anonymized trajectory dataset $\mathcal{T}'$ is shown in Table 4.3. The result shows that if a privacy threshold K value is increased, then the average users' trajectory information-loss is also increased. Therefore, the data publisher has to choose an appropriate K threshold value in which it maintains moderate utility of the dataset $\mathcal{T}'$ as well as users' privacy.

Further, we demonstrate the effectiveness of anonymized trajectory datasets in terms of user's points of interest (PoIs). To determine the users PoIs for the particular area, we need to consider the locations which are frequently visited by users and the span of time interval they stayed in those locations. For experiment, we consider a set of time intervals (eg. ≤ 5, ≤ 10, $\geq 5\& \leq 30$ and $\geq 10\& \leq 60$ minutes) and accordingly we find time specific visited PoIs in original dataset and the anonymized datasets published from the proposed method and other states of the art methods for analyzing average PoIs information loss or number of PoIs distorted. Figure 4.6 shows the average frequency of PoIs visited between the original dataset and the anonymized datasets published from the proposed method and other states of the art methods. The result shows that the proposed anonymized dataset has slightly high frequency of PoIs visited when compared with the other states of the art methods.

4.3.2 Comparison

The KCL-local[6], KCL- Global[5][31] and KCL-PPTD[15] are the recent anonymization approaches which provides users privacy against either single or combination of three

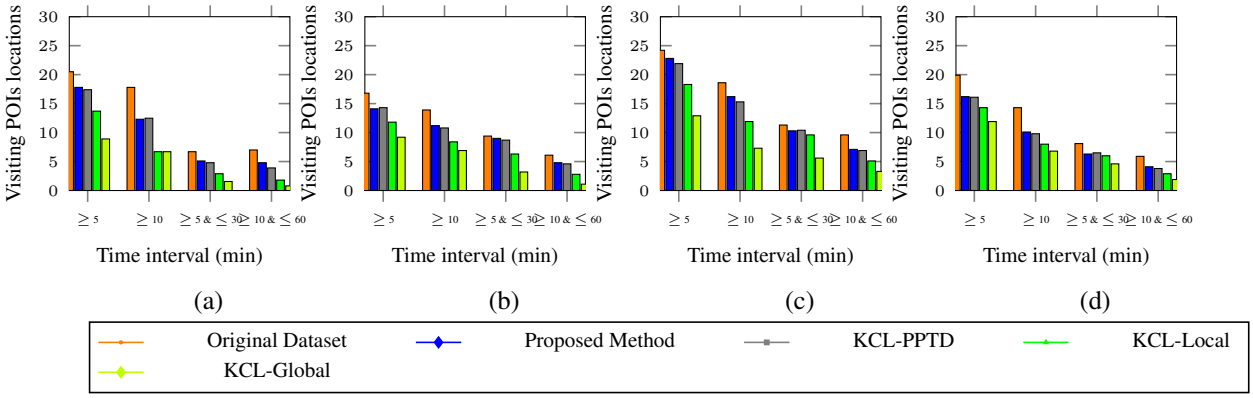

Figure 4.6: The average frequency of PoIs visited between the original dataset and the anonymized datasets published from the proposed method, KCL-PPTD, KCL-local and KCL-Global. a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset.

linkage attacks such as identity, attribute and similarity. We compare our proposed privacy approach with the states of the art methods to validate the efficiency of our method and we use exactly similar dataset (Metro100K), which is used in the above approaches for fair comparison. The experimental results exhibit that the proposed approach results in better performance with a significant reduction in information loss (includes both trajectory and sensitive values), as shown in Figure 4.7. The information loss in the proposed method is closely related to the KCL-PPTD method, because in KCL-PPTD, less trajectory information loss in user's trajectories and more sensitive information loss in the sensitive attribute. While in our method, no sensitive attribute loss and a little bit more trajectory information loss in user's trajectories. Therefore, the information loss of these two methods is comparatively similar. In contrast to KCL-PPTD, the proposed method prevents one extra linkage attack (correlated-records linkage attack) as well. Thus, the proposed method is a comparatively better approach than the existing approaches.

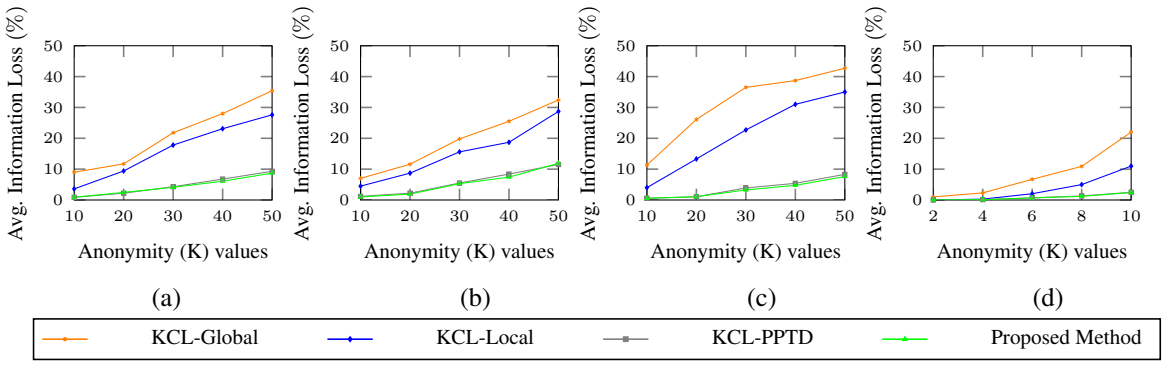

Figure 4.7: The average trajectory information loss in T' with various K threshold values a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset.

4.3.3 Complexity Analysis

The propose anonymization method consists of two phases. In the first phase, the sensitive value of critical data-records is replaced with virtual sensitive-value to generate virtual-trajectory dataset by using two algorithms $SaV()$ and $EVsV()$. The worst-case time complexity of $SaV()$ is $O(|\mathcal{T}|^2)$ and the worst-case time complexity of $EVsV()$ is $O(h)$, where h is a height of the taxonomy tree. Since h is small, it becomes $O(1)$. Therefore, the worst-case time complexity of the first phase is $O(|\mathcal{T}|^2)$. In the second phase, the moving-points of critical data-records are suppressed by using an algorithm $ATdb()$. The worst-case time complexity of $ATdb()$ is $O(\mathcal{S}n^{\rho}|\mathcal{T}|^{l}|\mathcal{T}|^{k})$, where the S be the set of all distinct sensitive values, n^{ρ} be the set of discrete sub-trajectory of length ρ, $|\mathcal{T}|^{l}$ is the total number of data-records of set A, $|A|= l$ and $|\mathcal{T}|^{k}$ is the total number of data-records of set B, $|B|= k$ (refer Algorithm 4.3) i.e., $|\mathcal{T}|^{l}+|\mathcal{T}|^{k}= |\mathcal{T}|$. Hence, the worst-case time complexity of the proposed approach is $O(\mathcal{S}n^{\rho}|\mathcal{T}|^{2})$. Figure 4.8 shows the result of running time performance between the proposed method and previous existing methods. We can observe that as Kvalues increases, the running time also increases because more number

of trajectories satisfies different K-anonymity values.

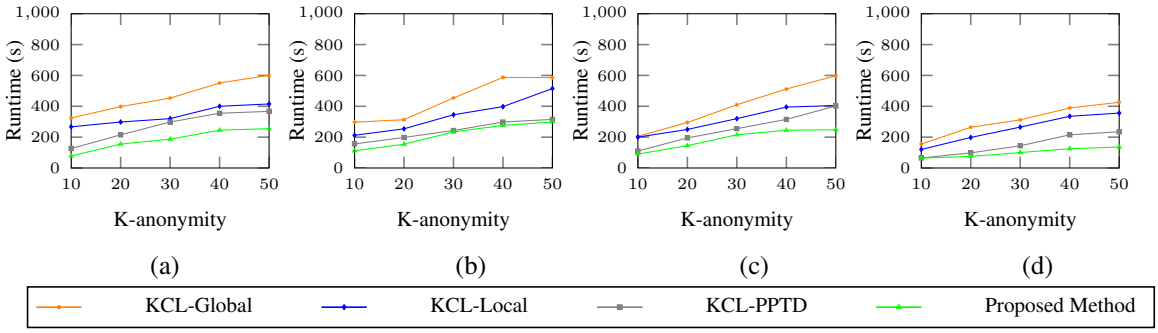

Figure 4.8: The run-time performance between the proposed privacy approach with KCL-PPTD, KCL-local and KCL-Global a) Geolife dataset b) T-Drive dataset c) Metro dataset d) Private Wi-Fi dataset.

4.4 Summary

In this work, we present a privacy preservation approach to prevent users' privacy from four different types of linkage attacks. Our method adopts an existing LK privacy model to fix the upper-bound to the adversary's background knowledge and lower-bound to the number of unique trajectories in the dataset. And we introduced a new privacy threshold called privacy-height, to represent the degree of privacy offered to the users. Further, the proposed approach is tested with four different trajectory datasets, namely Geolife, T-Drive, Metro100K and private Wi-Fi dataset. The result shows that the anonymized dataset is freed from all four linkage attacks as well as better performance with a significant reduction in the information loss when compared to other states of the art methods.

Chapter 5

Quantify the Impact of Data Correlation on Privacy Budget Allocation in ϵ-Differential Privacy

ϵ-Differential Privacy (ϵ-DP) [86] is a popular privacy mechanism. It is proved that DP provides strong privacy guarantees to users against an adversary with unbounded knowledge. It ensures that any user's privacy leakage is to be strictly bounded by at most a ϵ value, where ϵ is a user parameter. If the value of ϵ is small, it achieves a strong privacy guarantee and vice versa. The ϵ-DP releases a noisy output instead of true output for hiding user's sensitive information. This noisy output is computed by adding a random noise (derived from the Laplace distribution with scale λ) to the true output.

Recently, the ϵ-DP privacy notion has been applied in settings of continuous data publishing [16, 17, 18]. For example, the traffic surveillance system periodically publishes a count of people (or users) at each location per timestamps in privately. In the literature, there exist a few privacy approaches such as event-level privacy[19], user-level privacy[20]

and w-event privacy[17] for continuous private data publishing. Event-level privacy provides a ϵ-DP guarantee to each event's (or each timestamp's) count. In other words, it protects only a single data-point of the user's entire stream. However, by combining all event's count, the adversary can reconstruct the user's stream, which leads to an effect on users' privacy [17]. In contrast, user-level privacy guarantees a ϵ-DP to finite event's (or timestamp's) count. In other words, it protects only a finite length of users' stream. Due to this, the user-level privacy has limited applicability in most of the real-world applications. The w-event privacy mechanism has been proposed to address the limited use of event-level privacy and user-level privacy. This mechanism offers a strong privacy guarantee to any user stream within a window of w timestamps. A w-event privacy presents a sliding window methodology that involves a broad range of w-event private mechanisms. Each mechanism constructs a separate sub-mechanism per timestamp, and each sub-mechanism uses a certain privacy budget to control the noise (higher privacy budget, lower perturbation, or less noise added). The w-event privacy achieves ϵ-DP when the sum of all privacy budgets used in any window of w timestamps is at most total privacy budget ϵ.

However, the w-event privacy mechanism provides less privacy guarantee than traditional ϵ-DP, especially when the user's data-points are not independent (i.e., temporally correlated) between consecutive timestamps. It happens due to the allotted privacy budget at timestamps within a window of size w is not adequate, especially where the data-points of users' stream involve temporal correlation. Therefore, the privacy budget distribution strategies in w-event privacy such as Budget Distribution (BD) and Budget Absorption (BA) are not suitable in the presence of correlated datasets within a window. The following example illustrates how privacy guarantee is degrading when users' data-points have a temporal correlation.

Assume that a trusted curator collects users' location data-points in continuous times-

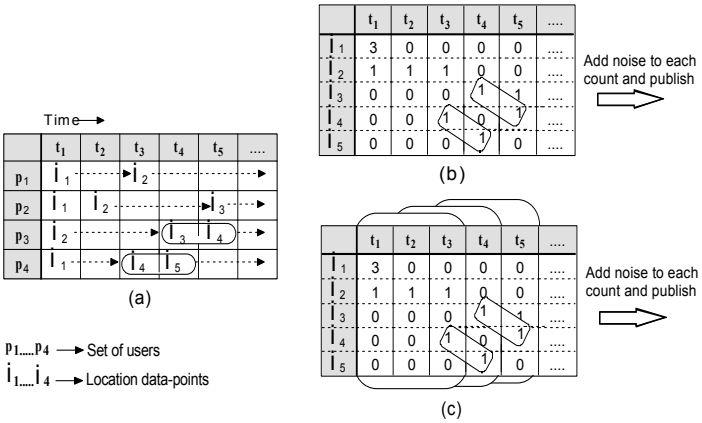

Figure 5.1: Illustration of example 1 (a) Collection of users location data-points in continuous timestamps (b) Statistics for event-level or user-level privacy (c) Statistics for w-event privacy while set w = 3

tamps, as shown in Figure 5.1(a). The curator aims to publish a statistic (*i.e., how many users are in each location)* at each timestamp without breach of any user privacy. According to the Laplace mechanism, the curator publishes private statistics at every timestamp using an independent random noise derived from the Laplace distribution with scale $Lap(1/\epsilon)$, where ϵ is a privacy budget. However, if the nature of two location data-points of users' stream at consecutive timestamps is temporally correlated, then the independent random noise with a scale $Lap(1/\epsilon)$ achieves 2ϵ-DP instead of ϵ-DP. It happens due to the modification or removal of one data point affects two counts in the published statistics (i.e., the global sensitivity (Δ) is 2), as shown in the Figure 5.1(b). Consequently, the presented two privacy budget distribution approaches in w-event privacy achieve $w\epsilon$-DP instead of ϵ-DP, especially when the nature of all location data-points of users' stream at consecutive timestamps are temporally correlated. Hence, the w-event privacy mechanism is not suitable for the publication of temporally correlated user streams, as shown in Figure 5.1(c).

Figure 5.2 shows that the distribution of total privacy budget ϵ to each timestamp within

a sliding window of size 3. In the first window of size 3, the traditional ϵ-DP privacy mechanism allots a privacy budget ($\epsilon/3$) uniformly to each timestamp. Then, compute noise lap$(1/(\epsilon/3))$ to perturb each timestamp counts by using the allotted privacy budget. Hence, the required ratio of the privacy budget at each timestamp is ϵ/3 to achieve ϵ-DP in the first sliding window. In the second sliding window, the required ratio of privacy budget at timestamps 2, 3 and 4 is ϵ/3, ϵ/3 and 2ϵ/3 respectively due to the presence of temporal correlation between the timestamps 3 and 4 as shown in the Figure 5.1(c). The sum of the privacy budgets in the second window is exceeded than the total privacy budget ϵ (i.e., $\epsilon/3 + \epsilon/3 + 2\epsilon/3 > \epsilon$). Thus, the second sliding window violates ϵ-DP privacy mechanism. Similarly, the third sliding window also violates ϵ-DP due to the presence of temporal correlation between the timestamps 4 and 5 (i.e., the required ratio of privacy budget at timestamp 5 is 2ϵ/3). Therefore, it is necessary to design a privacy budget distribution method for allocating a sufficient privacy budget to all timestamps within the sliding window of size w.

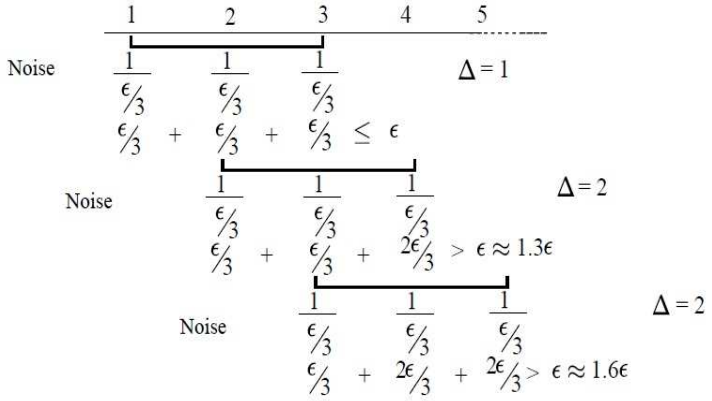

Figure 5.2: Distribution of privacy budget over timestamps(or event) within the sliding window of size $w = 3$.

Further, there is a limited state of art methods for distributing a privacy budget in continuous data publishing settings. There exist a few baseline approaches for allocating privacy

budgets in order to publish continuous location statistics privately. A Uniform method is to uniformly allocate a privacy budget to N timestamp's dataset. This approach achieves ϵ-DP since it combines all privacy budgets of N timestamp's datasets [75]. In our problem settings, the datasets at consecutive timestamps require more privacy budget than the baseline approach due to the presence of temporal-correlation. The fixed sampling [87, 20] is another approach for allocating a privacy budget at a given sampling interval I among N timestamps. Hence, the privacy budget at each interval I is $(\epsilon * I)/N$. It is also preserved ϵ-DP by combining the privacy budgets of all samples. This approach is not useful because pre-defined sampling intervals are not determined accurately if location data points arrive dynamically, and occur high perturbation errors if sampling intervals are too frequent.

In summary, the privacy budget distribution into the series of temporally correlated data-points in users stream remains unclear in the w-event privacy method. The contributions of this work are as follows.

1. We present a reformulated differential privacy definition for continuous data publication and prove that it can achieve ϵ-DP. Then we quantify the impact of temporal correlation on privacy leakage in reformulated ϵ-DP and analyze the privacy leakage in ϵ-DP with a numerical example.

2. We introduce a Privacy Budget Allocation method for allocating an adequate amount of privacy budget to each successive timestamps under the protection of ϵ-Differential privacy.

3. Finally, we evaluate the data utility of our method by computing the average error per timestamps through conducting a series of experiments on real and synthetic datasets.

The rest of this work is as follows. Section 5.1 introduces the necessary notations and

definitions of differential privacy under continual observation. In section 5.2, we analyze the impact of temporal correlation on privacy leakage with a numerical example. Section 5.3 proposes a Privacy Budget Allocation method (PBA) and presents a theoretical analysis of privacy leakage and utility leakage of this method. A numerical experiment conducted on various datasets to evaluate the data utility of our method is presented in section 5.4. Finally, the summary of this work is presented in section 5.5.

5.1 System Framework

5.1.1 Differential Privacy under continual observation

Let $\mathcal{M}$ be a privacy mechanism which takes stream prefixes S_t as input and produced a series of outputs $\omega = (\omega_1, \omega_2, ..., \omega_t) \in \Omega$ at each timestamp. The privacy mechanism $\mathcal{M}$ is said to be ϵ-Differentially private iff the following logarithmic function is to be bounded by maximum ϵ value for any adjacent stream prefixes S, Adj$(S_t, S_t^{'})$ and any possible output Ω of $Range(\mathcal{M})$.

$$\frac{P_r(\mathcal{M}(S_t) = (\omega_1, \omega_2, ..., \omega_t))}{P_r(\mathcal{M}(S_t^{'}) = (\omega_1, \omega_2, ..., \omega_t))} \leq \epsilon$$

Where the parameter ϵ quantifies the degree of a user privacy leakage. Suppose, $\mathcal{M}$ is decomposed into $(\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_t)$ sub-mechanisms. Each sub-mechanism $\mathcal{M}_i(D_i)$ produce an output ω_i with independent randomness. Hence, it holds $\frac{P_r(\mathcal{M}_i(D_i)=\omega_i)}{P_r(\mathcal{M}_i(D_i^{'})=\omega_i)} \leq e^{\epsilon_i}$ and

guarantees ϵ_i-DP. Therefore we derive

$$\frac{P_r(\mathcal{M}(S_t) = (\omega_1, \omega_2, ..., \omega_t))}{P_r(\mathcal{M}(S_t^{'}) = (\omega_1, \omega_2, ..., \omega_t))} = \prod_{i=1}^{t} \frac{P_r(\mathcal{M}_i(D_i) = \omega_i)}{P_r(\mathcal{M}_i(D_i^{'}) = \omega_i)}$$
$$\leq \prod_{i=1}^{t} e^{\epsilon_i} \leq exp(\sum_{i=1}^{t} \epsilon_i) \leq e^{\epsilon}$$

Let $A = \{A_i : i \in [1, k]\}$ be the set of adversaries with arbitrary knowledge and are interested in the user's private data. Consider an adversary A_i whose target is i^{th} user private data and has knowledge of all other users' private data except i^{th} user, i.e., A_i knows $S_t = S_t \backslash \{i\}$. The privacy leakage of privacy mechanism $\mathcal{M}$ (or i^{th} user) at timestamp t against A_i is as follows, in which l_i^t and $l_i^{t'}$ are two possible data points of i^{th} user at timestamps t.

$$\mathcal{L}_{A_i}(\mathcal{M}_t) = \sup_{\omega, l_i^t, l_i^{t'}} log \frac{P_r(\omega | l_i^t, S_t)}{P_r(\omega | l_i^{t'}, S_t^{'})}$$
$$\mathcal{L}(\mathcal{M}_t) = \max_{\forall A_i, i \in [k]} \mathcal{L}_{A_i}(\mathcal{M}_t)$$

The $\mathcal{L}(\mathcal{M}_t)$ is the maximum privacy leakage at timestamp t caused by any k adversary. Here, we considered a privacy budget ϵ as a metric of privacy leakage. If lesser ϵ value, then lesser the privacy leakage.

5.2 Temporal correlation (TC) privacy leakage analysis

5.2.1 Adversary's knowledge

In the stream data publication, it is fair to consider that an adversary knows the transition probability between the possible location data-points. In our settings, we adopted a

Markov chain process (MC) for modeling a transition probability between the location data-points (according to certain probabilistic rules) and is denoted as $\theta \in \Theta$, where Θ is a set of all transition probability distributions. In MC, the transition matrix describes the probabilities of transition from one data-point to another data-point, and the sum of transition probabilities in each row is equal to 1. Let consider a transition matrix of size 2 as shown in the Table 5.1(a). If a user i is at loc_1 (current location), then the probability of coming from loc_2 (previous location) is 0.4, represented as $P_r[l_i^{t-1} = loc_2 | l_i^t = loc_1] = 0.4$.

Table 5.1: Transition probability matrix and sample dataset

(a) Transition Matrix

	loc_1	loc_2
loc_1	0.6	0.4
loc_2	0.1	0.9

(b) Database D

	D_1	D_2
u_1	loc_1	loc_1
u_2	loc_1	loc_2

5.2.2 TC privacy leakage

Consider an adversary A_i with knowledge of $S_t = S_t \backslash \{i\}$ and transition probability distributions θ, named as A_i^θ. Let A_i^θ collects all private outputs which were published under the protection of ϵ-DP mechanisms $\mathcal{M}$ at each timestamps $t \in [1, T]$. Now, the aim of the adversary is to infer user $i's$ location data-point at timestamp t.

The TC privacy leakage $(\mathcal{TCL})$ of $\mathcal{M}_t$ w.r.t A_i^θ is the maximum ratio of two laplace distribution for all different values of $l_i^t, l_i^{t'}$ and for all possible transition probability distributions.

$$\mathcal{TCL}_{A_i^\theta}(\mathcal{M}_t) = \sup_{\omega, l_i^t, l_i^{t'}, \theta} log \frac{P_r(\omega \in \Omega | l_i^t, S_t, \theta)}{P_r(\omega \in \Omega | l_i^{t'}, S_t', \theta)} \tag{5.1}$$

The TC privacy leakage of $\mathcal{M}_t$ w.r.t any A_i^θ where $i \in [k]$ is less than or equal to ϵ, then we call $\mathcal{M}_t$ is ϵ-TC Differential privacy.

$$\sup_{\theta, \forall A_i, i \in [k]} \mathcal{TCL}_{A_i^\theta}(\mathcal{M}_t) \leq \epsilon \tag{5.2}$$

Further, to understand the impact of temporal correlation on privacy leakage in continuous data publish settings, Equation 5.1 is expanded and simplified by Bayes theorem, i.e.,

$$\begin{aligned}
\mathcal{TCL}_{A_i^\theta}(\mathcal{M}_t) &= \sup_{\omega_1,\ldots,\omega_t, l_i^t, l_i^{t'}, \theta} log \frac{P_r(\omega_1, ..., \omega_t | l_i^t, S_t, \theta)}{P_r(\omega_1, ..., \omega_t | l_i^{t'}, S_t', \theta)} \\
&= \sup_{\omega_1,\ldots,\omega_{t-1}, l_i^t, l_i^{t'}, \theta} log \frac{\sum_{l_i^{t-1}} P_r(\omega_1, ..., \omega_{t-1} | l_i^{t-1}, S_{t-1}) P_r(l_i^{t-1} | l_i^t)}{\sum_{l_i^{t-1'}} P_r(\omega_1, ..., \omega_{t-1} | l_i^{t-1'}, S_{t-1}') P_r(l_i^{t-1} | l_i^t)} \\
&+ \sup_{\omega_t, l_i^t, l_i^{t'}, \theta} log \frac{P_r(\omega_t | l_i^t, S_t)}{P_r(\omega_t | l_i^{t'}, S_t')}
\end{aligned} \tag{5.3}$$

There are three annotated terms in Equation 5.3. The first term determines privacy leakage at previous timestamp $t-1$; the second term indicates the probability of transition between the data-points of previous timestamp $(t-1)$ and current timestamp (t), and the last term is equal to the privacy leakage at time t. Hence, the privacy leakage at time t depends on the privacy leakage at time $t-1$, TC transition probability, and the privacy leakage at time t. Notice that, if $t = 1$, then $\mathcal{TCL}_{A_i^\theta}(\mathcal{M}_t) = \mathcal{L}_{A_i}(\mathcal{M}_1)$. Otherwise, if $t > 1$, then we have the following equation.

$$\mathcal{TCL}_{A_i^\theta}(\mathcal{M}_t) = \mathcal{TCL}_{A_i^\theta}(\mathcal{M}_{t-1}) + \mathcal{L}_{A_i}(\mathcal{M}_t) \tag{5.4}$$

The first term of the above equation 5.4 is calculated using the *temporal privacy loss function* given in [18]. We illustrate how the TC factor influences privacy leakage w.r.t adversary with and without knowledge of θ distribution through a numerical example.

Consider a query to find the user $i's$ location value l_i is either loc_m or loc_n at timestamp t, where loc_m and $loc_n \in L$. For simplicity, we assume that l_i^t is,

$$l_i^t = \begin{cases} 1 & \text{i's true location at time t,} \\ 0 & \text{Otherwise.} \end{cases} \tag{5.5}$$

Example: Let a database D of two users u_1 and u_2 (as shown in the Table 5.1(b)), A_1^θ and A_1 are the two adversaries with and without knowledge of θ distribution respectively and are interested in finding the location of u_1 at timestamp 2. Assume that both adversaries know the location information of u_2. According to the definition of $TC-DP$, we compute $\mathcal{TCL}_{A_1}(\mathcal{M})$ and $\mathcal{TCL}_{A_1^\theta}(\mathcal{M})$. For A_1 without knowledge of θ, we get

$$\begin{aligned}
\mathcal{TCL}_{A_1}(\mathcal{M}_2) &= \sup_{\omega_1,\omega_2} log \frac{P_r(\omega_1,\omega_2|l_1^2 = loc_1, l_2^2 = loc_2)}{P_r(\omega_1,\omega_2|l_1^{2'} = loc_2, l_2^2 = loc_2)} \\
&= \sup_{\omega_1} log \frac{\sum_{l_1^1} exp(-|\omega_1 - (l_1^1, l_2^1 = loc_1)|) P_r(l_1^1|l_1^2 = loc_1)}{\sum_{l_1^{1'}} exp(-|\omega_1 - (l_1^{1'}, l_2^1 = loc_1)|) P_r(l_1^{1'}|l_1^{2'} = loc_2)} \\
&+ \sup_{\omega_2} log \frac{exp(-|\omega_2 - (l_1^2 = loc_1, l_2^2 = loc_2)|)}{exp(-|\omega_2 - (l_1^{2'} = loc_2, l_2^2 = loc_2)|)} \\
&= 0 + \sup_{\omega_2} log \frac{exp(-|\omega_2 - 2|)}{exp(-|\omega_2 - 1|)} = 1
\end{aligned}$$

For A_1^θ with knowledge of θ, we get

$$\begin{aligned}\mathcal{TCL}_{A_1^\theta}(\mathcal{M}_2) &= \sup_{\omega} log \frac{P_r(\omega_1, \omega_2 | l_1^2 = loc_1, l_2^2 = loc_2)}{P_r(\omega_1, \omega_2 | l_1^{2'} = loc_2, l_2^2 = loc_2)} \\ &= \sup_{\omega_1} log \frac{\sum_{l_1^1} exp(-|\omega_1 - (l_1^1, l_2^1 = loc_1)|) P_r(l_1^1 | l_1^2 = loc_1)}{\sum_{l_1^{1'}} exp(-|\omega_1 - (l_1^{1'}, l_2^1 = loc_1)|) P_r(l_1^{1'} | l_1^{2'} = loc_2)} \\ &+ \sup_{\omega_2} log \frac{exp(-|\omega_2 - (l_1^2 = loc_1, l_2^2 = loc_2)|)}{exp(-|\omega_2 - (l_1^{2'} = loc_2, l_2^2 = loc_2)|)} \\ &= 0.55 + 1 = 1.55\end{aligned}$$

The above numeric analysis shows that TC has a significant influence on higher privacy leakage i.e., $\mathcal{TCL}_{A_1^\theta}(\mathcal{M}_2) > \mathcal{TCL}_{A_1}(\mathcal{M}_2)$. Hence, we can state that the curator (or data publisher) does not provide a strong privacy guarantee compared with traditional ϵ-DP in continuous data publication settings. In detail, a recent privacy method called w-event privacy allocates a ratio of privacy budget to each timestamp to achieve ϵ-DP guarantee of any user's stream within a window of size w by assuming the data-points in a user stream are independent. However, most of the location data-points are temporally correlated with a certain probability in real-time data collection. Due to this, the allotted privacy budget at timestamps within a window is not adequate to achieve ϵ-DP, resulting in more privacy leakage than the traditional ϵ-DP.

5.3 Proposed Method

This section discusses our Privacy Budget Allocation (PBA) mechanism, which is allowed to compute and allocate the quantity of privacy budget to each publication in a continuous data release setting. Then, we theoretically prove that our PBA mechanism achieves ϵ-DP and shows the data utility of PBA mechanism.

This mechanism is motivated by limited use of previous mechanisms such as Uniform, Sampling, and w-event privacy, which are discussed in the introduction. In this mechanism, we adopt a w-event privacy concept called sliding window methodology, and it follows that the window is moving one timestamp ahead after every w timestamps. A sliding window consists of w number of timestamps and each timestamp t is operated by a sub-mechanism $\mathcal{M}_t$. Since each $\mathcal{M}_t$ uses independent randomness, $\mathcal{M}_t$ achieves ϵ_t-DP for some ϵ_t. The sum of the privacy budgets within the sliding window of size w must be lesser than or equal to the total privacy budget ϵ. Note that, at any timestamp t, span of sliding window is $t - w + 1$ to t.

The PBA mechanism $\mathcal{M}$ consists of series of sub-mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k, \ldots, \mathcal{M}_t$, where each $\mathcal{M}_k$ takes dataset $S_t[k] = D_k$ as input and publishes a private statistic ω_k as output by using allotted privacy budget ϵ_k. Thus, $\mathcal{M}$ publishes a series of private statistics, namely $\omega_1, \omega_2, .., \omega_k, .., \omega_t$. In detail, the mechanism $\mathcal{M}$ involves two phases $\mathcal{M}^1$ and $\mathcal{M}^2$. These two phases operate sequentially by using half of the total privacy budget, i.e., ϵ^1 and ϵ^2. In the first phase, $\mathcal{M}^1$ allocates a ratio of privacy budget from ϵ^1 to each timestamp uniformly within a sliding window. At timestamp k, the sub-mechanism $\mathcal{M}_k^1$ calculates a dissimilarity value between the true statistic a_k and last release private statistic ω_l. The mean of absolute error (MAE) is a metric which measures the dissimilarity between a_k and ω_l and is formulated as $\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\omega_l[j] - a_k[j]|$ where a_k and ω_l are the vectors of length $|L|$. Then, the obtained dissimilarity value is forwarded into $\mathcal{M}_k^2$. In the second phase, $\mathcal{M}^2$ divides a privacy budget ϵ^2 into two parts; namely publication privacy budget and absorption privacy budget. The $\mathcal{M}^2$ allocates a publication privacy budget into each timestamp in an exponential decreasing fashion. At timestamp k, $\mathcal{M}_k^1$ forwards dissimilarity value to $\mathcal{M}_k^2$ to decide whether to publish a true publication with noise or null publication (last release private output). If $\mathcal{M}_k^2$ decides not to publish a true publication at timestamp k, then k^{th}

Algorithm 5.1 Pseudocode of PBA mechanism at k^{th} timestamp ($\mathcal{M}_k$)

INPUT: Dataset D_k, total privacy budget ϵ, $\epsilon_1^2, \epsilon_2^2, \ldots, \epsilon_{k-1}^2$ and $\epsilon_1^a, \epsilon_2^a, \ldots, \epsilon_{k-1}^a$
OUTPUT: Release Noisy output ω_k.

At sub-mechanism $\mathcal{M}_k$
  Compute noise for last release output ω_l
    Calculate $a_k = M(D_k)$
    Calculate $MAE(\omega_l, a_k)$
    Allocated budget at time k: $\epsilon_k^1 = \epsilon \cdot |L|/(2 \cdot w)$
    Compute noise $\lambda_k^1 = 1/\epsilon_k^1$
    Set $MAE() = MAE(\omega_l, a_k) + Lap(\lambda_k^1)$
  Compute noise for present publication ω_k
    Calculate remaining budget: $\epsilon_k^2 = (\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])/2$
    If Correlation exists
      $\epsilon_A = (\epsilon/4 - [\sum_{j=k-w+1}^{k-1}\epsilon_j^a])$
      Set $\epsilon_k^a = \epsilon_k^2$ (absorbed from ϵ_A)
      $\epsilon_k^2 = \epsilon_k^a + \epsilon_k^2$
    Compute noise $\lambda_k^2 = 1/\epsilon_k^2$
  If $MAE() > \lambda_k^2$
      return $\omega_k = a_k + \langle Lap(\lambda_k^2)\rangle^{|L|}$
  Else
      return $\omega_k = \omega_l$
 end for
end for

allotted publication privacy budget is become free and can be used in the future publication if necessary. In contrast, if $\mathcal{M}_k^2$ decides true publication at timestamp k, then it uses allotted publication privacy budget to publish statistics privately. Further, the absorption privacy budget allocates an extra privacy budget at timestamp k only when a correlation exists between the current timestamp k and the previous timestamp $k-1$. This is because a statistic at timestamp k requires more privacy budget compared to the normal publication. Algorithm 5.1 describes the mechanism of PBA in continuous data release settings. It takes Dataset D_k, total privacy budget ϵ, allotted budgets upto $(k-1)^{th}$ timestamp $(\epsilon_1^2, \epsilon_2^2, \ldots, \epsilon_{k-1}^2)$ and $(\epsilon_1^a, \epsilon_2^a, \ldots, \epsilon_{k-1}^a)$ are inputs and release a noisy statistic ω_k as an output. PBA aims to allocate an adequate amount of privacy budget at each timestamp within the sliding

window to achieve ϵ-DP (line 2-20). At any timestamp k, the sliding window follows two phases: the first phase is to calculate noisy dissimilarity value between ω_l and a_k (line 3-8), and the second phase is to decide whether publish a private statistics of current timestamp k if publications occur, otherwise publish ω_l (line 9-19). The sub-mechanism $\mathcal{M}_k$ computes true answer (a_k) from the dataset $S[k] = D_k$ (line 4), then calculates dissimilarity value between a_k and ω_l by using a metric called Mean of Absolute Error (line 5). After that, $\mathcal{M}_k$ utilizes an allotted privacy budget ϵ_k^1 to make *noisy dissimilarity value* shown in the lines (6-8). Finally, $\mathcal{M}_k$ computes MAE() value (MAE + noise) and forwards into phase 2 of $\mathcal{M}_k$. In the second phase, $\mathcal{M}_k$ starts with finding a remaining amount of privacy budget available at the time of k^{th} timestamp and assign half of the remaining budget to the phase 2 of k^{th} timestamp (line 10). If a correlation exists between the present and previous timestamps, then $\mathcal{M}_k$ adds extra budget from the absorbed privacy budget ϵ_A to the phase 2 of k^{th} timestamp, is shown in the lines (12-15). Once $\mathcal{M}_k$ computes noise (line 16), then it decides whether publish a_k with noise or ω_l based on the comparison between MAE() and computes noise λ_k^2 (lines 17-20). If MAE() is greater than λ_k^2, then it releases a_k with noise otherwise releases last release output ω_l.

Figure 5.3 shows the operation of PBA mechanism in continuous data release settings of 5 timestamps while assuming the size of $w = 3$. Assume that $\mathcal{M}$ publishes private outputs at timestamps $1, 3, 4, 5$ and last release private output at timestamp 2 i.e., the noisy output of timestamp 1. At each timestamp in phase 1, $\mathcal{M}$ allocates a fixed privacy budget i.e., $\epsilon/2 \cdot w = \epsilon/6$ (fix $w = 3$). Then $\mathcal{M}$ allocates half of the allotted privacy budget in phase 2 (i.e., $\epsilon/4$) in an exponential decreasing manner within the sliding window of size $w = 3$. In other words, at timestamp 1, it assigns $\epsilon_1^2 = (\epsilon/4 - 0)/2 = \epsilon/8$. At timestamp 2, $\epsilon_2^2 = 0$ because no output is generated at timestamp 2. At timestamp 3, $\epsilon_3^2 = (\epsilon/4 - (0 + \epsilon/8))/2 = \epsilon/16$. Since no correlation exists between the timestamps

within the first sliding window, it is not required to add extra budget to any timestamps. At timestamp 4, $\epsilon_4^2 = (\epsilon/4 - (0 + \epsilon/16))/2 = 3\epsilon/32$ and adds extra budget $3\epsilon/32$ to ϵ_4^2 due to the existence of correlation in between the timestamps of 4 and 3 as shown in Figure 5.1. Similarly at timestamp 5, the PBA assigns $\epsilon_5^2 = (\epsilon/4 - (\epsilon/16 + 3\epsilon/32))/2 + 3\epsilon/64$. Notice that the total sum of all privacy budgets in phases 1 and 2 of respective sliding windows is less than or equal to the total privacy budget ϵ.

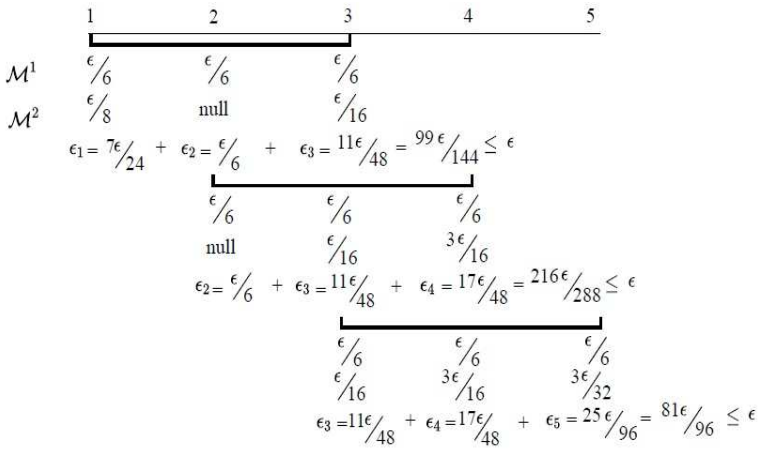

Figure 5.3: Distribution of privacy budget over timestamps(or event) within the sliding window of size $w = 3$.

5.3.1 Privacy Analysis

Theorem 5.3.1. *Privacy Budget Allocation algorithm (PBA) achieves ϵ-Differential privacy.*

Proof. A sub-mechanism $\mathcal{M}_k$ privately publishes either output of $q(D_k)$ or immediate last release output ω_l by utilizing a privacy budget ϵ_k. The sub-mechanism $\mathcal{M}_k$ has two phases that use independent privacy budgets, i.e., ϵ_k^1 and ϵ_k^2. Hence, we first prove that the sub-mechanism at phase 1 $\mathcal{M}_k^1$ satisfies ϵ_k^1-DP for $\epsilon_k^1 = \epsilon/2w$ and $\mathcal{M}_k^2$ satisfies ϵ_k^2-

DP for $\epsilon_k^2 = (\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])$ if it publishes output of $q(D_k)$, otherwise $\epsilon_k^2 = 0$. In phase 1, $\mathcal{M}_k^1$ publish private MAE() value i.e.,$q'(D_k) = \frac{1}{|L|}\Sigma_{j=1}^{|L|}|\omega_l[j] - a_k[j]|$. If add or remove a row from D_k, then the maximum alter in the result of $q'(D_k)$ is $1/|L|$. Hence, the sensitivity of q' is at most $1/|L|$. By using this sensitivity, $\mathcal{M}_k^1$ injects laplace noise with scale $\lambda_k^1 = \Delta(q')/\epsilon_k^1 = 2 \cdot w/(\epsilon \cdot |L|)$ to MAE() value. According to definition 2, $\mathcal{M}_k^1$ is ϵ_k^1-DP for $\epsilon_k^1 = \frac{1/|L|}{(2\cdot w)/(\epsilon\cdot|L|)} = \epsilon/(2 \cdot w)$. In the second phase, $\mathcal{M}_k^2$ publishes either private output of $q(D_k)$ value or null. In former differential privacy, if add or remove a row from D_k, then the maximum alter in the result of $q(D_k)$ is 1. Hence, the sensitivity of q is at most 1. The $\mathcal{M}_k^2$ injects laplace noise with scale $\lambda_k^2 = 2/(\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])$. In case if a correlation exists between the current timestamp and previous timestamp, then $\mathcal{M}_k^2$ borrow extra budget from the available $\epsilon/4$. So, $\mathcal{M}_k^2$ injects laplace noise with scale $\lambda_k^2 = 2/(\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j]) + 2/(\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j]))$ if a correlation exists; otherwise, noise is $\lambda_k^2 = 2/(\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])$. We assume that the mechanism $\mathcal{M}$ used an entire extra budget $(\epsilon/4)$ within a sliding window of size w. According to definition 2, $\mathcal{M}_k^2$ is ϵ_k^2-DP for $\epsilon_k^2 = (\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])/2 + (\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])/2) = (\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])$.

Subsequently, we must prove that PBA holds $\Sigma_{j=k-w+1}^{k}\epsilon_j \leq \epsilon$, for every k within the sliding window. From composition property, PBA holds at j^{th} privacy budget is $\epsilon_j = \epsilon_j^1 + \epsilon_j^2$, then it equals to $\Sigma_{j=k-w+1}^{k}\epsilon_j = \Sigma_{j=k-w+1}^{k}\epsilon_j^1 + \Sigma_{j=k-w+1}^{k}\epsilon_j^2$. Since every ϵ_j^1 is set to $\epsilon/2 \cdot w$, the total privacy budget within the sliding window is $\Sigma_{j=k-w+1}^{k}\epsilon_j = \epsilon/2 + \Sigma_{j=k-w+1}^{k}\epsilon_j^2$. Now, it is required to prove that $\Sigma_{j=k-w+1}^{k}\epsilon_j^2 \leq \epsilon/2$. In our settings $\Sigma_{j=k-w+1}^{k}\epsilon_j^2$ is $\Sigma_{j=k-w+1}^{k}(\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])/2 + (\epsilon/4 - [\Sigma_{j=k-w+1}^{k-1}\epsilon_j])/2)$. These two terms can be proved using inequality by induction. Since both terms are equal, we prove either one of the terms is lesser than or equal to $\epsilon/4$. and then we can say that another term is also lesser than or equal to $\epsilon/4$. In the induction part, first, simplify the term using geometric series and prove that the term is lesser than or equal to $(\epsilon/2)$ using inductive steps (for more

details in appendix A). Therefore, sub-mechanism $\mathcal{M}_k^2$ is always used up to half of the available privacy budget, i.e.,$(\epsilon/2)$.

5.3.2 Utility Analysis

In the PBA mechanism, the error at any timestamps depends on two reasons 1) a privacy budget utilized in true publications and 2) a privacy budget utilized in the last release publication, which is an approximated publication of the current timestamp. In detail, if publications occur at timestamp k, then $\mathcal{M}$ operates both the phases $\mathcal{M}_k^1$ and $\mathcal{M}_k^2$. We use the MAE metric of pair (ω_l, a_k) and pair (ω_k, a_k) for calculating error of publications at $\mathcal{M}_k^1$ and $\mathcal{M}_k^2$ respectively. If publication does not occur at timestamp k, then the mechanism $\mathcal{M}$ calculates the MAE metric of pair (ω_l, a_k) as a error of publications at timestamp k. Therefore, the mechanism $\mathcal{M}$ produces error per timestamps from one or both of the phases. Next, we show that the average error per timestamp in the PBA mechanism. Assume that there is an equal number of skipped publications between every occurrence of true publications, and n represents the total number of true publications that occur within a sliding window of size w.

Theorem 5.3.2. *The average error per timestamp in PBA is at most* $\frac{4}{n\epsilon} \cdot (2(2^n - 1) + 1) + \frac{2w}{\epsilon |L|}$

Proof. At timestamp k, the private dissimilarity value of $\mathcal{M}_k^1$ guides $\mathcal{M}_k^2$ for deciding to publish either true publication or null publication. Hence, we consider both $\mathcal{M}_k^1$ and $\mathcal{M}_k^2$ for computing average error per timestamp in PBA mechanism. The $\mathcal{M}_k^1$ induces error when its private dissimilarity value suggests $\mathcal{M}_k^2$ to make a wrong decision, i.e., wrongly skips a publication or wrongly performs true publication. If $\mathcal{M}_k^1$ suggest true publication occur at time k, then the error at timestamp k is the error induced by $\mathcal{M}_k^2$, which is discussed later. Alternatively, if $\mathcal{M}_k^1$ suggest true skipped publication occur at time k, then the error at timestamp k is an original dissimilarity value between ω_l and a_k,

it is bounded by the error of $\mathcal{M}_k^2$. However, if wrongly performs true publication at time k, then the original dissimilarity value is overrated due to noise of scale λ_k^1 added by the $\mathcal{M}_k^1$. Conversely, if wrongly skips a publication at time k, then the original dissimilarity value is underrated due to noise with scale λ_k^1 added by the $\mathcal{M}_k^1$. Therefore, the error induced by the $\mathcal{M}_k^1$ in PBA is at most $\frac{2w}{\epsilon|L|}$. Recall that PBA allocates privacy budget in exponential decreasing fashion in phase 2 of $\mathcal{M}$ i.e., $\epsilon/8, \epsilon/16, \epsilon/32, ...$ Hence, the error per each timestamp in phase 2 is $1/\epsilon_r$, where ϵ_r is a exponentially decreasing privacy budget. Moreover, $\mathcal{M}^2$ phase uses the extra budget from the absorbed privacy budget, so the error at $\mathcal{M}^2$ from the absorbed privacy budget is at most $(4/\epsilon)$. Therefore, the average error per timestamps within a sliding window of $\mathcal{M}^2$ in PBA is equal to

$$
\begin{aligned}
&= \frac{1}{n} \cdot (\frac{8}{\epsilon} + \frac{16}{\epsilon} + \cdots + \frac{2^{n+2}}{\epsilon}) + \frac{4}{n\epsilon} \\
&= \frac{8}{n\epsilon} \cdot (2^n - 1) + \frac{4}{n\epsilon} \\
&= \frac{4}{n\epsilon} \cdot (2(2^n - 1) + 1)
\end{aligned}
$$

The total average error per timestamps within the sliding window of size w in PBA mechanism is

$$
= \frac{4}{n\epsilon} \cdot (2(2^n - 1) + 1) + \frac{2w}{\epsilon|L|}
$$

5.4 Experimental Results

In this section, we conduct an experiment to demonstrate users' privacy risk, especially when the datasets have temporal-correlation at certain timestamps. Furthermore, we validate the effectiveness of our proposed algorithm with existing states of art methods.

We employed three real-time trajectory datasets such as $Geolife$ [88], $T-Drive$ [89, 90], $ShangHai$ [91], and a synthetic dataset, namely $Metro100K$ [31] in our experiments. A $Geolife$, $T-Drive$ and $Metro100K$ datasets are described in chapter 3. Apart from these three datasets, we used $ShangHai$ dataset in this work. It is a public trajectory data of about 5000 buses and taxis in Shanghai collected by the Hong Kong University of Science and Technology on February 20, 2007. The sampling interval of data is approximately 60 seconds. We optimized all real-time datasets for our experiment by considering that a user is located at most one location at each timestamp and collected all samples (user's location data-point) every 5 minutes. The above trajectory datasets consist of a series of tuples containing ID, timestamp, latitude, and longitude. We filter the real-time datasets using two minimum requirements, i.e., each user is located at most one location at each timestamp, and collect samples from the datasets according to the curator's pre-defined timestamp frequency (ex: every 5 min).

5.4.1 Metrics of Utility Evaluation

We used two popular metrics such as *Mean of absolute error(MAE)* and *Mean of square error(MSE)* [92] for quantifying the data utility of our mechanism with existing states of art methods. These two metrics are used to measure the dissimilarity (or error) between the measured value and actual value. Moreover, these two metrics have good mathematical properties, and also the MSE metric helps to find larger errors. In our settings, the PBA method protects each user's data point per timestamp (or user stream) within the sliding window of size w. Hence, we measure the error per timestamp by using MAE and MSE

metrics. The definition of MAE and MSE metric is as follows.

$$MAE() = \frac{1}{|T|}\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\omega_l[j] - a_t[j]|$$
$$MSE() = \frac{1}{|T|}\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\omega_l[j] - a_t[j]|^2$$

5.4.2 Result Analysis

According to our problem settings, we train the Markov model for modeling a transition probability between all possible location data-points. This transition probability matrix describes how a data-point is dependent on other possible remaining data-points. There are three types of temporal correlation, such as strong, moderate, and no correlation. Let assume that the temporal correlation between the data-points of a user stream is strong, i.e., $\theta^s = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The privacy-risk of users per timestamp is increasing linearly because the same datasets are released in contiguous timestamps, as shown in Figure 5.4. In other words, the location data-points of users are at time t to time 1 is equivalent, i.e., $loc^t = loc^{t-1} = \cdots = loc^1$. Hence, the privacy risk is increasing linearly at each timestamp. In another extreme case, is a moderate temporal-correlation between the data-points, say $\theta^m = \begin{bmatrix} 0.8 & 0.2 \\ 0 & 1 \end{bmatrix}$. The line with circle shape in Figure 5.4 shows users' privacy risk from timestamp 1 to t, which can be quantified by using equation 5.4. Finally, a line with a triangle shape in Figure 5.4 shows each timestamp achieves 1-DP while assuming $\epsilon = 1$ because no temporal-correlation exists.

Further, we demonstrate the effectiveness on TC privacy leakage when the transition probability matrix involves a large (or small) number of dimensions. Let consider a transition probability matrix with a moderate correlation and d be the number of dimensions in the transition matrix. If d is large, then the probability value on cells is well scattered in transition matrix. Figure 5.5 shows the variation of privacy leakage when the size of d varies in

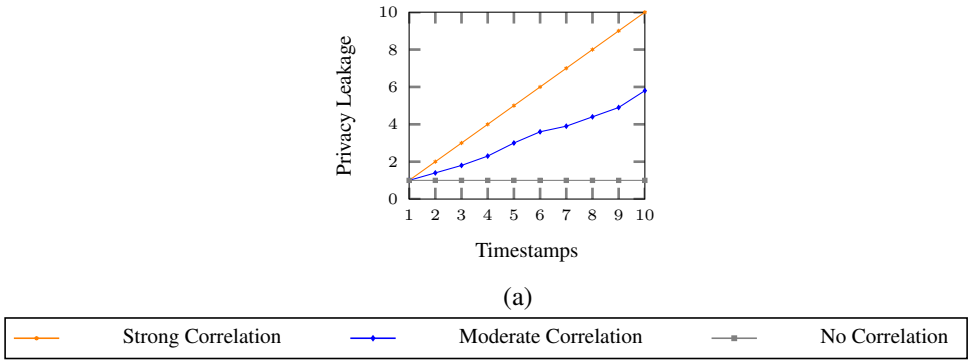

Figure 5.4: Analysis of privacy risk (or leakage) of 1-DP under different types of temporal correlation

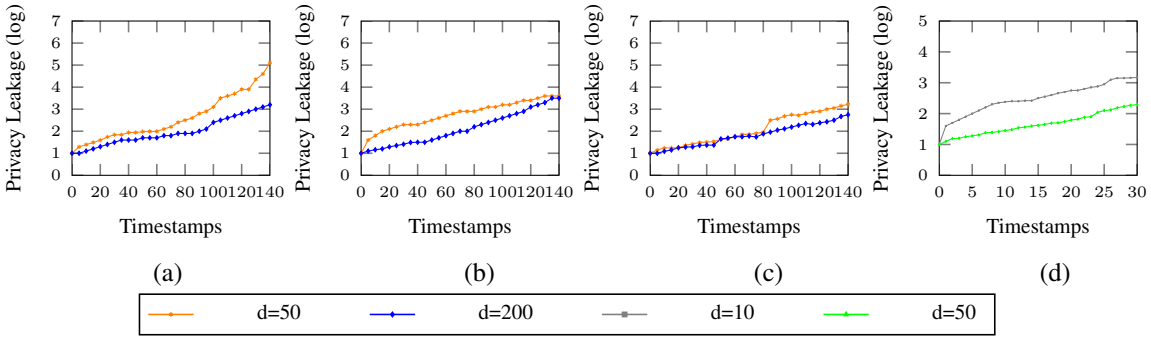

Figure 5.5: Privacy leakage versus different degrees of correlation while set $\epsilon = 1$ (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

the transition probability matrix. The results show that the ϵ-DP attains less privacy leakage when the vast number of dimensions in the transition probability matrix. It is depicted in the lines $d = 50$ and $d = 10$ of Figure 5.5 by considering $\epsilon = 1$. This is because the data points in a matrix are very close to the stronger correlation. In other words, a stronger correlation in the transition matrix results in more privacy leakage. The transition matrix involves a weaker correlation when the matrix dimension is larger, as shown in the lines $d = 200$, $d = 50$ of Figure 5.5.

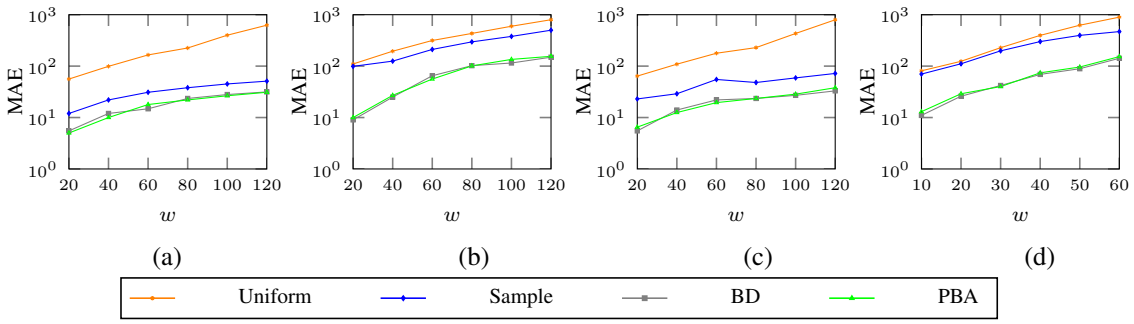

Figure 5.6: MAE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

5.4.3 Compare with Baseline approaches

Firstly, we started to compare our proposed method with baseline approaches such as Uniform, Sampling, and w-event privacy, to analyze the effectiveness of our method while varying the size of the sliding window w and ϵ value. Figures 5.6 and 5.7 show the result of MAE and MSE values between the PBA method with baseline approaches, while varying the size of the sliding window w. We observed that our PBA method outperforms with baseline approaches on all datasets. This is because the rate of allotted privacy budget within the sliding window is minimized in baseline approaches due to an increase in the number of timestamps in the sliding window. In other words, the adequate amount of privacy budget is allotted at temporally correlated timestamps in the PBA method compared to baseline approaches. Notice that MAE and MSE value in uniform method increases linearly when w increases because fixed privacy budget allotted at each timestamp. Similarly, a privacy budget is allotted only at a given sample interval in a sampling approach.

Furthermore, the MAE values in PBA and BD methods (see Figures 5.6(b), (c) and (d)) are approximately the same because both methods follow the same allocation scheme, i.e., exponential decreasing fashion. However, our PBA method achieves ϵ-DP even though the location data-points are temporally correlated at consecutive timestamps.

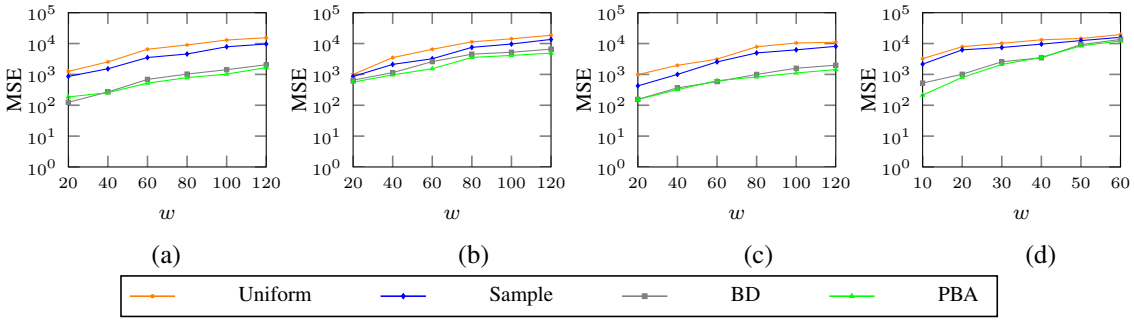

Figure 5.7: MSE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

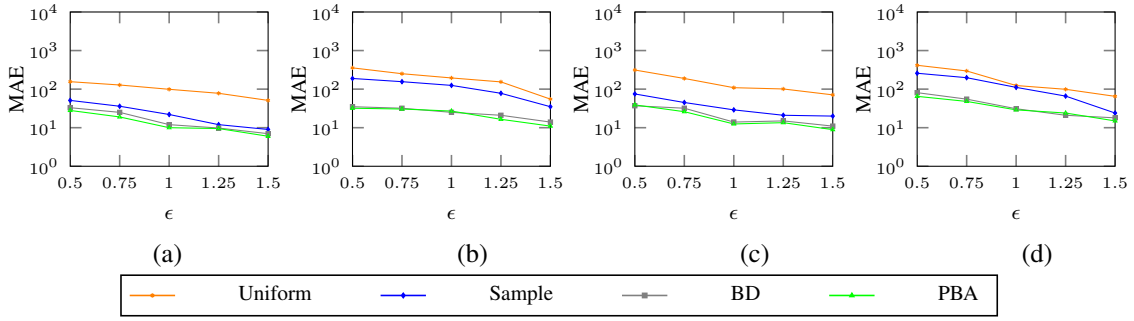

Figure 5.8: MAE vs. ϵ while fixing w=40 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

Further, we compared our PBA method with baseline approaches while varying ϵ values, as shown in the Figure 5.8 and 5.9. The result shows that the error rate of MRE and MSE is comparatively low while assigning a larger privacy budget at each timestamp. The MRE and MSE values of baseline approaches have more update errors while comparing our PBA method. This is because the uniform and sample methods use a fixed privacy budget at each timestamp even-though the location data-points are temporally correlated. Further, the error rate of the PBA and BD approach has almost similar because the PBA adopts a similar allocation scheme (i.e., exponential decreasing fashion at stage 2) as in the BD approach. However, our PBA method allocates an adequate amount of privacy budget

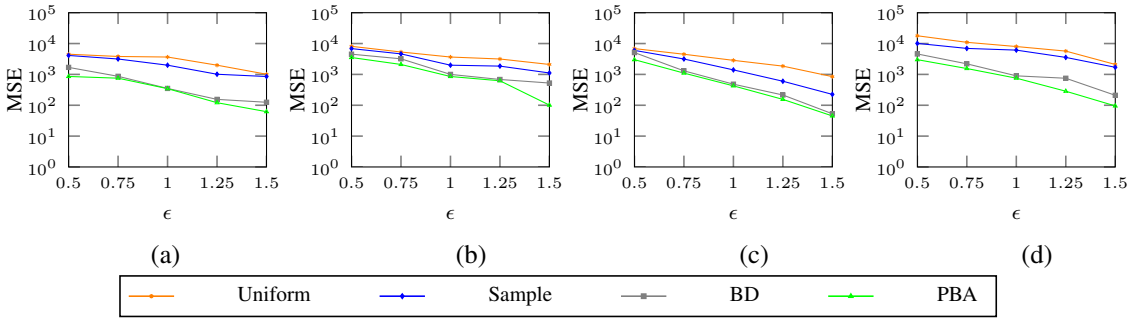

Figure 5.9: MSE vs. ϵ while fixing w=40 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

even though the location data-points are temporally correlated at consecutive timestamps.

5.4.3.1 Analysis and Evaluation

In literature, several works have been proposed that consider the correlation between the users in the dataset, i.e., user-user correlation, whereas we consider a correlation among the single user's location data-points at different timestamps (i.e., temporal correlation). We found limited research on finding the privacy risk of differential privacy under temporal correlation for the continuous location data release settings. However, Quantification [18] and Planar Isotropic Mechanisms (PIM) [59] are the recent privacy budget allocation methods for temporal correlation in continuous location data release settings. Figures 5.10 and 5.11 show the error rate of MAE and MSE between our PBA method with other states of art methods such as Quantification and PIM while varying the size of the sliding window w. The experimental results exhibit that the proposed PBA method provides significant data utility compared to the above two methods. The error rate of the quantification and PBA method are approximately similar. This is because the proposed method achieves ϵ-DP under a temporal correlation whereas the quantification method achieves α-DP under temporal correlation instead of ϵ-DP, where α is increased privacy leakage of

range $\epsilon \leq \alpha \leq T\epsilon$ (assume that the length of temporally correlated data-points in user's stream is T). In other words, the quantification method allocates more privacy budget (i.e., exceeds than the allotted budget of traditional ϵ-DP) to each timestamps. Hence, the error rate (MAE and MSE) of quantification is almost same to the PBA method under temporal correlation. And since the privacy budgets are assigned to each timestamps uniformly in PIM, the error rate (MAE and MSE) of PIM increases linearly, as shown in Figures 5.10 and 5.11.

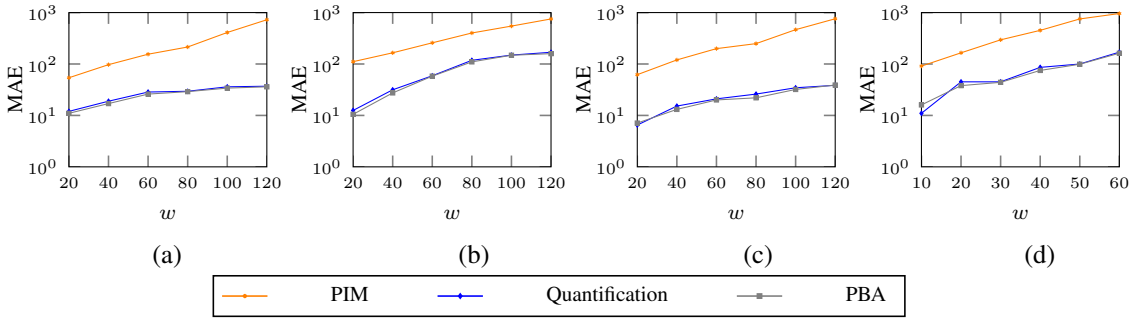

Figure 5.10: MAE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

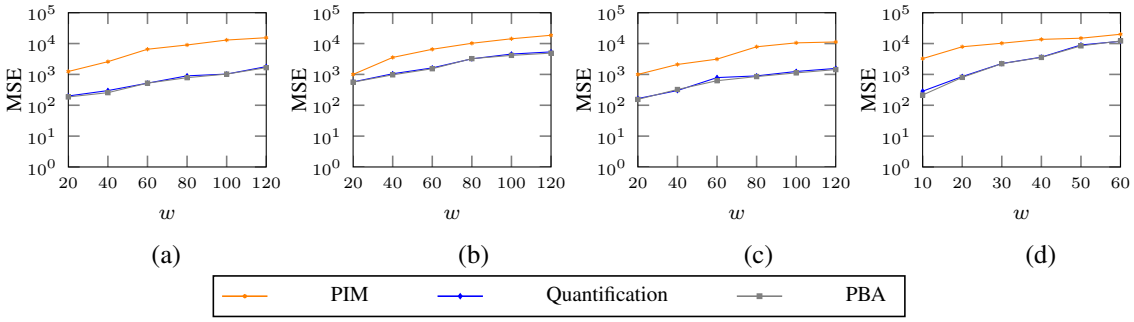

Figure 5.11: MSE vs. w while fixing ϵ=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

Figures 5.12 and 5.13 show the error rate of MAE and MSE between our PBA method with Quantification and PIM while varying ϵ values. The result shows that the MRE and

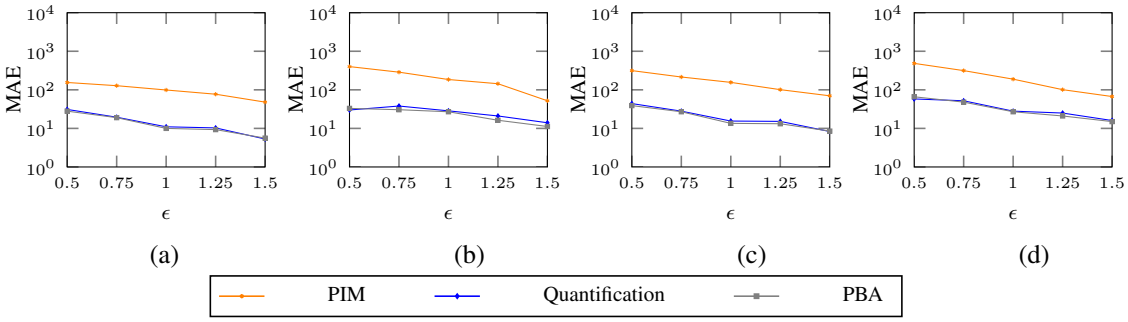

Figure 5.12: MAE vs. ϵ while fixing w=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

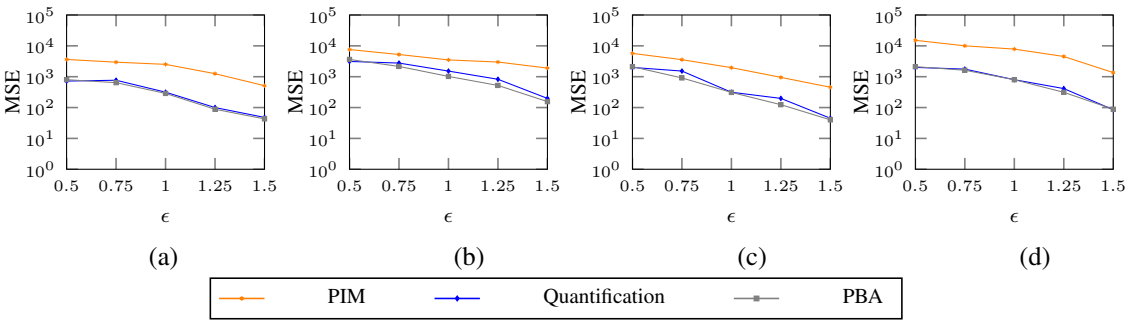

Figure 5.13: MSE vs. ϵ while fixing w=1 (a) Geolife (b) T-Drive (c) ShangHai and (d) Metro100K datasets.

MSE values decreasing when assigning more privacy budget at each timestamp while setting w=40. The MAE and MSE of the PBA method are closely related to the quantification method because the quantification method uses an increased privacy budget α, which leads to fewer update errors even though the user's stream involves temporal correlation. Another side, the PIM method follows a uniform approach for allocating privacy budgets that leads to having more update errors, as shown in Figures 5.12 and 5.13.

Table 5.2 shows that the privacy guarantee of various privacy budget allocation methods on temporally correlated data-points of length T. We observed that our PBA method achieves ϵ-DP under temporal correlation in continuous stream data publishing compared

Table 5.2: The privacy guarantee of various privacy budget allocation methods on temporally correlated data-points of length T

Privacy approaches	Privacy Budget Allocation scheme	Temporal correlation	privacy guarantee on T length user stream
Uniform[75]	✓	–	Tϵ-DP
Sampling[87]	✓	–	(T/I)ϵ-DP
Budget Distribution (BD)[17]	✓	–	wϵ-DP
PIM[59]	–	✓	Tϵ-DP
Quantification[18]	✓	✓	α-DP
Proposed method	✓	✓	ϵ-DP (i.e., $\Sigma_{j=k-w+1}^{k}\epsilon_j \leq \epsilon$)

with other DP approaches under temporal correlation. The existing approaches such as Uniform, Sampling, BD, PIM, and the Quantification method are provided less privacy guarantee under a temporal correlation, i.e., Tϵ-DP, (T/I)ϵ-DP, wϵ-DP, Tϵ-DP and α-DP respectively. In other words, the existing approaches require more privacy budgets ($\geq \epsilon$) in continuous temporally correlated stream data publishing to satisfy the definition of ϵ-DP. Even though BD and PBA methods follow the same allocation strategy, BD achieves $w\epsilon$-DP (assume that the length of temporally correlated data-points in the user's stream is w) instead of ϵ-DP under temporal correlation. Hence, the PBA method is a better approach for allocating privacy budgets to temporally correlated data-points.

5.5 Summary

In this work, we present the definition of differential privacy under temporal correlation to quantify the impact of temporal correlation on privacy leakage. Then, we illustrate and prove that the adversaries who have knowledge of temporal correlation can disclose more privacy leakage than the traditional ϵ-DP. Our analysis result shows that the privacy leakage increases over time in w-event privacy when the dataset involves temporal correla-

tion. Therefore, we introduce a privacy budget allocation (PBA) method for allocating an adequate amount of privacy budget to each successive timestamp under the protection of ϵ-differential privacy. This method protects any w length user stream that contains temporally correlated data-points. We conducted a series of experiments with real and synthetic datasets to evaluate the average error per timestamp for analyzing the data utility of our method.

Chapter 6

Quantify the Impact of Data Correlation on Privacy Leakage in ϵ-Local Differential Privacy

Local differential privacy (LDP)[10] is a variant of standard differential privacy. It addresses the privacy issue in the data collection phase (i.e., untrusted service provider) and is implemented in many applications, such as the Google Chrome browser[61], Apple's iOS-10 [62][63]. In particular, users make their location data-point private before sending it to the service provider. Hence, this privacy model promises a privacy guarantee to users even though the service provider is not trusted. Under the protection of LDP, the service provider can still compute the correct statistical results even though not collecting users' private location data-points. According to the ϵ-LDP, the adversary cannot infer users' private or sensitive data with high confidence (controlled by ϵ). Here, ϵ is a privacy budget that controls the level of privacy guarantee. If ϵ is small, then it signifies higher privacy protection or vice versa. Most of the existing methods under LDP focuses on one-time data

publishing [61][93][66]. These methods face the challenge of privacy degradation if the location data points are collected over time [64]. To overcome this challenge, in literature, few existing methods have been proposed for continuous data publishing, such as event level, user-level, and w-event privacy. The event-level privacy [44] achieves ϵ-DP at each timestamp's data (or location data-point). So, it protects only the user's single location data-point, not the entire user's stream. In contrast, user-level privacy [20] protects user's stream data up to a limited length. If the length of a user stream is long, it is required to add more noise, which reduces the data utility of the user stream. Both methods have limited applicability in real-world applications.

Recently w-event privacy has been proposed to address event level and user-level privacy issues. It offers privacy to the user's stream of length w; hence it is called the w-event privacy method. If w is set to 1, then w-event privacy becomes event-level privacy, and if it is set to infinity, it becomes user-level privacy. This privacy method adopts a concept called sliding window methodology of size w, and it moves one timestamp's data ahead after every w timestamp's data till it reaches the last timestamp's data of a user stream. Each sliding window of size w achieves ϵ-LDP by consuming a ratio of privacy budget to each timestamp's data within the sliding window of size w. Moreover, the sum of the privacy budgets of each timestamp's data within the window should not be more than the total privacy budget (ϵ). However, w-event privacy offers comparatively less privacy guarantee when a user stream's location data-points are correlated. This is because the privacy budget allots to each timestamp's data-point are not adequate due to the presence of correlated location data points within the user stream. Moreover, w-event privacy presents two budget allocation strategies, namely Budget Distribution (BD) and budget Absorption (BA), for allocating a budget to each timestamp's data within the window of size w for achieving ϵ-LDP. These two methods are not appropriate for allocating budgets because the

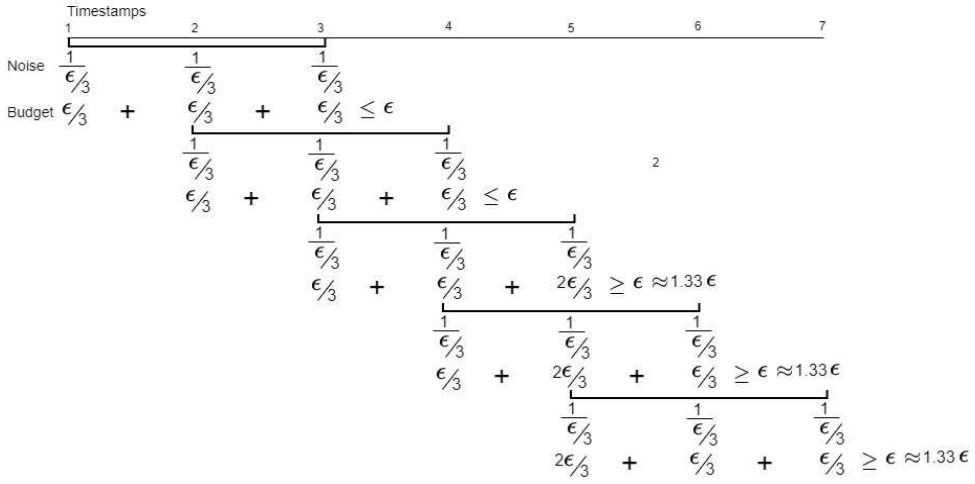

Figure 6.1: Distribution of privacy budget over timestamps(or event) when the user stream's data-points are correlated within the sliding window of size $w = 3$.

correlated location data-point requires more privacy budget than the normal timestamp's data-point. Therefore, it is necessary to design a privacy budget allocation scheme for allocating privacy budgets to the correlated location data-points within the user's stream. The following example illustrates how w-event privacy degrades a privacy guarantee when the user stream's data points are correlated within a sliding window of size w.

Consider a user stream of length 7, say $T_i = \{l_i | l_i \in L, 1 \leq i \leq 7\}$ where L is a universe of location data-points and l_i is a location at timestamp i. Let us assume that a mobility pattern, i.e.,*always reaches a location l_5 after visiting location l_4* which is represented as $P_r(l_5 = loc_5 | l_4 = loc_4) = 1$. According to the w-event privacy, the privacy budget ϵ is shared uniformly to each timestamp's data within the sliding window of size 3, as shown in Figure 6.1. At the first sliding window, the privacy mechanism allots a privacy budget uniformly (i.e., ϵ/3) to each timestamp's data to calculate noise Lap(1/(ϵ/3)). Hence the required ratio of privacy budget at timestamps 1,2, and 3 is ϵ/3. Since the total sum of the privacy budgets within the first sliding window is equal to ϵ, the first sliding window

achieves ϵ-LDP. Similar to the second sliding window at timestamps 2,3, and 4. A third sliding window, the required ratio of privacy budgets at timestamps 3, 4 and 5 is $\epsilon/3$,$\epsilon/3$ and $2\epsilon/3$, respectively, due to the presence of correlation between 4 and 5. Since the sum of the third window's privacy budgets is more than the total privacy budget ϵ, So the third sliding window does not achieve ϵ-LDP. Similarly, the fourth and fifth windows also do not satisfy ϵ-LDP because these windows require more budget than the total privacy budget ϵ.

In summary, the privacy budget allocation into a user stream for publication with privacy guarantees remains unclear, especially when the location data-points of a user stream are correlated. Our contribution of this work is as follows.

1. We present a definition ϵ-Local Differential Privacy for continuous data publication and prove that it can achieve ϵ-LDP. We quantify the privacy degradation when correlation exists in continuous data publication and analyze the privacy leakage with a numerical example.

2. We propose a Privacy Budget Allocation method on ϵ-Local Differential Privacy for distributing an adequate amount of privacy budgets to each timestamp's data under the protection of ϵ-LDP.

3. Finally, we demonstrate the effectiveness of our proposed method in terms of data utility with existing allocation methods by considering real and synthetic datasets.

The rest of this work is as follows. Section 6.1 describes the definition of ϵ-Local Differential Privacy for continuous stream data publication and analyzes the privacy leakage in ϵ-LDP with and without presence of correlation between the location data-points of a user stream. Section 6.2, we propose a privacy budget allocation method for correlated location data-points in user stream publication and present a theoretical analysis of privacy leakage

and utility leakage of our proposed method. We demonstrate the effectiveness of our proposed method in terms of data utility is presented in section 6.3. Finally, the summary of this work is presented in section 6.4.

6.1 System Framework

In this section, we describe the basic definition of Local-Differential privacy (LDP) under continual observation. Also, we illustrate the privacy leakage of LDP under temporal correlation with a numerical example.

6.1.1 Local differential privacy under continual observation

Let $\mathcal{M}$ be the randomized mechanism that takes the user's trajectory $(T_i)^t$ as input and produced a series of perturbed location data-points $(\widehat{x}_i^1, \widehat{x}_i^2, \ldots, \widehat{x}_i^t)$ at each timestamp. Suppose the mechanism $\mathcal{M}$ is decomposed into several sub-mechanisms $(\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_t)$ and each sub-mechanism $\mathcal{M}_t$ takes a location data-point x_i^t and produces a perturbed location data-point $\widehat{x}_i^t$. A sub-mechanism $\mathcal{M}_t$ is satisfied ϵ_t-local differential privacy (LDP) if and only if for any two input location data-points x_i^t and $x_i^{t'}$, and for any possible outputs $\widehat{x}_i^t$ of range($\mathcal{M}_t$), $\mathcal{M}_t$ holds

$$log \frac{P_r(\mathcal{M}_t(x_i^t) = \widehat{x}_i^t)}{P_r(\mathcal{M}_t(x_i^{t'}) = \widehat{x}_i^t)} \leq \epsilon_t \tag{6.1}$$

Where ϵ_t is a privacy parameter that quantifies the degree of user's privacy leakage at timestamp t. According to equation, the curator does not distinguish the location data-point of a user i is whether x_i^t or $x_i^{t'}$ with high confidence by seeing the answer $\widehat{x}_i^t$.

The mechanism $\mathcal{M}$ achieves ϵ-LDP if each user encodes his vector bits of his location data-point (say $x_i^j[l], l = (1, 2, \ldots, L)$) before sending into curator. These perturbed vector bits

of location data-point protects the user's location information from the malicious curator. In this fashion, a user publishes location data-points privately $(\widehat{x}_i^1, \widehat{x}_i^2, \ldots) = \widehat{x}_i$ under the protection LDP at each timestamp. Each location data-point at time t uses independent randomness for achieving ϵ-LDP. So, it holds

$$\begin{aligned} log\frac{P_r(\mathcal{M}((T_i)^t) = \widehat{x}_i^1, \widehat{x}_i^2, \ldots, \widehat{x}_i^t)}{P_r(\mathcal{M}((T_i)^{t'}) = \widehat{x}_i^1, \widehat{x}_i^2, \ldots, \widehat{x}_i^t)} &= \prod_{t=1}^{t} \frac{P_r(\mathcal{M}_t(x_i^t) = \widehat{x}_i^t)}{P_r(\mathcal{M}_t(x_i^{t'}) = \widehat{x}_i^t)} \\ &\leq \prod_{t=1}^{t} e^{\epsilon_t} \leq exp(\sum_{t=1}^{t} \epsilon_t) \\ &\leq e^{\epsilon} \end{aligned}$$

Mechanism $\mathcal{M}$ adopts the local randomize method to perturb each location data-point of a user. RAPPOR is a basic method for local randomizer and is widely used in statistics aggregation. This method follows two steps for achieving $\mathcal{M}$ satisfies ϵ-LDP. The steps are

1. Each bit of x_i^t is flipped into either 0 or 1 according to the following rules

$$\widehat{x}_i^t[l] = \begin{cases} x_i^t[l] & \text{with prob. } 1-p, \\ 0 & \text{with prob. } \frac{1}{2}p, \\ 1 & \text{with prob. } \frac{1}{2}p, \end{cases} \tag{6.2}$$

Where $p \in [0, 1]$ is a parameter that measures the level of randomness for local differential privacy, $\widehat{x}_i^t[l]$ where $l = (1, 2, \ldots, L)$ are the fake bits of vector x_i at timestamp t.

2. After deriving all fake bits of x_i i.e., $\widehat{x}_i^t =< \widehat{x}_i^t[1], \widehat{x}_i^t[2], \ldots, \widehat{x}_i^t[L] >$, then transmit it into the curator.

In this fashion, the curator collects all user location data-points privately, and further, the

curator can use or publish the location data-points for other useful purposes. LDP's main property is that a user can take full control of his/her privacy by independently perturbing location data-points.

6.2 Threat model: privacy leakage analysis

The main objective of an adversary is to learn the user's true location at a specific timestamp. Consider an adversary A_i with knowledge of all location data-points of user i except t^{th} time location data-point i.e., A_i knows at time t is $(T_i)^{t_k} = (T_i)^t \backslash \{x_i^t\}$ and he wants to find the location data-point of i^{th} user at timestamp t. The privacy leakage of privacy mechanism $\mathcal{M}_t$ (or i^{th} user at timestamp t) against A_i is as follows, we have $(T_i)^{t_k} = (T_i)^t \backslash \{x_i^t\}$ and x_i^t and $x_i^{t'}$ are two possible location data-points of i^{th} user at timestamps t. Then the privacy leakage of $\mathcal{M}_t$ is

$$\mathcal{L}_{A_i}(\mathcal{M}_t) = \sup_{x_i^t, x_i^{t'}, \widehat{x}_i^t} log \frac{P_r(\widehat{x}_i | x_i^t, (T_i)^{t_k})}{P_r(\widehat{x}_i | x_i^{t'}, (T_i)^{t_k})}$$

The $\mathcal{L}_{A_i}(\mathcal{M}_t)$ is the maximum privacy leakage of user i at timestamp t caused by adversary A_i. If $\mathcal{L}(\mathcal{M}_t)$ is lesser than or equal to ϵ, we say that $\mathcal{M}_t$ achieves ϵ-LDP. Here, we considered a privacy budget ϵ as a metric for measuring the privacy leakage. If lesser ϵ value, then lesser the privacy leakage.

6.2.1 Temporal correlation (TC) privacy leakage analysis

It is fair to assume that adversaries also have the knowledge of transition probabilities between the location data-points. In our settings, we adopt a Markov chain model to describe the transition probabilities between the data-points and is represented by a variable called Θ. For instance, consider a transition probability matrix of size 2 for a trajectory of user

i is $T_i = (x_i^1 = l_1, x_i^2 = l_2, x_i^3 = l_2)$ as shown in Figure 6.2. If the location of x_i^{t+1} is l_2, then the probability of coming from l_1 at time t (i.e., $x_i^t = l_1$) is 0.4.

		$t+1$	
	x_i	l_1	l_2
t	l_1	0.6	0.4
	l_2	0.1	0.9

Figure 6.2: Transition probabilities between the data-points

Consider an adversary A_i with knowledge of Θ and $(T_i)^{t_k} = (T_i)^t \backslash \{x_i^t\}$, so A_i is named as A_i^Θ. Assume that A_i^Θ collects all perturbed location data points except t^{th} location data point of user i and A_i^Θ wants to infer t^{th} location data point of user i. The privacy leakage analysis with respect to Θ is named as temporal privacy leakage ($\mathcal{TCL}$). The $\mathcal{TCL}$ of LDP mechanism at timestamp t with respect to A_i^Θ is, the maximum ratio of two distribution for all different values of x_i^t and $x_i^{t'}$ and all possible outputs $\hat{x}_i$ of range($\mathcal{M}$), we have

$$\mathcal{TCL}_{A_i^\Theta}(\mathcal{M}_t) = \sup_{\widehat{x}_i, x_i^t, x_i^{t'}, \Theta} log \frac{P_r(\widehat{x}_i^1, ..., \widehat{x}_i^t | x_i^t, (T_i)^{t_k}, \Theta)}{P_r(\widehat{x}_i^1, ..., \widehat{x}_i^t | x_i^{t'}, (T_i)^{t_k}, \Theta)} \tag{6.3}$$

The $\mathcal{TCL}$ of mechanism $\mathcal{M}_t$ w.r.t A_i^Θ satisfies ϵ-local differential privacy iff the above equation 6.3 is bounded by atmost ϵ value for any adversary A_i, $i \in n$ where n is set of adversaries.

$$\sup_{\Theta, \forall A_i, i \in n} \mathcal{TCL}_{A_i^\Theta}(\mathcal{M}_t) \leq \epsilon$$

To understand the impact of temporal correlation on privacy leakage in ϵ-local differential privacy, we expand the above equation 6.3 using the Bayesian theorem. The $\mathcal{TCL}_{A_i^\Theta}(\mathcal{M}_t)$

is equal to

$$= \sup_{\widehat{x}_i^1,.....,\widehat{x}_i^{t-1},x_i^t,x_i^{t'},\Theta} log\frac{\sum_{x_i^{t-1}} P_r(\widehat{x}_i^1,.....,\widehat{x}_i^{t-1}|x_i^{t-1},(T_i)^{t_{k-1}},\Theta) * P_r(x_i^{t-1}|x_i^t)}{\sum_{x_i^{t-1'}} P_r(\widehat{x}_i^1,.....,\widehat{x}_i^{t-1}|x_i^{t-1'},(T_i)^{t_{k-1}},\Theta) * P_r(x_i^{t-1'}|x_i^{t'})} + \sup_{\widehat{x}_i^t,x_i^t,x_i^{t'},\Theta} log\frac{P_r(\widehat{x}_i^t|x_i^t,(T_i)^{t_k},\Theta)}{P_r(\widehat{x}_i^t|x_i^{t'},(T_i)^{t_k},\Theta)} \quad (6.4)$$

The above expression involves three terms: 1) the first term is to evaluate the privacy leakage at previous timestamp (t-1), 2) the second term finds the probability of transition between the data-points of previous timestamp $(t-1)$ and current timestamp (t) and 3) third is to find the privacy leakage at current timestamp t. Hence, we state that the privacy leakage at current timestamp t depends on the privacy leakage at $t-1$, TC transition probability, and the privacy leakage at time t. Informally, we can write that equation 6.4 is

$$\mathcal{TCL}_{A_i^\Theta}(\mathcal{M}_t) = \mathcal{TCL}_{A_i^\Theta}(\mathcal{M}_{t-1}) + \mathcal{L}_{A_i^\Theta}(\mathcal{M}_t) \quad (6.5)$$

According to equation 6.4, the first two terms determine temporal privacy leakage and this can be calculated using linear fractional programing (LFP), discussed in [18]. In detail, let q and d are the two distinct rows of transition probability matrix Θ and α be the privacy parameter that quantifies the level of temporal privacy leakage. According to LFP, the maximum value of objective function is

$$\mathcal{TCL}_{A_i^\Theta}(\mathcal{M}_{t-1}) = \max_{q,d\in\Theta} log\frac{q(e^\alpha-1)+1}{d(e^\alpha-1)+1} \quad (6.6)$$

In third term of equation 6.4, $\widehat{x}_i^t$ is conditionally independent of $u_i^{t_k}$ if given x_i^t. So, the

third term of equation 6.4 is as follows.

$$\mathcal{L}_{A_i^{\Theta}}(\mathcal{M}_t) = \sup_{\widehat{x}_i^t, x_i^t, x_i^{t'}} log \frac{P_r(\widehat{x}_i^t | x_i^t)}{P_r(\widehat{x}_i^t | x_i^{t'})}$$

Based on the steps of local randomizer, the relevant probabilities are

$$P_r(\widehat{x}_i^t[l] = 1 | x_i^t[l] = 1) = \frac{1}{2}p + 1 - p = 1 - \frac{1}{2}p \quad P_r(\widehat{x}_i^t[l] = 0 | x_i^t[l] = 1) = \frac{1}{2}p$$

$$P_r(\widehat{x}_i^t[l] = 1 | x_i^t[l] = 0) = \frac{1}{2}p \quad P_r(\widehat{x}_i^t[l] = 0 | x_i^t[l] = 0) = \frac{1}{2}p + 1 - p = 1 - \frac{1}{2}p$$

To satisfy the differential privacy condition, the ratio two conditional probabilities with distinct values of x_i^t, x and x' i.e., $\frac{P_r(\widehat{x}_i^t = \widehat{x} | x_i^t = x)}{P_r(\widehat{x}_i^t = \widehat{x} | x_i^{t'} = x')}$ should be bounded by e^{ϵ}.

$$\frac{P_r(\widehat{x}_i^t = \widehat{x} | x_i^t = x)}{P_r(\widehat{x}_i^t = \widehat{x} | x_i^{t'} = x')} = \frac{\sum_{\widehat{x} \in \{0,1\}} P_r(\widehat{x}_i^t = \widehat{x} | x_i^t = x)}{\sum_{\widehat{x} \in \{0,1\}} P_r(\widehat{x}_i^t = \widehat{x} | x_i^{t'} = x')}$$

Let $a, b \geq 0$ and $c, d \geq 0 : \frac{a+b}{c+d} \leq max(\frac{a}{c}, \frac{b}{d})$, the above equation is

$$\begin{aligned}
&\leq \max_{\widehat{x} \in \{0,1\}} \frac{P_r(\widehat{x}_i^t = \widehat{x} | x_i^t = x)}{P_r(\widehat{x}_i^t = \widehat{x} | x_i^{t'} = x')} \\
&= \frac{\left(\frac{1}{2}p\right)^{\widehat{x}_1} \left(1 - \frac{1}{2}p\right)^{1-\widehat{x}_1} \times \ldots \times \left(\frac{1}{2}p\right)^{\widehat{x}_l} \left(1 - \frac{1}{2}p\right)^{1-\widehat{x}_l}}{\left(\frac{1}{2}p\right)^{1-\widehat{x}_1} \left(1 - \frac{1}{2}p\right)^{\widehat{x}_1} \times \ldots \times \left(\frac{1}{2}p\right)^{1-\widehat{x}_l} \left(1 - \frac{1}{2}p\right)^{\widehat{x}_l}} \\
&\frac{\times \left(1 - \frac{1}{2}p\right)^{\widehat{x}_{l+1}} \left(\frac{1}{2}p\right)^{1-\widehat{x}_{l+1}} \times \ldots \times \left(1 - \frac{1}{2}p\right)^{\widehat{x}_L} \left(\frac{1}{2}p\right)^{1-\widehat{x}_L}}{\times \left(1 - \frac{1}{2}p\right)^{\widehat{x}_{l+1}} \left(\frac{1}{2}p\right)^{1-\widehat{x}_{l+1}} \times \ldots \times \left(1 - \frac{1}{2}p\right)^{\widehat{x}_L} \left(\frac{1}{2}p\right)^{1-\widehat{x}_L}} \\
&= \left(\frac{1}{2}p\right)^{2(\widehat{x}_1 + \widehat{x}_2 + \ldots + \widehat{x}_l - \widehat{x}_{l+1} - \ldots - \widehat{x}_L)} \\
&\times \left(1 - \frac{1}{2}p\right)^{2(\widehat{x}_{l+1} + \widehat{x}_{l+2} + \ldots + \widehat{x}_L - \widehat{x}_1 - \ldots - \widehat{x}_l)}
\end{aligned}$$

The sensitivity is maximized when $\widehat{x}_{l+1} = \widehat{x}_{l+2} = \ldots = \widehat{x}_L = 1$ and $\widehat{x}_1 = \widehat{x}_2 = \ldots = \widehat{x}_l = 0$, then

$$= \left(\frac{1 - \frac{1}{2}p}{\frac{1}{2}p}\right)^{2h}$$

$$\epsilon = 2hln\left(\frac{1 - \frac{1}{2}p}{\frac{1}{2}p}\right)$$

The local randomizer method satisfies ϵ-local differential privacy where $\epsilon = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$. Therefore the total privacy leakage under temporal correlation in continuous data release is as follows.

$$\begin{aligned}\mathcal{TCL}_{A_i^{\Theta}}(\mathcal{M}_t) &= \mathcal{TCL}_{A_i^{\Theta}}(\mathcal{M}_{t-1}) + \mathcal{L}_{A_i^{\Theta}}(\mathcal{M}_t) \\ &= \max_{q,d \in \Theta} log\frac{q(e^{\alpha}-1)+1}{d(e^{\alpha}-1)+1} + 2hln\left(\frac{1 - \frac{1}{2}p}{\frac{1}{2}p}\right)\end{aligned}$$

We illustrate how the TC factor influences privacy leakage w.r.t adversary with and without knowledge of transition probability distribution Θ through a numerical example. Consider a query to find the user $i's$ location value l_i^t is either l_1 or l_2 at timestamp t, where l_1 and $l_2 \in L$. For simplicity, we assume that l_i^t is,

$$l_i^t = \begin{cases} 1 & \text{i's true location at time t,} \\ 0 & \text{Otherwise.} \end{cases} \tag{6.7}$$

Consider a user $i's$ trajectory of length 2 timestamps, say $T_i = (x_i^1 = l_1, x_i^2 = l_2)$ and the transition probability matrix Θ for user $i's$ trajectory as shown in Figure 6.2. Let A_i^{Θ} and A_i are the two adversaries with and without knowledge of Θ distribution respectively and are interested in finding the location of U_i at timestamp 2 (i.e., the location value of x_i^2).

According to equation 6.5, we compute $\mathcal{TCL}_{A_i}(\mathcal{M}_2)$ and $\mathcal{TCL}_{A_i^\theta}(\mathcal{M}_2)$. For adversary A_i without knowledge of Θ, we get

$$\begin{aligned}\mathcal{TCL}_{A_i}(\mathcal{M}_2) &= \sup_{\widehat{x}_i^1,\widehat{x}_i^2} log\frac{P_r(\widehat{x}_i^1,\widehat{x}_i^2|x_i^2=l_2)}{P_r(\widehat{x}_i^1,\widehat{x}_i^2|x_i^{2'}=l_1)}\\ &= \sup_{\widehat{x}_i^1} log\frac{exp(-|\widehat{x}_i^1-(x_i^1=l_1)|)P_r(x_i^1|x_i^2=l_2)}{exp(-|\widehat{x}_i^1-(x_i^{1'}=l_2)|)P_r(x_i^{1'}|x_i^{2'}=l_1)}\\ &+ \sup_{\widehat{x}_i^2} log\frac{exp(-|\widehat{x}_i^2-(x_i^2=l_2)|)}{exp(-|\widehat{x}_i^2-(x_i^{2'}=l_1)|)}\\ &= 0+\sup_{\widehat{x}_i^2} log\frac{exp(-|\widehat{x}_i^2-1|)}{exp(-|\widehat{x}_i^2-0|)}=1\end{aligned}$$

For adversary A_i with knowledge of Θ (i.e., A_i^Θ), we get

$$\begin{aligned}\mathcal{TCL}_{A_i^\Theta}(\mathcal{M}_2) &= \sup_{\widehat{x}_i^1,\widehat{x}_i^2} log\frac{P_r(\widehat{x}_i^1,\widehat{x}_i^2|x_i^2=l_2)}{P_r(\widehat{x}_i^1,\widehat{x}_i^2|x_i^{2'}=l_1)}\\ &= \sup_{\widehat{x}_i^1} log\frac{exp(-|\widehat{x}_i^1-(x_i^1=l_1)|)P_r(x_i^1|x_i^2=l_2)}{exp(-|\widehat{x}_i^1-(x_i^{1'}=l_2)|)P_r(x_i^{1'}|x_i^{2'}=l_1)}\\ &+ \sup_{\widehat{x}_i^2} log\frac{exp(-|\widehat{x}_i^2-(x_i^2=l_2)|)}{exp(-|\widehat{x}_i^2-(x_i^{2'}=l_1)|)}\\ &= 0.55+1=1.55\end{aligned}$$

The above privacy analysis shows that privacy leakage is increased when the user's location data-points are temporally correlated. Hence, we can claim that the curator (or data publisher) has more chance to disclose sensitive data (or location data-point) than traditional ϵ-LDP in continuous data publications settings. It happens due to an inadequate supply of privacy budget for perturbing the temporally correlated location data-points of a user stream. A recent privacy method called w-event privacy for allocating a ratio of privacy budget to each location data-point to achieve ϵ-LDP guarantee of any user's stream

of length w. However, this privacy method achieves ϵ-LDP by assuming the location data-points of a user stream are independent. In real-time data collection, most of the location data-points are temporally correlated of certain probability. Due to this, w-event privacy has more privacy leakage as compared to traditional ϵ-LDP.

6.3 The Proposed Algorithm

In this section, we discuss our proposed Privacy Budget Allocation (PBA) mechanism in local settings. It allows to compute and allocate the quantity of privacy budget to each publication in continuous data release settings. Further, we theoretically prove that our PBA mechanism achieves ϵ-LDP and shows the data utility of our PBA mechanism.

Let $\mathcal{M}$ be the PBA mechanism, which consists of series of sub-mechanisms, say $\mathcal{M}_1$, $\mathcal{M}_2$, . . . , $\mathcal{M}_k$, . . . , $\mathcal{M}_t$. Each sub-mechanism $\mathcal{M}_k$ takes a location data-point of a user stream as an input and generates a noisy location data-point $\widehat{x}_i^k$ as an output. Each sub-mechanism $\mathcal{M}_k$ uses some amount of privacy budget (say ϵ_k) from total privacy budget ϵ for perturbing K^{th} location data-point to achieve ϵ_k-LDP at sub-mechanism $\mathcal{M}_k$. In this fashion, $\mathcal{M}$ publishes a series of private location data-points of a user stream namely $(\widehat{x}_i^1, \widehat{x}_i^2, \ldots, \widehat{x}_i^k, \ldots, \widehat{x}_i^t)$ to the data curator or the data collector.

In our PBA mechanism $\mathcal{M}$, we adopt a sliding window methodology used in the w-event privacy model. This privacy model promises to protect any sequence of location data-points occurring within a sliding window of size w. If $\mathcal{M}$ to satisfy w-event privacy, it ensures the following condition. Let us assume that the sliding window of size w consists of w number of timestamps or location data-points and the span of any timestamp k within the sliding window is from $k - w + 1$ to k. Each timestamp k is operated by a respective sub-mechanism $\mathcal{M}_k$. Since $\mathcal{M}_k$ uses independent randomness, so $\mathcal{M}_k$ achieves ϵ_k-LDP

for some ϵ_k. The sum of privacy budgets used for randomness by each sub-mechanism within the sliding window must be less than or equal to the total privacy budget ϵ. Then the PBA mechanism $\mathcal{M}$ achieves w-event privacy. Moreover, the sliding window is moving one timestamp ahead after every w timestamps. Hence, mechanism $\mathcal{M}$ achieves user-level privacy of any length of the user's stream.

The PBA mechanism $\mathcal{M}$ within the sliding window involves two phases. These two phases operate sequentially by using independent randomness. At any timestamp k within the window, the first phase of the $\mathcal{M}$ compute dissimilarity between current location data-point (say x_i^k) and last release noisy data-point from the noisy data-points of timestamp 1 to $k-1$ i.e., $(\widehat{x}_i^l, \widehat{x}_i^2, \ldots, \widehat{x}_i^l, \ldots, \widehat{x}_i^{k-1})$. The dissimilarity value is made privately by using the allotted privacy budget at k^{th} timestamp, and this private dissimilarity value is forwarded into phase 2 of $\mathcal{M}$. The second phase of $\mathcal{M}$ uses it to decide whether to publish the current location data-point or not. If the second phase of $\mathcal{M}$ decides not to publish the current location data-point at k (named as skipped publication), then the k^{th} allotted privacy budget becomes free and can be used for future publication if necessary. On the other hand, $\mathcal{M}$ decides to publish the current location data-point at k, them $\mathcal{M}$ absorbs one extra privacy budget that became available from previous skipped publication if and only if the correlation exists between the current location data-point and last published location data-point; otherwise, it uses only allotted privacy budget for publication at k^{th} timestamp. The quantity of extra privacy budget is the same as the privacy budget allotted at the respective timestamp. Note that whenever the mechanism absorbs an extra privacy budget from previous timestamps, then the mechanism must nullify immediately succeeding timestamp's publication, i.e., the result of $\widehat{x}_i^{k+1} = null$ if k^{th} timestamp uses an extra budget. This null publication's privacy budget can not be used in any future publication, because it exceeds the maximum privacy budget ϵ when the sliding window slides over

Algorithm 6.1 Pseudocode of $\mathcal{M}_k$ in PBA

INPUT: Data-point x_i^k, total privacy budget ϵ and sliding window size w.
OUTPUT: Release noisy data-point $\widehat{x}_i^k$.

At sub-mechanism $\mathcal{M}_k$
  Phase1: Compare last release noisy data-point $\widehat{x}_i^l$ and current data-point x_i^k
    Compute x_i^k and retrieve $\widehat{x}_i^l$
    Calculate dissimilarity $d(x_i^k, \widehat{x}_i^l) = \frac{1}{|L|}\Sigma_{j=1}^{|L|}|\widehat{x}_i^l[j] - x_i^k[j]|$
    Calculate privacy budget $\epsilon_k^1 = \epsilon/(2 \cdot w) = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$, where $p \in [0, 1]$
    Convert $n_i^k = d(x_i^k, \widehat{x}_i^l)$ into $\widehat{n}_i^k$ using Local Randomizer.
  Phase2: Compute noise for current data-point x_i^k
    Calculate privacy budget at k^{th} timestamp $\epsilon_k^2 = \epsilon/(2 \cdot w) = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$, where $p \in [0, 1]$
    Compute $N_{nullP} = \frac{\epsilon_l^2}{\epsilon/(2 \cdot w)} - 1$
    If $k - l \leq N_{nullP}$
      $x_i^k = null$
    Else
      Find $N_{AbsorbedP} = k - (l + N_{nullP})$
      $\epsilon_{extra} = \epsilon/(2 \cdot w) * N_{AbsorbedP} - 1$
      $\epsilon_{total} = \epsilon_{total} + \epsilon_{extra}$
      If correlation exists
        $\epsilon_{total} = \epsilon_{total} - \epsilon/(2 \cdot w)$
        $\epsilon_k^2 = 2 \cdot \epsilon/(2 \cdot w)$
      Compute noise $\lambda_k^2 = 1/\epsilon_k^2$
      If $\widehat{d}_i^k > \lambda_k^2$
        Return $\widehat{x}_i^k$
      Else
        Return $\widehat{x}_i^l$
  end for
end for

time. Although, any window utilizes some extra privacy budgets for publications if and only if the window contains an equal or more number of skipped publications occurred. However, both phases of $\mathcal{M}$ must be made private even-though phase 1 of $\mathcal{M}$ appears within the internal process of the PBA mechanism. This is because the adversary knows that phase 1 computes dissimilarity value, and this value affects the decision taken by phase 2 of $\mathcal{M}$. Hence, phase 1 of $\mathcal{M}$ must add proper noise to dissimilarity value.

Algorithm 6.1 shows the pseudocode for allocating a privacy budget at k^{th} timestamp

within the sliding window of size w by using the PBA method. A sub-mechanism $\mathcal{M}_k$ takes a k^{th} location data-point as an input and generates a noisy location data-point $\widehat{x}_i^k$ as an output. As we know that the PBA mechanism $\mathcal{M}$ involves two phases, namely $\mathcal{M}^1$ and $\mathcal{M}^2$. These two phases operate sequentially by using half of the total privacy budget, such as $\epsilon^1 = \epsilon/2$ and $\epsilon^2 = \epsilon/2$, respectively. In the first phase of $\mathcal{M}$, $\mathcal{M}^1$ uniformly allocates a ratio of privacy budget from ϵ^1 to each timestamp within a sliding window. At timestamp k, $\mathcal{M}_k^1$ computes the current location data-point (x_i^k) and retrieves the last release noisy location data-point ($\widehat{x}_i^l$) (line 3). Then it calculates a dissimilarity value between x_i^k and $\widehat{x}_i^l$. To find dissimilarity value, we use a metric called mean of absolute error (MAE) and it is formulated as $\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\widehat{x}_i^l[j] - x_i^k[j]|$, where $x_i^k[j]$ and $\widehat{x}_i^l[j]$ are the vectors of length $j = 1, 2, \ldots, |L|$ (line 4). After, the obtained dissimilarity value $d(x_i^k, \widehat{x}_i^l)$ is made in private using local randomizer i.e., $\widehat{n}_i^k$ (lines 5-6), and forward it into the sub-mechanism $\mathcal{M}_k^2$. At phase 2, $\mathcal{M}^2$ uniformly allocates a ratio of privacy budget from ϵ^2 to each timestamp within a sliding window (line 8). At timestamp k, $\mathcal{M}_k^2$ computes the number of null publications' privacy budget absorbed by the last release publication. So that the mechanism must nullify the same number of immediately succeeding timestamp's publications. At k, if the last release publication absorbs one extra privacy budget, then k^{th} timestamp must be null (lines 9-11). Otherwise, find the number of skipped publications before k^{th} timestamp and add their privacy budgets into the absorbed privacy budget variable named as ϵ^A (lines 13-15). If a correlation exists between the k^{th} timestamp and last release publication, then k^{th} timestamp requires an extra budget from ϵ^A and add into the allotted privacy budget at timestamp k ϵ_k^2 (lines 16-18). Then, $\mathcal{M}_k^2$ compute noise λ_k^2 for k^{th} publication, if the private dissimilarity value at phase 1 is greater than noise λ_k^2 computed for k^{th} publication at phase 2, then $\mathcal{M}_k$ publishes x_i^k with noise or otherwise last release noisy location data-point $\widehat{x}_i^l$.

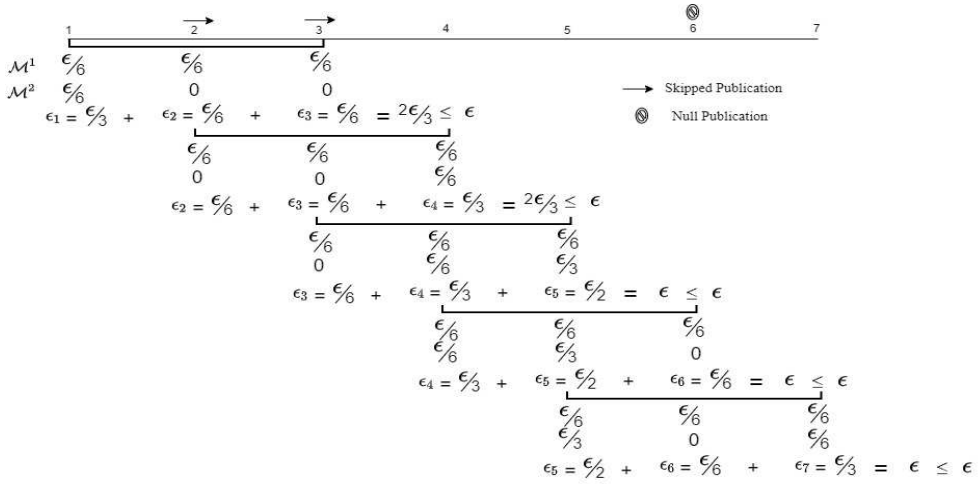

Figure 6.3: Distribution of privacy budget over timestamps(or event) within the sliding window of size $w = 3$.

Figure 6.3 shows that the PBA method allocates the privacy budget into each timestamp within the sliding window of size 3 while assuming the user trajectory length is 7. The PBA method allocates $\epsilon/(2.w) = \epsilon/6$ privacy budget for both sub-mechanism $\mathcal{M}_k^1$ and $\mathcal{M}_k^2$. At the first timestamp, PBA publishes the private location data point by using $\mathcal{M}_1^2 = \epsilon/6$. Suppose at timestamp 2 and 3, PBA decides not to publish private location data points i.e., the output of timestamps 2 and 3 is null, and the outputs are approximated by $\widehat{x}_1^1$. The budgets of timestamps 2 and 3 become available for future publications i.e., $\epsilon/6 + \epsilon/6 = \epsilon/3$. At timestamp 4, PBA publishes a private location data point using allotted privacy budget $\mathcal{M}_4^2 = \epsilon/6$at time 4. At time 5, the mechanism decides to publish a private location data point, but the location value at time 5 correlates with the location value at time 4. Thus $\mathcal{M}_5^2$ uses its own privacy budget and an extra privacy budget that became available from the previous skipped publication (say timestamp 2 and 3). Since $\mathcal{M}_5^2$ uses an extra privacy budget, the publication at timestamp 6 must be null (i.e., $\mathcal{M}_6^2 = 0$). Hence the sub-mechanism $\mathcal{M}_6$ outputs null. Finally, at timestamp 7, the mechanism decides to publish a

private location data point using the allotted privacy budget $\mathcal{M}_7^2 = \epsilon/6$. For any window, the sum of the privacy budgets is atmost ϵ. Note that if PBA mechanism does not make null at timestamp 6, the sum of the privacy budgets in the sliding window (of timestamp 4-6) is $7\epsilon/6$, which violates ϵ-LDP.

6.3.1 Privacy Analysis

Theorem 6.3.1. *A privacy budget allocation(PBA) method satisfies w-event local differential privacy.*

Proof. Any sliding window of size w, the mechanism $\mathcal{M}$ assigns half of the privacy budget into phase1 and the remaining into phase2 of PBA method. Therefore, we should prove that, at any timestamp k in w, $\mathcal{M}_k^1$ satisfies ϵ_k^1-LDP where $\epsilon_k^1 = \epsilon/2w = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$ and $\mathcal{M}_k^2$ satisfies ϵ_k^2-LDP where ϵ_k^2 is atmost $2 * 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$ if publication occurs at timestamp k, otherwise ϵ_k^2 is zero. Here, ϵ_k^2 is calculated according to the following situation such as, if statistics at timestamp k involves correlation then ϵ_k^2 requires $2 * 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$ privacy budget, else otherwise requires $2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$. In phase1, $\mathcal{M}_k^1$ publishes a dissimilarity value i.e., $q(d) = d(x_i^k, \widehat{x}_i^l) = |\widehat{x}_i^l[j] - x_i^k[j]|, j = (1, 2, ..., L)$ in privately. Since the one bit of vector x_i^k is one, the maximum difference of $q(d)$ and $q(d')$ is two bits. Hence, the sensitivity of q is 2. Then, the noise of $\mathcal{M}_k^1$ is $\lambda_k^1 = 2/2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$. According to the definition LDP, $\mathcal{M}_k^1$ satisfies ϵ_k^1-LDP where $\epsilon_k^1 = \frac{2}{2/2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)} = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right) = \epsilon/2w$ when $p = \frac{2}{e^{\frac{\epsilon}{4wh}+1}}$. Moreover, the sum of the privacy budget used by each $\mathcal{M}_k^1$ within the window of size w is equal to $\epsilon/2$. Therefore, we next prove that each $\mathcal{M}_k^2$ falls on $0 \leq \Sigma_{k=i}^{w+i-1}\epsilon_k^2 \leq \epsilon/2$. In phase2, $\mathcal{M}_k^2$ uses ϵ_k^2 budget, which quantifies is purely based on statistic value at time k. Suppose k be the timestamp which absorbs extra budget(E_b) of quantity $E_b = 1$, due to the existence of data correlation. According to the PBA method, the quantity of extra budget at k^{th} timestamp totally depends on the number of null pub-

lications $(N_{nullP}) \geq E_b$ appears before k^{th} timestamp within the window and the same E_b quantity of publication should be nullified after k^{th} timestamp. Then, the sum of the privacy budget of k along with N_{nullP} publications is atmost $(E_b + 1) * \epsilon/2w$ i.e., each of these $N_{nullP} + 1$ publications receives privacy budget $\frac{(E_b+1)*\epsilon/2w}{N_{nullP}+1} \leq \epsilon/2w$. This applies any timestamp k' which receives extra budget within the window of size w. Therefore, $\Sigma_{k=i}^{w+i-1}\epsilon_k^2 = \Sigma_{k=i}^{w+i-1}\epsilon/2w \leq \epsilon/2$. Hence, $\mathcal{M}$ satisfies ϵ-LDP where $\epsilon = \epsilon^1 + \epsilon^2$.

Theorem 6.3.2. *The average error per timestamp in PBA is at most* $\frac{1}{2E_b+1} \cdot (\frac{2w}{(E_b+1)\epsilon} + E_b \cdot error_{null}) + \frac{2w}{\epsilon|L|}$

Proof. At timestamp k, the private dissimilarity value at $\mathcal{M}_k^1$ guides $\mathcal{M}_k^2$ to decide whether to publish either true publication or null publication. Hence, we consider both sub-mechanism's errors to compute the average error per timestamp in PBA. The $\mathcal{M}_k^1$ induces error when its private dissimilarity value suggests $\mathcal{M}_k^2$ to make a wrong decision, i.e., $\mathcal{M}_k^2$ performs wrongly skips a publication or wrongly performs the true publication. If $\mathcal{M}_k^1$ made the correct decision (i.e., true publication occurs at time k), then the error at timestamp k is the error induced by $\mathcal{M}_k^2$. Or if the publication is correctly skipped at time k, the error at timestamp k is an original dissimilarity value between x_i^k and $\widehat{x}_i^l$, which is bounded by the error of $\mathcal{M}_k^2$. However, if $\mathcal{M}_k^1$ wrongly performs true publication at time k, then the original dissimilarity value is underrated due to noise of scale λ_k^1 added by the $\mathcal{M}_k^1$. Or if $\mathcal{M}_k^1$ wrongly skips a publication at time k, then the original dissimilarity value is overrated due to noise with scale λ_k^1 added by the $\mathcal{M}_k^1$. The expected under/overrated of dissimilarity is equal $\frac{2w}{\epsilon|L|}$ due to the noise of $\mathcal{M}_k^1$. Therefore, the error induced by the $\mathcal{M}_k^1$ in PBA is at most $\frac{2w}{\epsilon|L|}$.

At $\mathcal{M}_k^2$ phase, if publication occurs, then the publication receives privacy budget provided by the $\mathcal{M}_k^2$ and also it may associated with the same amount of extra budget extracted from previous null publication. So the k^{th} publication receives privacy budget $\frac{(E_b+1)\epsilon}{2w}$.

Hence, it's error is $\frac{2w}{(E_b+1)\epsilon}$. As we know that, the immediate succeeding timestamp of k^{th} timestamp must be null publication and it introduces error $error_{null}$. For calculate error at k^{th} timestamp, we consider both and averaging it by $2E_b + 1$. The average error per timestamp of $\mathcal{M}_k^2$ in PBA is $\frac{1}{2E_b+1} \cdot (\frac{2w}{(E_b+1)\epsilon} + E_b \cdot error_{null})$. Adding error induced by $\mathcal{M}_k^1$, we get average error per timestamp in PBA is $\frac{1}{2E_b+1} \cdot (\frac{2w}{(E_b+1)\epsilon} + E_b \cdot error_{null}) + \frac{2w}{\epsilon|L|}$

6.4 Experimental Results

In this section, we start with an experiment to analyze the impact of different types of correlation on privacy leakage. Then, we conduct an experiment to demonstrate the data utility of our proposed method with the existing states of art methods.

We employed three real data sets and one synthetic dataset in our experiment for measuring the effectiveness of our proposed PBA method with the existing allocation methods. A $Gowalla$ dataset is used in this experiment along with $Geolife$, $T-Drive$ and $Metro100K$ datasets which are described in chapter 3. A $Gowalla$ dataset is real-time dataset, is consists of more than 600000 users' check-in history from November 2010 to December 2011. We optimized the real-time datasets in our experiment by considering a user is located at most one location at each timestamp and collected all samples (user's location data-point) every 5 minutes interval.

6.4.1 Impact of correlation on privacy leakage

We present the behavior of different types of correlation and their impact on privacy leakage of our proposed method.

6.4.1.1 Correlation Behavior

According to our problem settings, we train the Markov model for modeling a transition probability between all possible location data-points. This transition probability matrix describes how a location data-point is dependent on other possible remaining location data-points. There are three types of correlation in the transition probability matrix: strong, moderate, and no correlation. Now, we analyze the privacy leakage under the protection of traditional ϵ-LDP when the dataset involves correlation. Figure 6.4 shows that the privacy leakage of ϵ-LDP under different types of correlation. Let assume that a strong correlation occurs in the transition probability matrix, say $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, then the privacy leakage of ϵ-LDP increases linearly due to the similar kind of location data-point is released in all timestamps, shown in Figure 6.4(a). If a moderate correlation occurs in-between location data-points, say $\begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$, then the privacy leakage of ϵ-LDP from timestamp 1 to t, as shown in Table 6.4(b). The privacy leakage under moderate correlation is quantified by using equation 6. Finally, there is no correlation between the data-points, then each timestamp's data-point achieves 0.1-LDP while assuming ϵ=0.1, as shown in Figure 6.4(c).

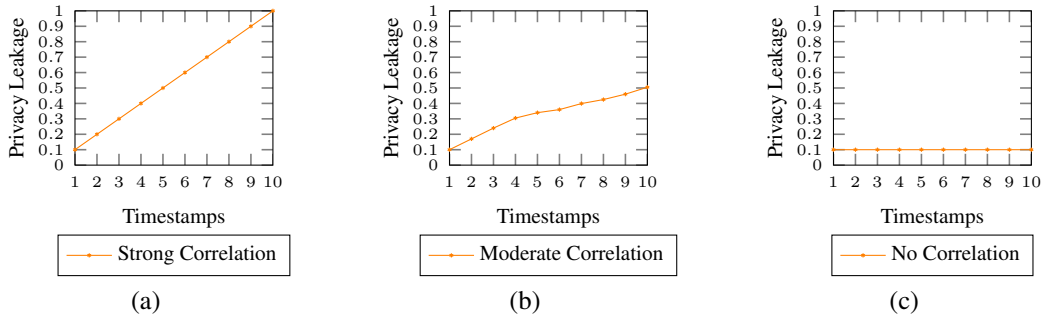

Figure 6.4: Privacy leakage vs Timestamps under different types of correlation (a) Strong correlation (b) Moderate correlation (c) No correlation

6.4.1.2 Privacy leakage under different degrees of correlation

Let assume that a transition probability matrix with a moderate correlation in which the probability value on cells is scattered or uniformizes the probabilities by performing Laplacian smoothing. Let d be the number of dimensions in the transition matrix. If d is large, then the probability value on cells is well scattered. The figure shows the degradation of privacy leakage when the size of d varies in the transition probability matrix.

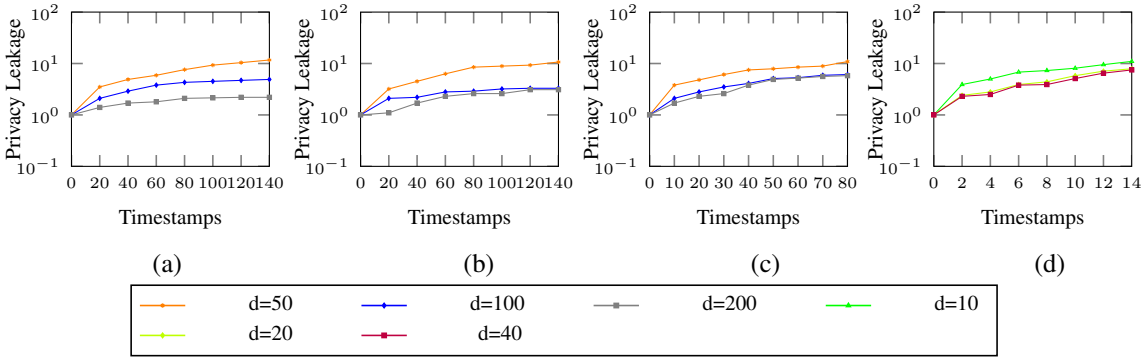

Figure 6.5: Privacy leakage vs different degrees of correlation while set $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

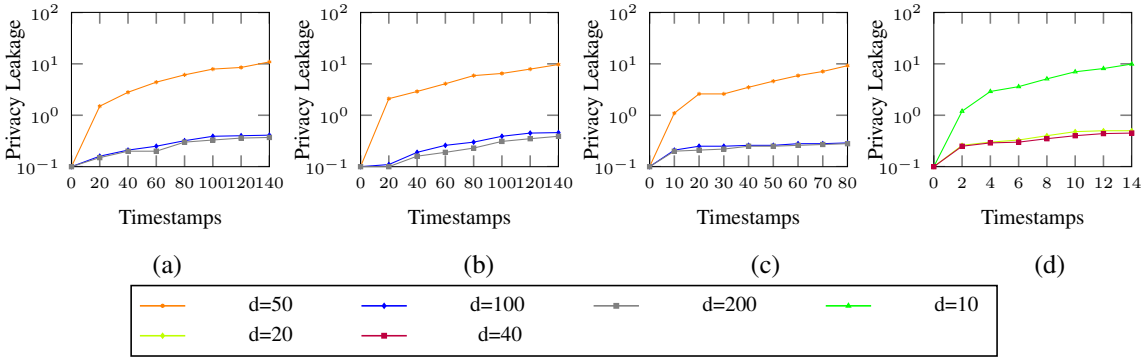

Figure 6.6: Privacy leakage vs different degrees of correlation while set $\epsilon = 0.1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

Privacy leakage versus d: The result of the privacy leakage increases when the size of d

decreases, as shown in the lines $d = 50$ and $d = 10$ of Figures 6.5 and 6.6. This is because the data points in a matrix are very close to the stronger correlation. In other words, a stronger correlation in the transition matrix results in more privacy leakage. The transition matrix involves a weaker correlation when the matrix dimension is larger, as shown in the lines $d = 100$, $d = 200$, $d = 20$, and $d = 40$ of Figures Figures 6.5 and 6.6.

Privacy leakage versus ϵ: We found that the growth of privacy leakage is significantly detained in the case of $\epsilon = 0.1$. For example, taking $d = 100$, the growth of privacy leakage increases, and it continues for approximately 80 timestamps when $\epsilon = 1$ (Fig 6.5). In contrast, it continues for almost 70 timestamps when $\epsilon = 0.1$ (Fig 6.6). However, after an over of timestamps, the privacy leakage in $\epsilon = 0.1$ is not comparatively lower than that of $\epsilon = 1$ under strong correlation. This is because, even though a small privacy budget eliminates privacy leakage at each timestamp, the adversary can study sufficient information from the continuous data releases.

6.4.2 Utility Evaluation

To evaluate publishing statistic's data utility, we used two metrics such as Mean of absolute error(MAE) and Mean of square error(MSE). These two metrics measure dissimilarity (or error) value between the published statistics and true statistics. Moreover, the MSE metric helps to find larger errors. According to our PBA method, each timestamp acquires a piece of the privacy budget to publish statistics under the protection of ϵ-LDP. Thus, we measure dissimilarity (or error) values per timestamp using MAE and MSE metrics. The definition of MAE and MSE metric is as follows.

$$MAE() = \frac{1}{|T|}\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\widehat{x}_i^l[j] - x_i^k[j]|$$

$$MSE() = \frac{1}{|T|}\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\widehat{x}_i^l[j] - x_i^k[j]|^2$$

Where x_i^k and $\widehat{x}_i^l$ are the location data-point of user i at time k and perturbed location data-point of user i at time l (i.e., last release data-point) respectively. And both data-points are the vectors of length L. T be the total number of location data-points of user i.

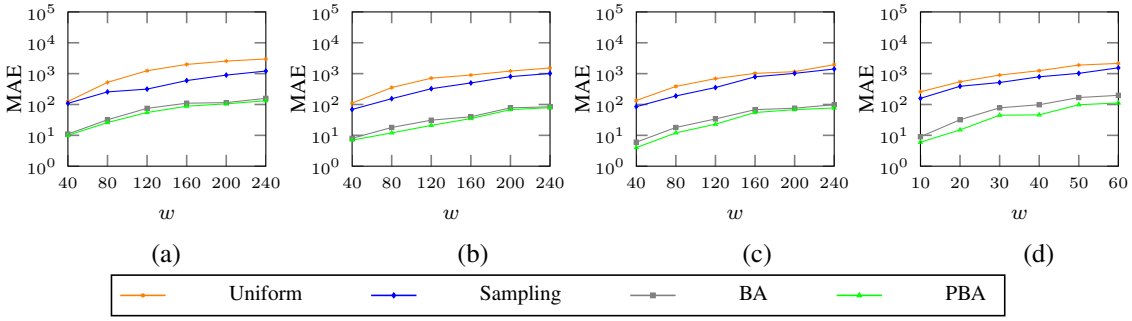

Figure 6.7: MAE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

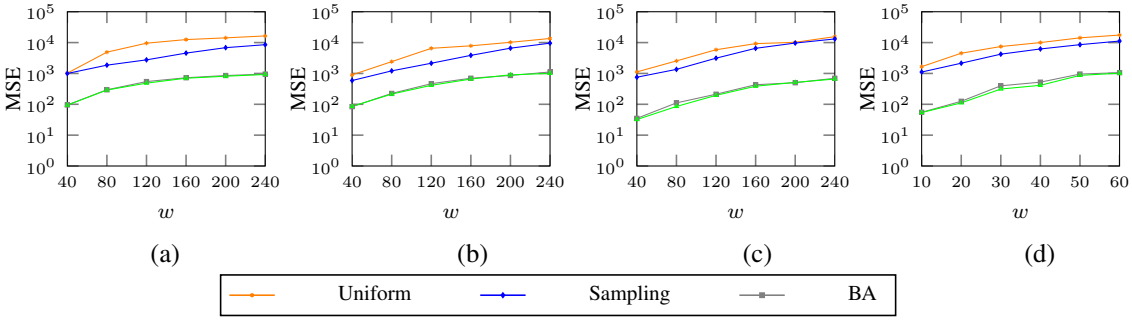

Figure 6.8: MSE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

There exist a few baseline methods for allocating privacy budgets to each timestamp in literature, such as Uniform, Sampling, and Budget Absorption (BA). These methods allocate privacy budgets using different strategies under the protection of ϵ-LDP. We compare our proposed method with the above baseline approaches to analyze the effectiveness of

our method while varying the size of the sliding window w and ϵ value. Figures 6.7 and 6.8 show the error rate of MAE and MSE between the PBA method and the baseline approaches while varying the sliding window size w. We observed that the error rate of MAE and MSE increases while increasing the sliding window size w. Because the rate of allotted budget to each timestamp is minimized when the large size of the sliding window. However, the PBA method's error rate is comparatively less as compared to the baseline approaches. This is because the adequate amount of privacy budget is allotted at temporally correlated timestamps in the PBA method compared to baseline approaches. The error rate of MAE and MSE between the PBA and BA methods are approximately similar because both methods follow the same privacy budget allocation strategy (i.e., sliding window methodology). However, the BA method allows non-temporally correlated user stream, while the PBA method allows temporally correlated user stream and proves that the PBA method achieves ϵ-LDP under temporal correlation. Notice that the Uniform method's error rate increases linearly when w increases because of a fixed privacy budget allotted to each timestamp, which leads to more error rates. Similarly, a sampling approach allots privacy budget at a given sample interval of the entire user stream.

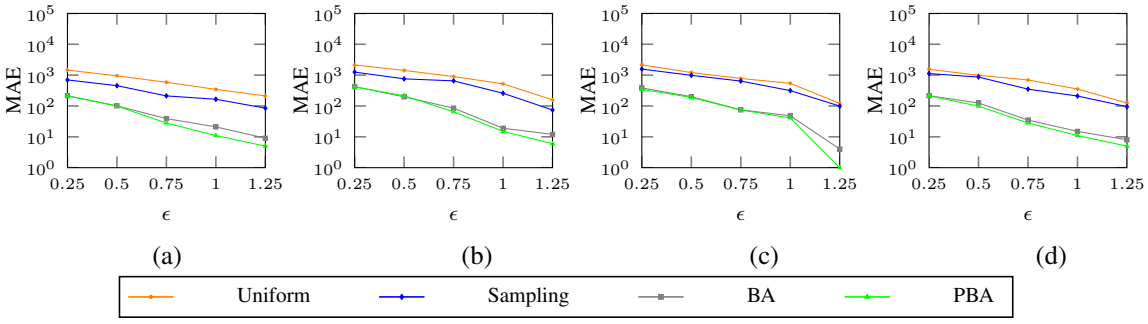

Figure 6.9: MAE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

Further, we compared our proposed method with the above baseline approaches while

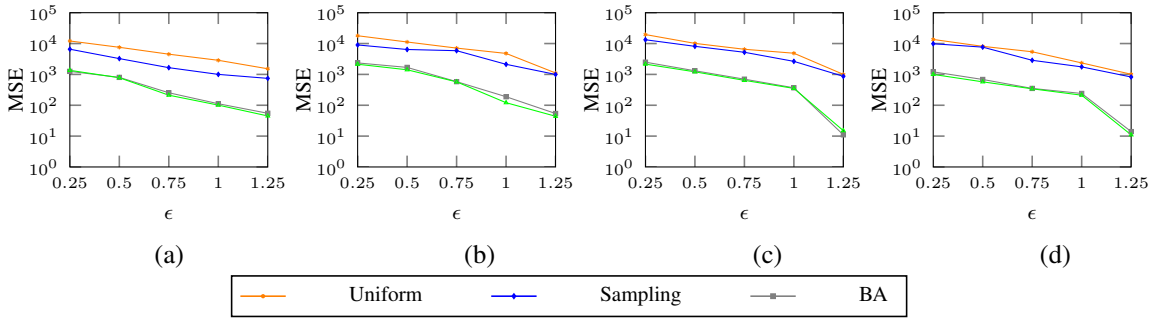

Figure 6.10: MSE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

varying ϵ values, as shown in Figures 6.9 and 6.10. We observed that the error rate of MAE and MSE decreases when the ϵ value increases because a lower ϵ value causes more privacy level and obtain lower accuracy. However, the error rate of the baseline approaches is higher than our PBA method because baseline methods use a fixed privacy budget even though the user stream involves temporal correlation. Notice that the error rates of the PBA and BA method are approximately similar because both methods follow the same allocation strategy and note that the BA method provides ϵ-LDP when the user stream does not involve temporal correlation, whereas the PBA method allows the user stream with temporal correlation and proves that the PBA achieves ϵ-LDP under temporal correlation.

In literature, there is no related work that directly solves our problem settings as per the best of our knowledge. We found three existing state-of-the-art methods that are most relevant to our problem setting. [59] proposed a Planar Isotropic Mechanism (PIM) for preserving a single trace of user's privacy, and it achieves differential privacy against the adversaries with knowledge of temporal correlation. [73] presents a local randomizer framework based on a randomized matrix (abbreviated as LRM) to protect a user location data-point at each timestamp in local settings. [74] introduced a generalized randomized response (GRR) mechanism, which achieves ϵ-LDP under temporal correlation in spatio

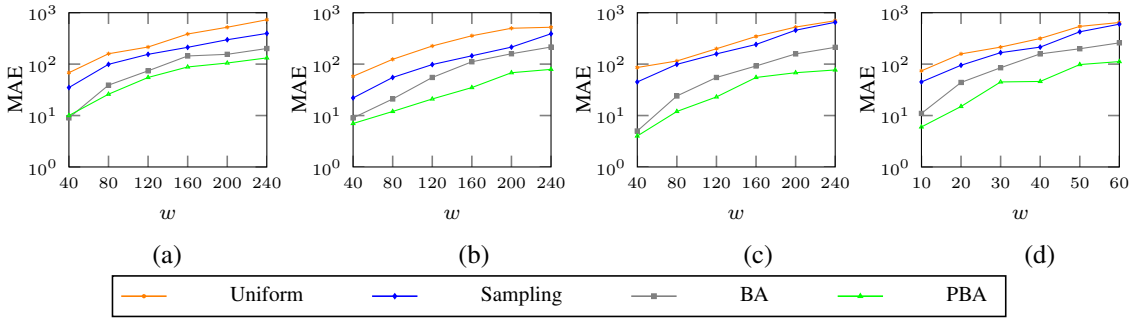

Figure 6.11: MAE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

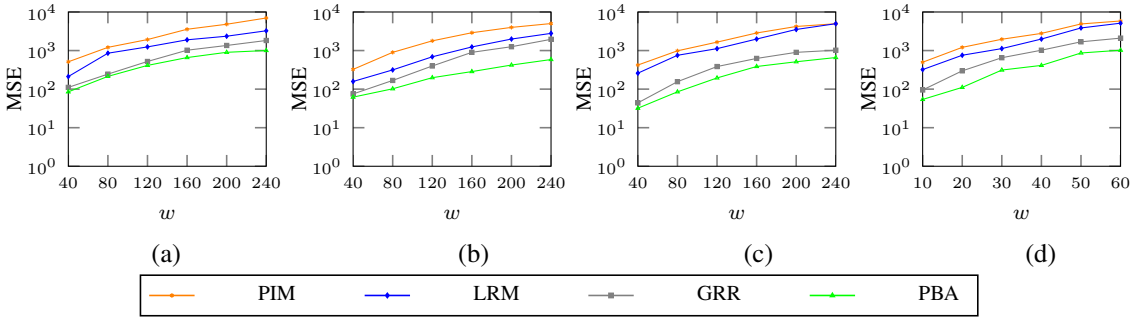

Figure 6.12: MSE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

temporal data. We conducted a set of experiments to tested PIM, LRM, GRR, and PBA methods at each timestamp in Geolife, T-Drive, ShangHai, and Metro100K datasets. Figures 6.11 and 6.12 show the performance of MAE and MSE between our PBA method with PIM, LRM, and GRR over time while varying the size of the sliding window (w) and a constant $\epsilon = 1$. The experimental results exhibit that both metric (MAE and MSE) error rate increases while increasing the size of the sliding window w. This is because larger w may have more temporally correlated data-points, which leads to more error rates. The error rate of the proposed PBA method provides a significant data utility compared with other state-of-the-art methods. This is because the PBA method allocates a privacy budget

only when the publication occurs at the timestamp. On the other side, the above state-of-the-art methods allocate budgets uniformly at every timestamp.

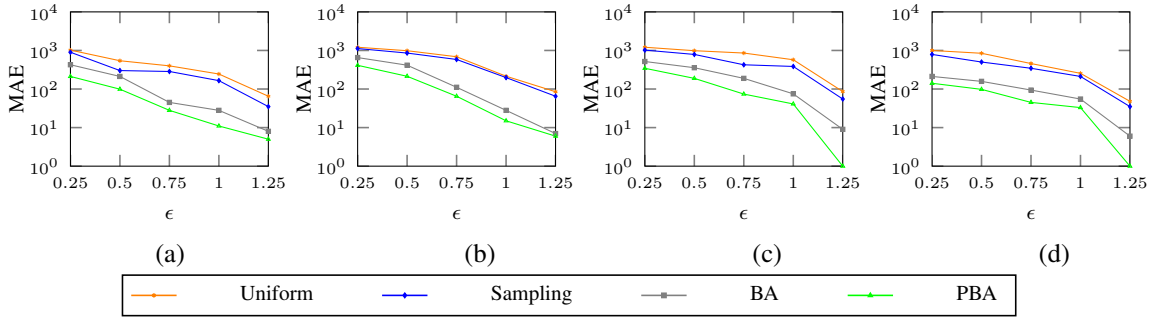

Figure 6.13: MAE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

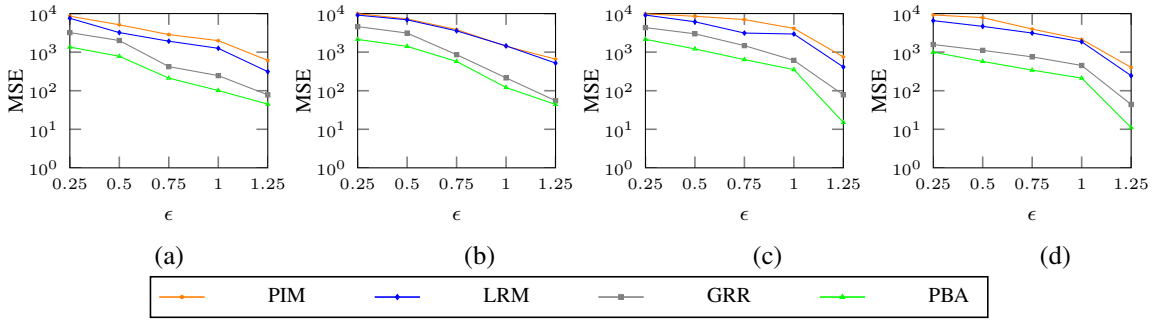

Figure 6.14: MSE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

Figures 6.13 and 6.14 show the error rate's performance between our proposed method and existing state-of-the-art methods while varying the ϵ values and a constant size $w = 40$. The experimental results show that the error rate decreasing while increasing the ϵ values. This is because a large privacy budget value (ϵ) leads to better data utility (or less error rate). The error rate of our proposed method is comparatively better than other existing methods. This is because the PBA method utilizes ϵ value only when the publications occur over time, and the remaining skipped publication's ϵ value will be used for future

publication over timestamps. So, the PBA method gives better data utility compared to other state-of-the-art methods. On the other hand, the existing state-of-the-art methods utilize constant e value over timestamp, leading to achieving ϵ-LDP but less data utility than our proposed method.

6.5 Summary

This work presents the definition of Local differential privacy under temporal correlation to study the impact of temporal correlation on privacy leakage. Then, we illustrate and show that the adversaries who know temporal correlation can disclose more privacy leakage than the traditional ϵ-LDP. The outcome of our study shows that the privacy leakage in w-event privacy increases over time when the temporal correlation is involved in the user stream. Therefore, we proposed a Privacy Budget Allocation (PBA) method for allocating an adequate amount of privacy budget to each successive timestamp under the protection of ϵ-LDP. Our proposed method provides a privacy guarantee to any w length temporally correlated user stream. Finally, we conduct experiments with real and synthetic datasets to compare the effectiveness of our proposed method with state-of-the-art methods. The demonstration results show that the proposed method is comparatively better under temporal correlation than the existing state-of-art methods.

Chapter 7

Compare the Impacts of Data Correlation on Privacy Leakage in a Combined Privacy Preserving Approaches

Many applications require either single or both phases: data collection and data sharing to provide better social benefits to users, such as intelligent healthcare system [94], intelligent traffic control systems [95], and online advertisements [96]. The service provider collects the user's private data and provides the services to users or service provider shares that collected private (or sensitive) data with other service providers for providing better social benefits to users. However, the collection and sharing of users' private data may compromise the user's privacy, leading to disclosing the user's sensitive information. Many privacy preservation methods have been proposed for providing a privacy guarantee at the time of data collection and data sharing phases [97, 98]. Initially, the anonymization

[99, 26] technique is a popular method to preserve user privacy in data publishing. However, it is challenging to balance the trade-off between data utility and user privacy. To this end, Differential Privacy (DP) [9] is a novel privacy preservation approach with a mathematical standard. This approach provides a strong privacy guarantee with the assumption that service providers are trustworthy. It is difficult to presume that all service providers are trusted, resulting in untrusted service providers misusing the information collected for other purposes [11, 12]. To overcome this, recently proposed a variant of standard Differential privacy for local settings, called Local Differential Privacy (LDP) [10]. It promises that it provides a strong privacy guarantee even though the service provider is not trusted.

There are a few privacy approaches in the literature under the protection of ϵ-DP and ϵ-LDP for continuous data release settings [93, 66, 17]. For continuous data release settings, Dwork et al. [100] presented two privacy approaches: event-level privacy and user-level privacy. Event-level privacy achieves ϵ-DP to each event, protecting a single data-point of a user trajectory. At the same time, user-level privacy achieves ϵ-DP to the finite length of a user trajectory. As a result, user-level privacy is limited usage in most real-world applications. To overcome these limitations, recently, a w-event privacy mechanism [17] has been introduced to address the limited applicability in real-world applications. This privacy mechanism adopts a sliding window methodology to achieve ϵ-DP to infinite length of user trajectory.

However, if the data collector or curator is malicious, users must send their sensitive information privately using the LDP mechanism. Or if the curator has the dataset and wants to share it with third parties, then the curator applies the DP mechanism and sends it to the third parties. Or the curator wants to release statistics for which it requires both private data sent by the user (using LDP) and collected stored data, and then the curator releases aggregated statistics using the DP mechanism, so the curator involves both mechanisms in

order to release aggregate statistics. In other words, few applications require either a privacy guarantee by the data collector (i.e., DP) or a privacy guarantee by the data provider itself (i.e., LDP) or a privacy guarantee by both data provider and data collector (i.e., LDP and DP). Many authors studied the privacy leakage of the only DP involved in the application or only LDP involved in the application, but not both mechanisms involved the application. So, it is necessary to study the privacy leakage of a combined approach.

Furthermore, when the data points are not correlated, w-event privacy protects users' privacy from the bounded knowledge of the adversary. However, in case the data points are correlated, then it provides less privacy guarantee than the traditional w-event privacy mechanism (either DP or LDP). This is because the quantity of privacy budgets which is allotted to each timestamp is not sufficient in continuous data release settings. Recently Privacy Budget Allocation strategies have been proposed to address this limitation in a continuous data release setting under the DP and LDP protection. However, it shows the impact of data correlation on privacy leakage in either DP or LDP mechanism but not shown the impact of data correlation on privacy leakage in a combined (LDP+DP) approach. Since this combined (LDP+DP) approach having many possible combinations by considering with or without temporal correlation, such as either it requires *traditional LDP + traditional DP* or *LDP with temporal correlation (TC)+ traditional DP* or *traditional LDP + DP with temporal correlation* or *LDP with temporal correlation + DP with temporal correlation*, it is necessary to study the impact of data correlation on privacy leakage in all possible combinations of the combined (LDP+DP) approach. Depends on the type of query, the curator chooses one from the different combinations of combined approaches to releasing statistics privately.

In summary, compare the privacy leakage of all combined approaches irrespective of whether the data-points involve temporally correlated or independent in continuous data

release setting with other states of the art methods. The contributions of this work are as follows.

1. We quantify the impact of data correlation on privacy leakage of all cases of a combined approach in continuous data release settings.

2. We performed a series of experiments to determine the average error rate per timestamp for evaluating the data utility of a combined approach (LDP with TC+DP with TC) with other states of the art methods.

The organization of this work is as follows. Section 7.1 presents a reformulated definitions of Differential Privacy in continuous data release settings and Local Differential Privacy in continuous data release settings. And also presented privacy leakage analysis of both DP and LDP mechanisms for correlated and non correlated datasets. Section 7.2 presents the comparative analysis of all possible combinations of combined approach and shows the effectiveness of data correlation on privacy leakage in a combined privacy approach. In section 7.3, conduct an experiment to evaluate the data utility of the various combination of LDP and DP mechanisms with other existing state-of-the-art methods. Finally, the summary of this work is presented in section 7.4.

7.1 System Framework

In this section, we presents a reformulated definitions of Differential Privacy in continuous data release settings and Local Differential Privacy in continuous data release settings. And also we describe the the privacy leakage of DP and LDP under temporal correlation.

7.1.1 Differential Privacy under continual observation

A privacy mechanism $\mathcal{M}$ is satisfied ϵ-DP, if and only if for any two neighboring stream prefixes differ by atmost one user trajectory i.e., S_t and $S_t^{'}$, and for any possible outputs ω of range($\mathcal{M}$), $\mathcal{M}$ holds

$$log\frac{P_r(\mathcal{M}(S_t) = (\omega_1, \omega_2, ..., \omega_t))}{P_r(\mathcal{M}(S_t^{'}) = (\omega_1, \omega_2, ..., \omega_t))} \leq \epsilon$$

Where ϵ represents the degree of privacy offered to users. The most common approach used for achieving ϵ-DP is Laplace mechanism. For instance, let a query q, according to the ϵ-DP, add random noise which is derived from Laplace distribution with scale $Lap(\lambda)$ to the true answer. The density function of Laplace distribution is as follows.

$$P(r) = \frac{1}{2\lambda}exp(-|r|/\lambda) \tag{7.1}$$

Where $\lambda = \Delta q/\epsilon$, Δq is a maximum difference between the outputs over the neighboring stream prefixes S_t and $S_t^{'}$ i.e., $\Delta q = \max_{S_t, S_t^{'}}||q(S_t) - q(S_t^{'})||$.

7.1.2 Local Differential Privacy under continual observation

A privacy mechanism $\mathcal{M}$ is satisfied ϵ-LDP, if and only if for any two any pair of input value x_i^t and $x_i^{t'}$ and for any possible outputs $\widehat{x}$ of range($\mathcal{M}$), $\mathcal{M}$ holds

$$log\frac{P_r(\mathcal{M}(x_i^t) = \widehat{x}_i^t)}{P_r(\mathcal{M}(x_i^{t'}) = \widehat{x}_i^t)} \leq \epsilon$$
$$log\frac{P_r(\mathcal{M}((T_i)^t) = (\widehat{x}_i^1, \widehat{x}_i^2, ..., \widehat{x}_i^t))}{P_r(\mathcal{M}((T_i)^{t'}) = (\widehat{x}_i^1, \widehat{x}_i^2, ..., \widehat{x}_i^t))} \leq \epsilon$$

Where ϵ_t represents the degree of privacy offered at timestamp t. In local differential pri-

vacy, each user encodes his vector bits of his location data-point (say $x_i^j[l], l = (1, 2, \ldots, L)$) before sending into curator. Thus the data collector cannot access the original data of the contributors or the users.

7.1.3 Privacy leakage analysis for non-correlated dataset of DP and LDP

Consider an adversary A_i, wants to infer i^{th} user location data-point at timestamp t, which means A_i knows other users' locatoion data-points except i^{th} user location data-point, i.e., A_i knows $S_t = S_t \backslash \{i\}$ or A_i knows $(T_i)^{t_k} = (T_i)^t \backslash \{l_i^t\}$.

The privacy leakage of DP and LDP mechanism for non-correlated dataset at timestamp t against A_i is as follows, assume that l_i^t and $l_i^{t'}$ are the i^{th} user location data-points at timestam t.

$$\mathcal{L}_{DP}(\mathcal{M}_t, A_i) = \sup_{\omega, l_i^t, l_i^{t'}} log \frac{P_r(\omega | l_i^t, S_t)}{P_r(\omega | l_i^{t'}, S_t')}$$

$$\mathcal{L}_{LDP}(\mathcal{M}_t, A_i) = \sup_{l_i^t, l_i^{t'}, \widehat{x}_i} log \frac{P_r(\widehat{x}_i | l_i^t, (T_i)^{t_k})}{P_r(\widehat{x}_i | l_i^{t'}, (T_i)^{t_k})}$$

The above equations are lesser ϵ value, then lesser the privacy leakage and vice-verse. Here, we considered a privacy budget ϵ as a metric for privacy leakage.

7.1.4 Privacy leakage analysis for temporally correlated dataset of DP and LDP

With the knowledge of adversary from other sources, It is reasonable to believe that an adversary knows the transition probability between all possible position data-points. For modeling a transition probability between location data-points (according to some proba-

bilistic rules), we used a Markov chain process (MC), which is denoted as $\theta \in \Theta$, where Θ is a transition probability distributions of all possible position data-points. For instance, A_i knows $S_t = S_t \backslash \{i\}$ or $(T_i)^{t_k} = (T_i)^t \backslash \{l_i^t\}$ and θ, so A_i is represented as A_i^θ. The privacy leakage of DP and LDP mechanism for temporally correlated ($\mathcal{TC}$) dataset against A_i at time t is as follows, where l_i^t and $l_i^{t'}$ are two i^{th} user data-points at timestamps t.

$$\mathcal{TCL}_{DP}(\mathcal{M}_t, A_i^\theta) = \sup_{\omega, l_i^t, l_i^{t'}, \theta} log \frac{P_r(\omega | l_i^t, S_t, \theta)}{P_r(\omega | l_i^{t'}, S_t', \theta)}$$

$$\mathcal{TCL}_{LDP}(\mathcal{M}_t, A_i^\Theta) = \sup_{\widehat{x}_i, l_i^t, l_i^{t'}, \Theta} log \frac{P_r(\widehat{x}_i^1, ..., \widehat{x}_i^t | l_i^t, (T_i)^{t_k}, \Theta)}{P_r(\widehat{x}_i^1, ..., \widehat{x}_i^t | l_i^{t'}, (T_i)^{t_k}, \Theta)}$$

The privacy leakage of DP and LDP mechanisms for $\mathcal{TC}$ dataset at timestamp t w.r.t A_i^θ where $i \in [k]$ are less than or equal to privacy leakage metric (ϵ). Then, the adversary cannot distinguish whether the data point of user i is l_i^t or $l_i^{t'}$ with high confidence in DP and same as in LDP also. If the ϵ value is less, then less privacy leakage or vice-versa. The privacy leakage of DP and LDP for $\mathcal{TC}$ datasets attain more privacy leakage than the traditional DP and LDP for non-correlated datasets. To understand the impact of temporal correlation on privacy leakage in DP in continuous data publish settings, the above Equation is expanded and simplified by Bayes theorem, i.e.,

$$\begin{aligned}
\mathcal{TCL}_{DP}(\mathcal{M}_t, A_i^\theta) &= \sup_{\omega_1, ..., \omega_t, l_i^t, l_i^{t'}, \theta} log \frac{P_r(\omega_1, ..., \omega_t | l_i^t, S_t, \theta)}{P_r(\omega_1, ..., \omega_t | l_i^{t'}, S_t', \theta)} \\
&= \sup_{\omega_1, ..., \omega_{t-1}, l_i^t, l_i^{t'}, \theta} log \frac{\sum_{l_i^{t-1}} P_r(\omega_1, ..., \omega_{t-1} | l_i^{t-1}, S_{t-1}, \theta) P_r(l_i^{t-1} | l_i^t)}{\sum_{l_i^{t-1'}} P_r(\omega_1, ..., \omega_{t-1} | l_i^{t-1'}, S_{t-1}', \theta) P_r(l_i^{t-1} | l_i^t)} \\
&\quad + \sup_{\omega_t, l_i^t, l_i^{t'}, \theta} log \frac{P_r(\omega_t | l_i^t, S_t, \theta)}{P_r(\omega_t | l_i^{t'}, S_t', \theta)} \qquad (7.2)
\end{aligned}$$

Similarly in ϵ-LDP,

$$= \sup_{\widehat{x}_i^1,.....,\widehat{x}_i^{t-1},x_i^t,x_i^{t'},\theta} log\frac{\sum_{x_i^{t-1}} P_r(\widehat{x}_i^1,.....,\widehat{x}_i^{t-1}|x_i^{t-1},(T_i)^{t_{k-1}},\theta) * P_r(x_i^{t-1}|x_i^t)}{\sum_{x_i^{t-1'}} P_r(\widehat{x}_i^1,.....,\widehat{x}_i^{t-1}|x_i^{t-1'},(T_i)^{t_{k-1}},\theta) * P_r(x_i^{t-1'}|x_i^{t'})} + \sup_{\widehat{x}_i^t,x_i^t,x_i^{t'},\theta} log\frac{P_r(\widehat{x}_i^t|x_i^t,(T_i)^{t_k},\theta)}{P_r(\widehat{x}_i^t|x_i^{t'},(T_i)^{t_k},\theta)} \quad (7.3)$$

The first two terms of the above equations in ϵ-DP and ϵ-LDP are to determine temporal privacy leakage and this can be calculated using linear fractional programing (LFP), discussed in [18]. In detail, let q and d are the two distinct rows of transition probability matrix Θ and α be the privacy parameter that quantifies the level of temporal privacy leakage. According to LFP, the maximum value of objective function is

$$\mathcal{TCL}_{DP/LDP}(\mathcal{M}_{t-1}, A_i^\theta) = \max_{q,d\in\Theta} log\frac{q(e^\alpha - 1)+1}{d(e^\alpha - 1)+1} \quad (7.4)$$

The local randomizer method satisfies ϵ-local differential privacy where $\epsilon = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$. Therefore the total privacy leakage under temporal correlation in continuous data release is as follows.

In ϵ-DP:

$$\mathcal{TCL}_{DP}(\mathcal{M}_t, A_i^\theta) = \max_{q,d\in\Theta} log\frac{q(e^\alpha - 1)+1}{d(e^\alpha - 1)+1} + \epsilon \quad (7.5)$$

In ϵ-LDP:

$$\mathcal{TCL}_{LDP}(\mathcal{M}_t, A_i^\theta) = \max_{q,d\in\Theta} log\frac{q(e^\alpha - 1)+1}{d(e^\alpha - 1)+1} + 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right) \quad (7.6)$$

The above privacy analysis shows that privacy leakage is increased when the user's lo-

cation data-points are temporally correlated. Hence, we can claim that the curator (or data publisher) has more chance to disclose sensitive data (or location data-point) than traditional ϵ-LDP in continuous data publications settings. However, the w-event privacy mechanism achieves ϵ-DP and ϵ-LDP in continuous data release settings, assuming the data-points in the trajectory are independent. In real-time data collection and data sharing phases, the location data-points are temporally correlated with a certain probability. Due to the presence of temporal correlation, the ratio of privacy budget to each timestamp in w-event privacy is not adequate. Hence, w-event privacy fails to achieve ϵ-DP and ϵ-LDP, especially when the data-points are temporally correlated. Recently, Privacy Budget Allocation (PBA) [101] strategies have been proposed to address this limitation in a continuous data release setting under the DP and LDP protection.

There are few applications that require either a privacy guarantee by the data collector (i.e., DP) or a privacy guarantee by the data provider itself (i.e., LDP) or a privacy guarantee by both data provider and data collector (i.e., LDP and DP). If the dataset involves temporal correlation ($\mathcal{TC}$), then the applications require different cases of mechanisms such as only *DP under* $\mathcal{TC}$ or *only LDP under* $\mathcal{TC}$ or *LDP under* $\mathcal{TC}$ *and traditional DP* or *traditional LDP and DP under* $\mathcal{TC}$ or *LDP under* $\mathcal{TC}$ *and DP under* $\mathcal{TC}$. However, the recently proposed PBA approach shows the impact of data correlation on privacy leakage in either the DP or LDP mechanism but not shown the impact of data correlation on privacy leakage in a combined (LDP+DP) approach. Since this combined (LDP+DP) approach having many possible combinations by considering with or without temporal correlation, it is necessary to study the impact of data correlation on privacy leakage in all possible combinations of the combined (LDP+DP) approach.

Figure 7.1 shows the framework of combined (LDP+DP) approach by considering with or without temporal correlation. Let assume that a curator receives LDP statistics from a user

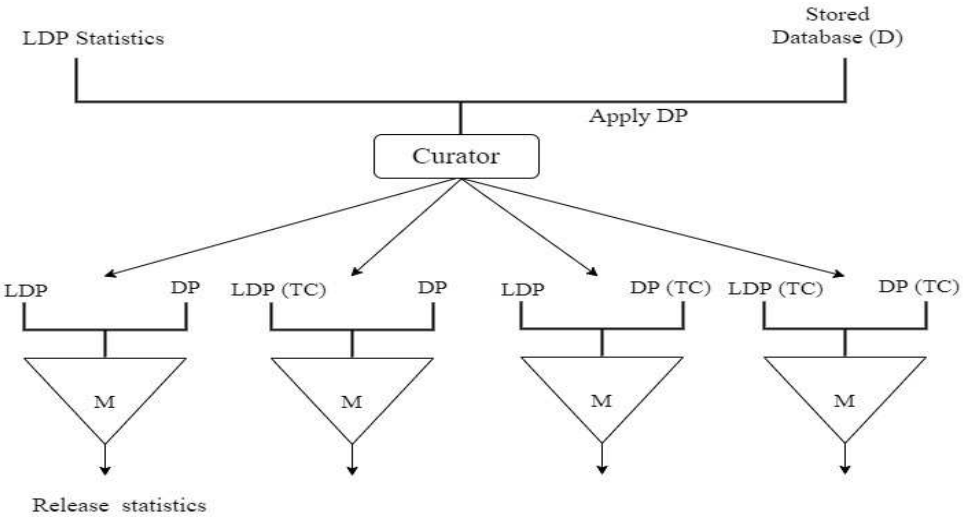

Figure 7.1: Different combinations of combined (LDP+DP) approach assuming with or without temporal correlation

and use the stored database, to answer a query. The curator uses either a stored database directly or made it private by using the DP mechanism depends on the query requirements. There are four possible cases of combined (LDP+DP) approach in our framework. Based on the type of datasets (either temporally correlated or not), the curator can choose a suitable case of mechanisms to answer a statistical query. For instance, a curator wants to release an answer to count query, which requires both LDP statistics and stored database D. If the stored database involves correlation, the curator chooses either $LDP + DP(\mathcal{TC})$ mechanism. If the query is correlated with the released existing query, the curator must choose $LDP(\mathcal{TC}) + DP(\mathcal{TC})$ mechanism.

7.2 Comparative analysis

This section analyzes the privacy leakage of all cases of combined approach DP and LDP mechanisms with or without considering temporal correlation.

7.2.1 LDP and DP

To answer a query, the curator receives perturbed data sent by the user using Local Differential privacy (called LDP statistic), and the stored datasbase. Since the curator uses stored database to answer a query, curator uses both traditional (LDP+DP) mechanism releases aggregate statistics. The privacy leakage of the combined approach at timestamp t against A_i is as follows.

$$\begin{aligned}\mathcal{PL}_{(LDP+DP)}(\mathcal{M}_t, A_i) &= \mathcal{L}_{LDP}(\mathcal{M}_t, A_i) + \mathcal{L}_{DP}(\mathcal{M}_t, A_i) \\ &= 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right) + \epsilon\end{aligned}$$

The first term of the above equation is a value which satisfies ϵ-LDP at timestamp t, where $\epsilon = .2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$.

7.2.2 LDP($\mathcal{TC}$) and DP

To answer a query, the curator receives perturbed data sent by the user using Local Differential privacy (called LDP statistic), and the stored datasbase. However, the data-points sent by the user at different timestamp may be temporally correlated. If the data-points are correlated, then the traditional ϵ-LDP in a continuous data release setting does not provide expected privacy guarantee. Therefore, the curator uses a combined approach of *LDP($\mathcal{TC}$) and DP* combination in order to release aggregate statistics. The privacy leakage of this combined approach at timestamp t against A_i is as follows.

$$\begin{aligned}\mathcal{PL}_{(LDP(\mathcal{TC})+DP)}(\mathcal{M}_t, A_i) &= \mathcal{TCL}_{LDP}(\mathcal{M}_t, A_i^{\Theta}) + \mathcal{L}_{DP}(\mathcal{M}_t, A_i) \\ &= \left(\max_{q,d\in\Theta} log\frac{q(e^{\alpha}-1)+1}{d(e^{\alpha}-1)+1} + 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)\right) + \epsilon\end{aligned}$$

The privacy leakage of $\mathcal{TCL}_{LDP}$ involves two terms; the first term is to determine the temporal privacy leakage between two adjacent data-points at timestamp t and $t-1$. This term has derived by using linear fractional programming (LFP), discussed in [18]. The second term is a value which satisfies ϵ-LDP at timestamp t, where $\epsilon = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$.

7.2.3 LDP and DP($\mathcal{TC}$)

To answer a query, the curator receives perturbed data sent by the user using Local Differential privacy (called LDP statistic), and the stored datasbase. However, the data-points at different timestamp in stored database may be temporally correlated. If the data-points are correlated, then the traditional ϵ-DP in a continuous data release setting does not provide expected privacy guarantee. Therefore, the curator uses a combined approach of *LDP and DP($\mathcal{TC}$)* combination in order to release aggregate statistics. The privacy leakage of this combined approach at timestamp t against A_i is as follows.

$$\begin{aligned}\mathcal{PL}_{(LDP+DP(\mathcal{TC}))}(\mathcal{M}_t, A_i) &= \mathcal{L}_{LDP}(\mathcal{M}_t, A_i) + \mathcal{TCL}_{DP}(\mathcal{M}_t, A_i^{\Theta}) \\ &= 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right) + \left(\max_{q,d\in\Theta} log\frac{q(e^{\alpha}-1)+1}{d(e^{\alpha}-1)+1} + \epsilon\right)\end{aligned}$$

The privacy leakage of $\mathcal{TCL}_{DP}$ involves two terms; the first term is to determine the temporal privacy leakage between two adjacent data-points at timestamp t and $t-1$. This term has derived by using linear fractional programming (LFP), discussed in [18] The second term is a value which satisfies ϵ-LDP at timestamp t, where $\epsilon = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$.

7.2.4 LDP($\mathcal{TC}$) and DP($\mathcal{TC}$)

To answer a query, the curator receives perturbed data sent by the user using Local Differential privacy (called LDP statistic), and the stored datasbase. However, the data-points

sent by the user at different timestamp may be temporally correlated and the data-points at different timestamp in stored database may be temporally correlated. If the data-points are correlated in both datasets, then the traditional ϵ-LDP and traditional ϵ-DP in a continuous data release setting does not provide expected privacy guarantee. Therefore, the curator uses a combined approach of *LDP($\mathcal{TC}$) and DP($\mathcal{TC}$)* combination in order to release aggregate statistics. The privacy leakage of this combined approach at timestamp t against A_i is as follows.

$$\begin{aligned}\mathcal{PL}_{(LDP(\mathcal{TC})+DP(\mathcal{TC}))}(\mathcal{M}_t, A_i) &= \mathcal{TCL}_{LDP}(\mathcal{M}_t, A_i^{\Theta}) + \mathcal{TCL}_{DP}(\mathcal{M}_t, A_i^{\Theta}) \\ &= \left(\max_{q,d\in\Theta} log\frac{q(e^{\alpha}-1)+1}{d(e^{\alpha}-1)+1} + 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)\right) \\ &+ \left(\max_{q,d\in\Theta} log\frac{q(e^{\alpha}-1)+1}{d(e^{\alpha}-1)+1} + \epsilon\right) \\ &= \max_{q,d\in\Theta} log\frac{q(e^{\alpha}-1)+1}{d(e^{\alpha}-1)+1}\left(2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right) + \epsilon\right)\end{aligned}$$

The first term in both $\mathcal{TCL}_{LDP}$ and $\mathcal{TCL}_{DP}$ is to determine the temporal privacy leakage between two adjacent data-points at timestamp t and $t-1$. This term has derived by using linear fractional programming (LFP), discussed in [18]. The second term in $\mathcal{TCL}_{LDP}$ and $\mathcal{TCL}_{DP}$ is a value which satisfies ϵ-LDP and ϵ-DP at timestamp t, where $\epsilon = 2hln\left(\frac{1-\frac{1}{2}p}{\frac{1}{2}p}\right)$ and ϵ respectively.

7.3 Experimental Results

In this section we conduct a series of experiments to examine the effect data correlation on privacy leakage of combined (LDP+DP) approach. Also we evaluate the data utility of combined (LDP+DP) approach by adopting various state-of-the-art methods od LDP and DP mechanisms.

In this experiment we used three real data sets and one synthetic dataset to test the efficiency of the various combination of LDP and DP mechanisms. A $Geolife$ is a real-time dataset that includes GPS user trajectories obtained over a span of one year from 182 users in real-time. A real-time $T-Drive$ dataset includes a 10357 taxi GPS trajectory over a span of one year. A $Gowalla$ real-time dataset consists of a check-in background of more than 600000 users from November 2010 to December 2011. Finally, a $Metro100K$ is a simulated dataset containing trajectories of 100000 gathered in 24 hours from a metropolitan area with 26 towns. A collection of tuples containing an user Id, longitude, latitude, and timestamp are involved in these four datasets. In our experiment, we optimized the real-time datasets by assuming that a user is located at most one location at each timestamp and collected all samples (user location data point) per interval of 5 minutes.

We used two metrics for measuring the data utility of published statistics, such as Mean of absolute error(MAE) and Mean of square error (MSE). These two metrics calculate the dissimilarity (or error) between the published statistic and the actual statistic. According to ϵ-(LDP+DP), each timestamp receives a piece of the privacy budget ϵ in order to achieve ϵ-(LDP+DP) in continuous data release settings. So, we calculate the dissimilarity (or error) value per timestamp - the description of the MAE and MSE metrics are,

$$MAE() = \frac{1}{|T|}\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\widehat{x}_i^l[j] - x_i^k[j]|$$

$$MSE() = \frac{1}{|T|}\frac{1}{|L|}\Sigma_{j=1}^{|L|}|\widehat{x}_i^l[j] - x_i^k[j]|^2$$

Where x_i^t and $\widehat{x}_i^l$ are the location data-point of user i at time t and perturbed location data-point of user i at time l (i.e., last release data-point) respectively. The T be the total number of location data-points of user i.

There are few common baseline methods in the literature under the protection of ϵ-LDP

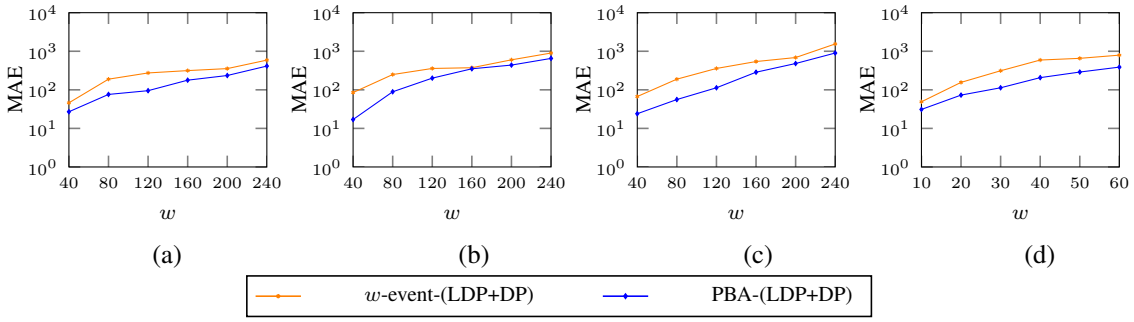

Figure 7.2: MAE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

and ϵ-DP in continuous data release settings, such as Uniform, Sampling, w-event privacy. These baseline methods allocates privacy budgets to each timestamp in continuous data release settings under the protection differential privacy. But, these methods are assuming the data-points in user streams are independent. Hence, these methods are not consider the data-points which are temporally correlated. Whereas, a recently proposed a privacy budget allocation (PBA) method considers temporally correlated data-points while achieving ϵ-LDP and ϵ-DP. Since Uniform and sampling methods are basic methods and attain more error rates than the recent approach w-event privacy thus, we measure the combined approach's error rates (LDP+DP) under the protection of w-event privacy and PBA method by using MAE and MSE metrics. Figures 7.2 and 7.3 shows the error rate of combined approach (LDP+DP) under the protection of w-event privacy and PBA while varying the sliding window size w. We observed that the error rate of w-event (LDP+DP) and PBA-(LDP+DP) increases while increasing the sliding window size w because each timestamp's privacy budget is minimized when the large size of the sliding window. However, the PBA-(LDP+DP) method's error rate is comparatively less than the w-event (LDP+DP) because the appropriate amount of privacy budget is allotted at temporally correlated timestamps in the PBA method than the baseline approaches.

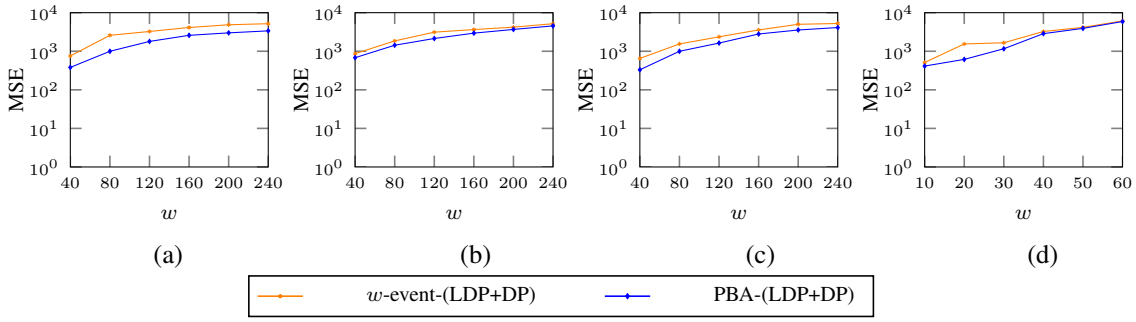

Figure 7.3: MSE vs. w while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

We compared the w-event (LDP+DP) and PBA-(LDP+DP) methods while varying ϵ values, as shown in Figures 7.4 and 7.5. We found that the error rate of w-event (LDP+DP) and PBA-(LDP+DP) decreases as the ϵ value increases because a lower ϵ value causes more privacy level and achieves lower accuracy. However, the error rate of the w-event (LDP+DP) is higher than the PBA-(LDP+DP) method because the w-event (LDP+DP) method uses a fixed privacy budget even though the user stream involves temporal correlation.

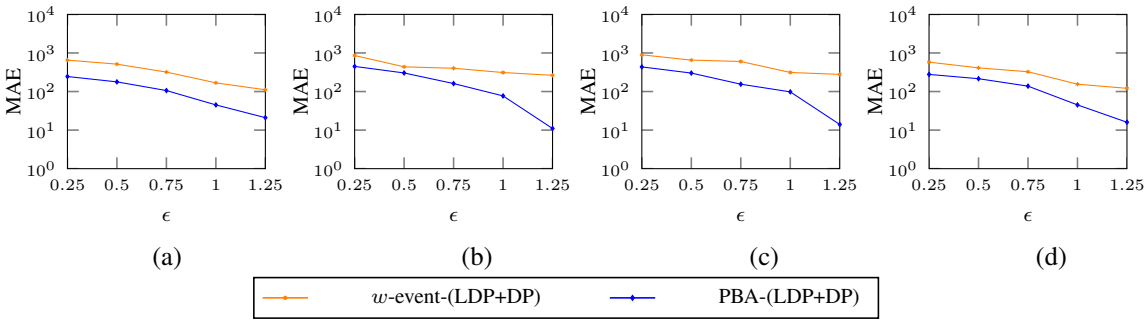

Figure 7.4: MAE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

There are four possible cases in a combined approach, such as (traditional LDP+traditional DP), (traditional LDP+DP with TC), (LDP with TC+traditional DP), and (LDP with TC+DP

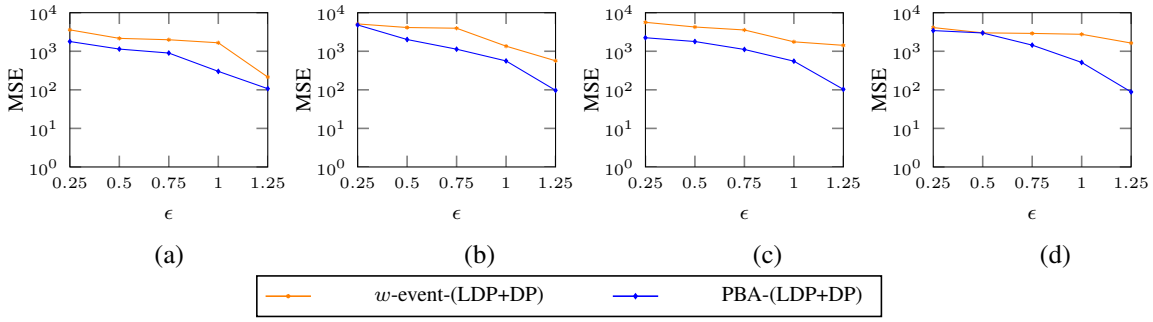

Figure 7.5: MSE vs. ϵ while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

with TC). Here, we analyze the privacy leakage of only (LDP with TC+DP with TC) combined approach with existing state-of-the-art methods. There is no related work that directly solves the combined approach (LDP+DP) and also limited research on solving temporal correlation under ϵ-LDP and temporal correlation under ϵ-DP. We found three existing state-of-the-art methods under ϵ-DP and four methods under ϵ-LDP that are most relevant to our problem setting. Then we compare the error rates of the combined approach by adapting existing methods under ϵ-LDP and ϵ-DP and analyze which combination of approaches give better data utility. The Planar Isotropic Mechanism (PIM), Local Randomizer Framework (LRM), Generalized Randomized Response (GRR), and Privacy Budget Allocation (PBA) are the recent approaches for temporal correlation under ϵ-LDP and the Quantification and Planar Isotropic Mechanism (PIM) are the methods for temporal correlation under ϵ-DP. Figures 7.6 and 7.7 shows the error rates of MAE and MSE between the combined approaches such as PIM under LDP with quantification under DP (LDP(PIM)+DP(quantification)), LRM under LDP with quantification under DP (LDP(LRM)+DP(quantification)), GRR under LDP with quantification under DP (LDP (GRR)+DP(quantification)) and PBA under LDP with quantification under DP (LDP(PBA)+ DP(quantification)) while varying the size of the sliding window (w) and a

constant $\epsilon = 1$.

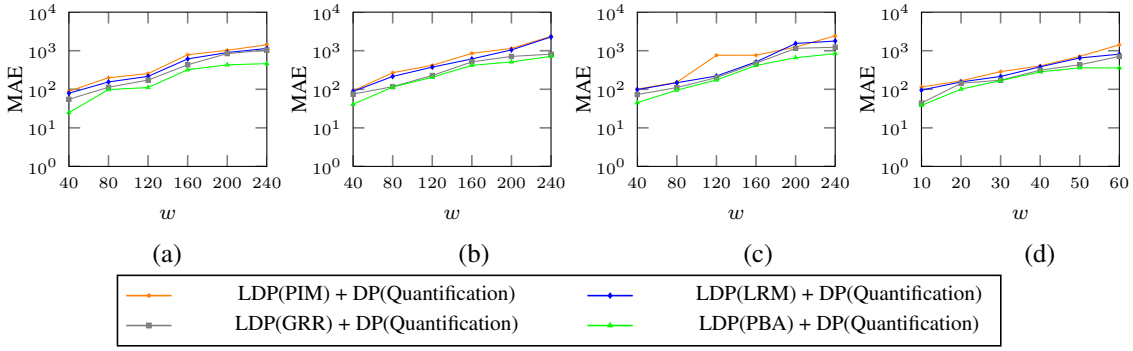

Figure 7.6: MAE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

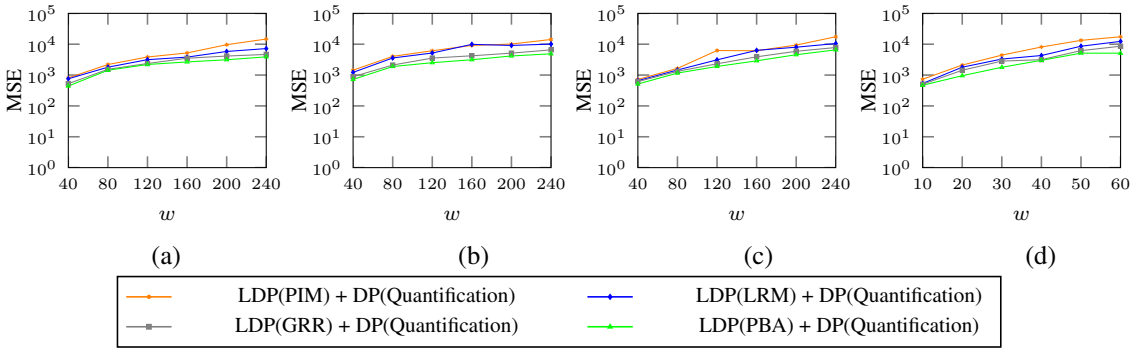

Figure 7.7: MSE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

The result exhibits that the above-combined approaches' error rate increases while increasing the size of the sliding window w. This is because larger w may have more temporally correlated data-points, which leads to more error rates. The error rate of (LDP(PBA)+DP(quantification)) combined approach provides a significant data utility compared with other combined approaches because the LDP(PBA) method allocates a

privacy budget only when the publication occurs at the timestamp and the quantification under DP allocates more privacy budget when the data-points are temporally correlated. On the other side, the remaining combined approaches PIM, LRM, GRR under LDP, allocate privacy budget uniformly at every timestamp. Hence, these approaches' error rate is not better than LDP(PBA)+DP(quantification).

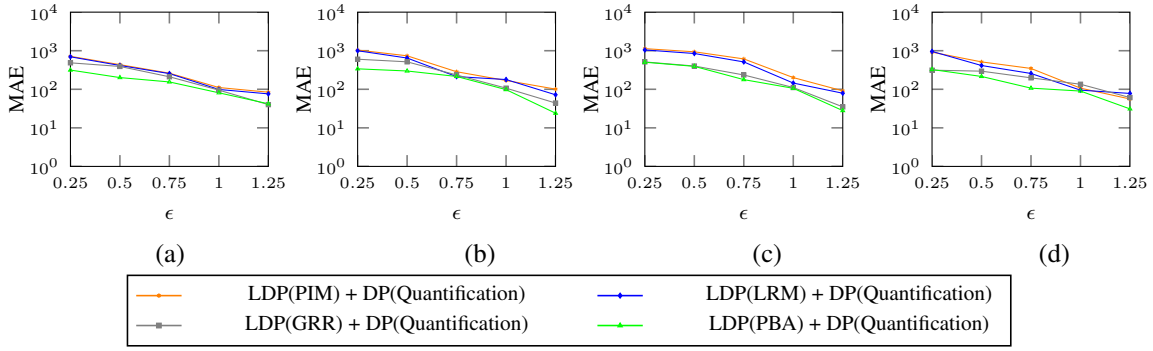

Figure 7.8: MAE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

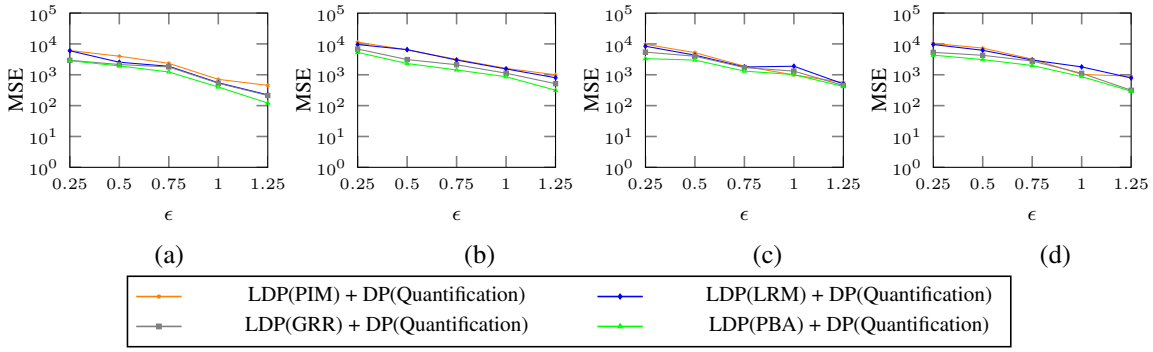

Figure 7.9: MSE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

Figures 7.8 and 7.9 show the error rates of the combined approaches LDP(PIM) + DP (quantification), LDP(LRM) + DP(quantification), LDP(GRR) + DP(quantification), LDP

(PBA) + DP(quantification) while varying the ϵ values and a constant size $w = 40$. The experimental results show that the error rates of the above-combined approaches decreasing while increasing the ϵ values because a high privacy budget value (ϵ) contributes to improved data utility (or less error rate). The error rate of LDP(PBA)+DP(quantification) is is comparatively better than the remaining combined approaches because the PBA under LDP uses ϵ value only when the publications occur over time, and the remaining skipped publication's ϵ value will be used for future publication over timestamps, and the quantification under DP uses an increased privacy budget $\alpha \geq \epsilon$, which leads to decrease the error rate even though the data-points involves temporal correlation. So, the PBA method under LDP and quantification under DP provide better data utility than other remaining combined approaches.

Figures 7.10 and 7.11 show the error rates of the combined approaches LDP(PIM)+DP(PIM), LDP(LRM)+DP(PIM), LDP(GRR)+DP(PIM), LDP(PBA)+DP(PIM) while varying the w values and a constant $\epsilon = 1$. The error rate of combined approach LDP(PBA)+DP(PIM) provides a significant data utility compared with other combined approaches because the LDP(PBA) method allocates a privacy budget only when the publication occurs at the

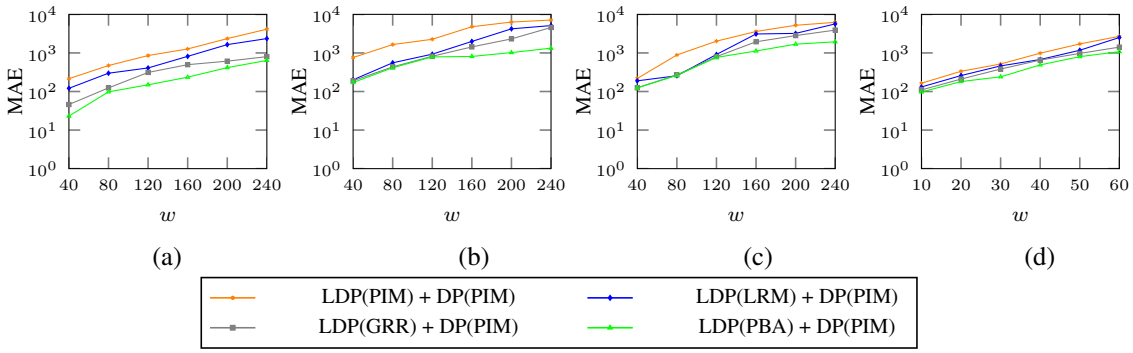

Figure 7.10: MAE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

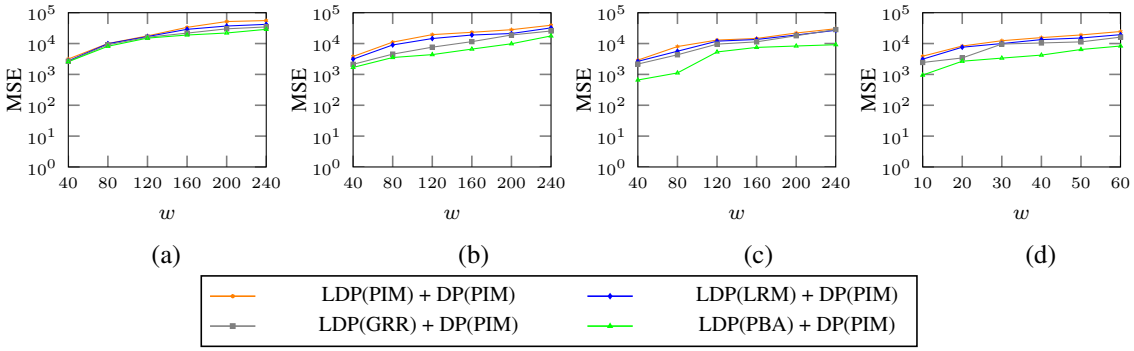

Figure 7.11: MSE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

timestamp and the PIM under DP allocates privacy budget uniformly to each timestamp. Even though PIM under DP error rate increases linearly due to uniform allocation, because PBA under LDP, the combined approach LDP(PBA)+DP(PIM) provides better data utility.

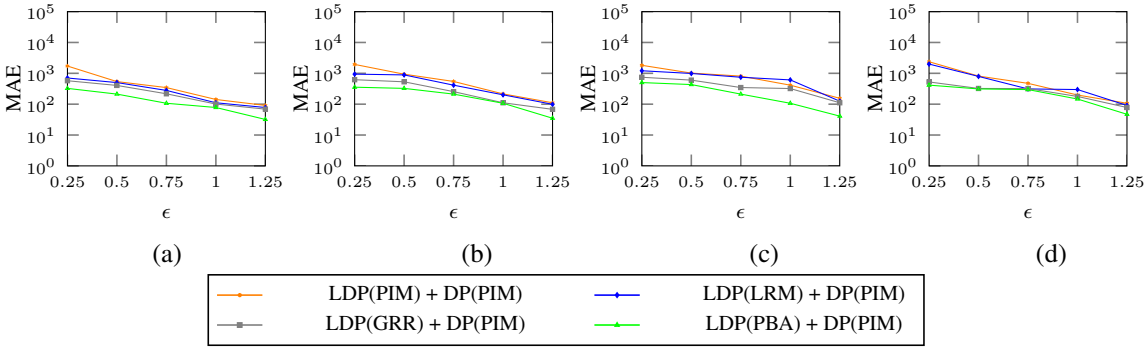

Figure 7.12: MAE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

Figures 7.12 and 7.13 show the error rates of the combined approaches LDP(PIM)+DP (PIM), LDP(LRM)+DP(PIM), LDP(GRR)+DP(PIM), LDP(PBA)+DP(PIM) while varying the ϵ values and a constant size $w = 40$. The error rate of LDP(PBA)+DP(PIM) is comparatively better than the remaining combined approaches because the PBA under

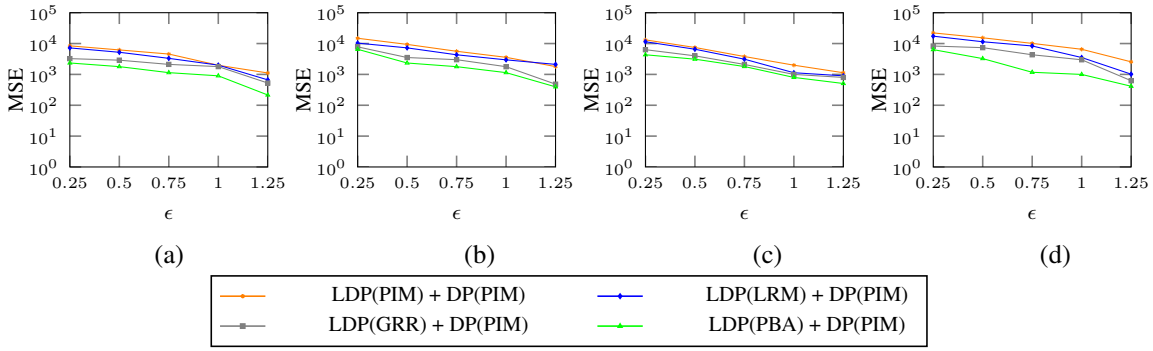

Figure 7.13: MSE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

LDP uses ϵ value only when the publications occur over time, and the remaining skipped publication's ϵ value will be used for future publication over timestamps, and the PIM under DP follows a uniform approach for allocating privacy budget, which leads to increase the error rate. Even though PIM under DP error rate increases linearly due to uniform allocation, because PBA under LDP, the combined approach LDP(PBA)+DP(PIM) provides better data utility while varying an ϵ compared to other combined approaches.

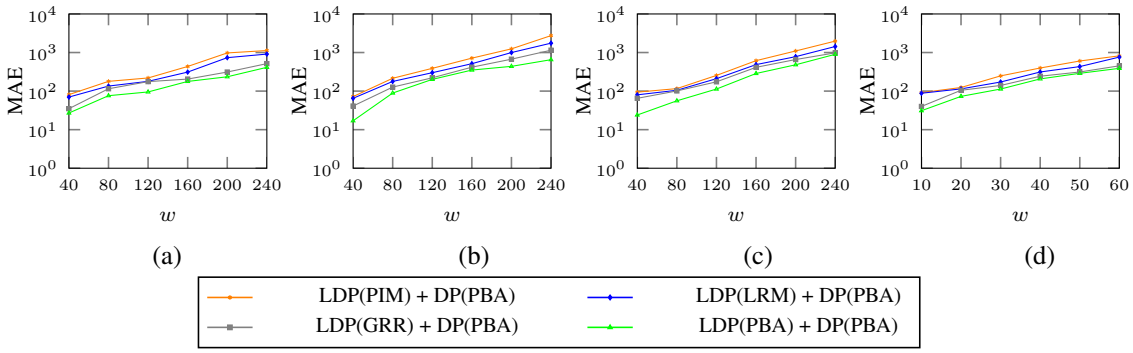

Figure 7.14: MAE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

Figures 7.14 and 7.15 show the error rates of the combined approaches LDP(PIM)+DP(PBA),

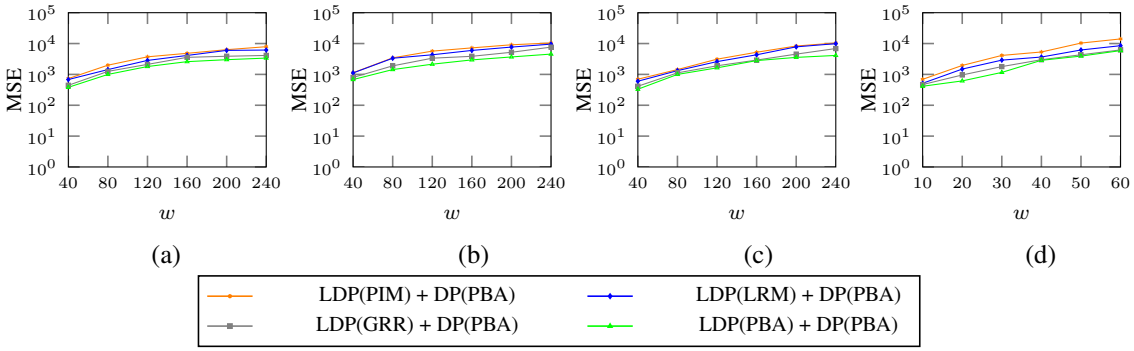

Figure 7.15: MSE vs w: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $\epsilon = 1$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

LDP(LRM)+DP(PBA), LDP(GRR)+DP(PBA), LDP(PBA)+DP(PBA) while varying the w values and a constant $\epsilon = 1$. The error rate of combined approach LDP(PBA)+DP(PBA) offers a significant data utility compared with other combined approaches because the LDP(PBA) method allocates a privacy budget only when the publication occurs at the timestamp and the DP(PBA) method also follows the same strategy that is it assign privacy at timestamp only the publication occurs. Since both privacy mechanisms (LDP and DP) adopt the PBA method, the combined approach LDP(PBA)+DP(PBA) offers better data utility.

Figures 7.16 and 7.17 show the error rates of the combined approaches LDP(PIM)+DP(PBA), LDP(LRM)+DP(PBA), LDP(GRR)+DP(PBA), LDP(PBA)+DP(PBA) while varying the ϵ values and a constant size $w = 40$. The error rate of LDP(PBA)+DP(PIM) is comparatively better than the remaining combined approaches because the PBA under LDP uses ϵ value only when the publications occur over time, and the remaining skipped publication's ϵ value will be used for future publication over timestamps, and the PBA under DP also follows the same allocation strategy that is it uses privacy budget value only when the publications occur, and any skipped publication's ϵ value is collected for future publica-

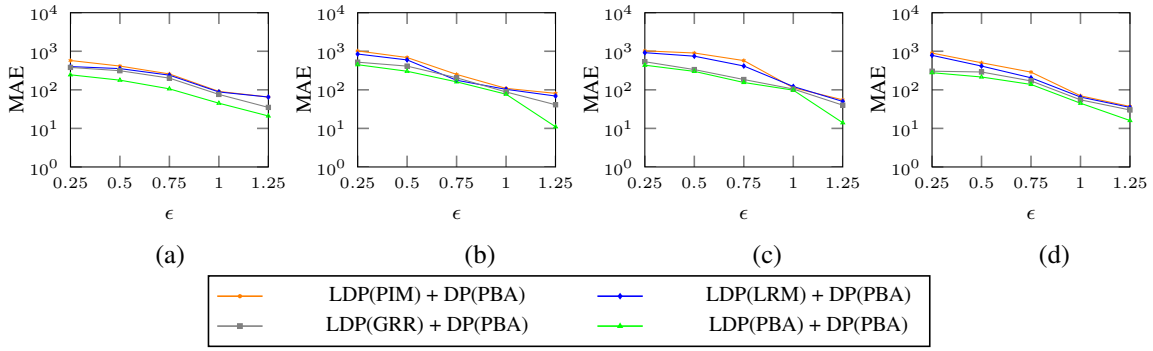

Figure 7.16: MAE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

tion. Since both privacy mechanisms (LDP and DP) adopt the PBA method, the combined approach LDP(PBA)+DP(PBA) offers better data utility than other combinations of the combined approach.

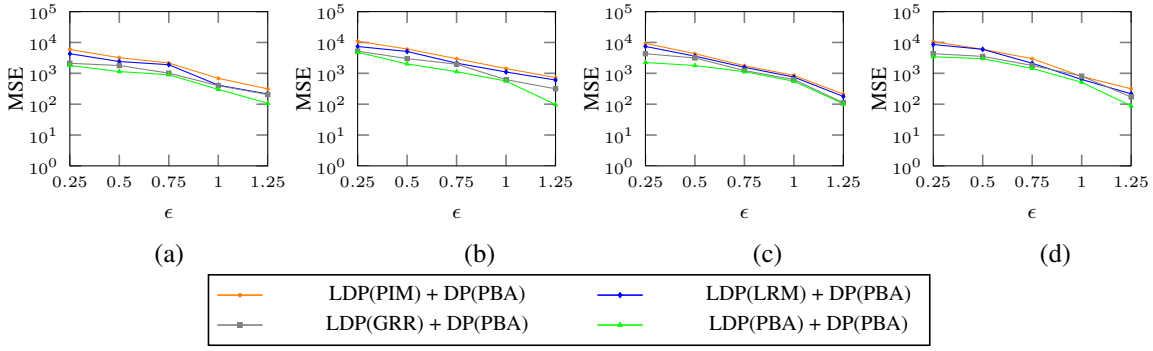

Figure 7.17: MSE vs ϵ: compare all combined approaches by adopting state-of-the-art methods of LDP and DP while fixing $w = 40$ (a) Geolife (b) T-Drive (c) Gowalla and (d) Metro100K datasets.

Table 7.1 shows that the privacy guarantee of combined approaches by adopting various methods of LDP and DP under temporal correlation. Assume that the length of temporally correlated data-points within user stream is T. We observed that the LDP(PBA)+DP(PBA) method achieves 2ϵ-(LDP+DP) under temporal correlation compared with other combined

Table 7.1: The privacy guarantee of combined approaches by adopting various methods of LDP and DP under temporal correlation

Privacy approaches	Temporal correlation	privacy guarantee on T length user stream
w-event-(LDP+DP)	−	$2w\epsilon$-(LDP+DP)
PBA-(LDP+DP)	✓	2ϵ-(LDP+DP)
LDP(PIM)+DP(Quantification)	−	$(T\epsilon + \alpha)$-(LDP+DP)
LDP(LRM)+DP(Quantification)	−	$(T\epsilon + \alpha)$-(LDP+DP)
LDP(GRR)+DP(Quantification)	✓	$((\epsilon, \delta) + \alpha)$-(LDP+DP)
LDP(PBA)+DP(Quantification)	✓	$(2\epsilon + \alpha)$-(LDP+DP)
LDP(PIM)+DP(PIM)	−	$(2T\epsilon$-(LDP+DP)
LDP(LRM)+DP(PIM)	−	$(2T\epsilon$-(LDP+DP)
LDP(GRR)+DP(PIM)	✓	$((\epsilon, \delta) + T\epsilon)$-(LDP+DP)
LDP(PBA)+DP(PIM)	✓	$\epsilon(2 + T)$-(LDP+DP)
LDP(PIM)+DP(PBA)	−	$(\epsilon(T + 2))$-(LDP+DP)
LDP(LRM)+DP(PBA)	−	$(\epsilon(T + 2))$-(LDP+DP)
LDP(GRR)+DP(PBA)	✓	$((\epsilon, \delta) + 2\epsilon)$-(LDP+DP)
LDP(PBA)+DP(PBA)	✓	(2ϵ)-(LDP+DP)

approaches under temporal correlation. The existing LDP and DP approaches under temporal correlation such as PIM, LRM, GRR, w-event privacy, and quantification methods achieve $T\epsilon$-LDP, $T\epsilon$-LDP, (ϵ, δ)-LDP, $w\epsilon$-DP and α-DP, respectively. Due to more privacy leakage in existing LDP and DP approaches, the combined approaches under existing LDP and DP approaches has more privacy leakage. Hence, the combined approach that adopts the PBA method is a better approach for allocating privacy budgets to temporally correlated data-points.

7.4 Summary

In this work, we present the combined privacy approach by adopting both LDP and DP mechanisms. There are four possible combinations of the combined approach, such as (tra-

ditional LDP+traditional DP), (traditional LDP+DP with TC), (LDP with TC+traditional DP), and (LDP with TC+DP with TC). All these approaches achieves ϵ-(LDP+DP). These combinations may apply to any real-time applications depend on the application's requirements or depends on query requirement. However, it is necessary to analyze the privacy leakage of all cases of combined approach (LDP+DP) and quantify the impact of data correlation on privacy leakage of all cases of combined approach (LDP+DP) in continuous data release settings. In this work, we describe the privacy leakage under temporal correlation of all cases of the combined approach. Then we performed a series of experiments with real and synthetic datasets to determine the average error rate per timestamp for evaluating the data utility of our combined approach (LDP with TC+DP with TC) with other states of the art methods.

Chapter 8

Conclusion and Future Scope

This thesis investigates the design and development of privacy preserving methods which provides a strict privacy guarantee against an adversary who has the knowledge of the correlation either between the users or between the data-points within the user stream.

8.1 Conclusion of the thesis

In this thesis, we quantified the privacy risk in all three privacy preserving models such as Data anonymization, ϵ-Differential privacy and ϵ-Local Differential privacy under correlation. Then we present solutions for preserving privacy against an adversary with bounded and unbounded background knowledge. The chapter 3 to 5 describes each privacy model and proposed a privacy preserving solutions under correlation for preserving users privacy. The chapter 6 compares the privacy leakages under correlation in combined privacy preserving models.

The first work of this thesis is to present a data anonymization approach that prevents users' privacy from four different types of linkage attacks, namely identity, attribute, similarity, and correlated-records linkage attacks.Our data anonymization approach adopts an

existing LK privacy model to fix the upper bound to the adversary's background knowledge and lower bound to the number of unique trajectories in the dataset. Also, we introduced a new privacy threshold called privacy-height to represent the degree of privacy offered to the users. The experimental result shows that the anonymized dataset produced by our proposed method is freed from all four linkage attacks. It shows better performance with a significant reduction in the information loss compared to other states of the art methods.

The second work of this thesis is to presented the definition of differential privacy under temporal correlation in continuous data release settings. The reformulated differential privacy helps to quantify the impact of temporal correlation on privacy leakage in traditional ϵ-DP. Also, we illustrate that the adversary with knowledge of temporal correlation could disclose more privacy leakage than the traditional ϵ-DP. This analysis result shows that the privacy leakage increases over time in w-event privacy when the dataset involves temporal correlation. Therefore, we presented a privacy budget allocation (PBA) method for allocating an adequate amount of privacy budget to each successive timestamp under the protection of ϵ-differential privacy. This method protects any w length user stream that contains temporally correlated data points. Further, we evaluate the average error per timestamp for analyzing the data utility of our proposed method. The result shows that the PBA method's data utility is significantly better than other state-of-the-art methods.

The third work of this thesis is to illustrated the impact of temporal correlation on privacy leakage in ϵ-Local Differential Privacy using a numerical example. This analysis shows that the adversaries with knowledge of temporal correlation can disclose more privacy leakage than the traditional ϵ-LDP. In continuous data release settings, the privacy leakage in w-event privacy increases over time when the temporal correlation is involved in the user stream. Therefore, we proposed a Privacy Budget Allocation (PBA) method for allocating an adequate amount of privacy budget to each successive timestamp under the

protection of ϵ-LDP. It provides a strict privacy guarantee to any w length temporally correlated user stream. Also, we conduct an experiment for evaluating the data utility of our proposed method, and the results show that the proposed method is comparatively better under temporal correlation than the existing state-of-art methods.

Finally the fourth work is to compared the a combined privacy preserving approach by adopting both LDP and DP mechanisms. There are four possible combinations of a combined approach, namely (traditional LDP+traditional DP), (traditional LDP+DP with TC), (LDP with TC+traditional DP), and (LDP with TC+DP with TC). Depends on the query requirement, any combination of a combined approach may apply in order to answer the query. This chapter quantifies the impact of data correlation on privacy leakage of all cases of combined approach (LDP+DP) in continuous data release settings. Further, we conducted series of experiments to determine the average error rate per timestamp for evaluating the data utility of our combined approach (LDP with TC+DP with TC) with other states of the art methods.

In this thesis, we addressed the correlation challenge in privacy preserving models and proposed possible privacy preserving methods under Data anonymization, ϵ-DP, and ϵ-LDP models against the correlation issues. Finally, we evaluate the data utility of all proposed methods by presenting experimental results for real and synthetic data sets.

8.2 Future Scope

- The presented data anonymization approach prevents only four linkage attacks. It is necessary to consider all possible linkage attacks in trajectory data publishing and propose a unified privacy approach that prevents all different types of possible linkage attacks in data publishing.

- The future research direction in chapter 4 is to investigate the impact of temporal correlation on privacy leakage combining with other types of correlation models. Using our methodology, strengthen the previous research that ignored the impact of temporal correlations in continuous data releases.

- In LDP, the server allocates the same privacy budget to each user in order to perturb their data. However, it is unfeasible to use the same privacy budget allocated by the server because different users have different privacy requirements for their data. The future research direction in chapter 5 is to design a personalized privacy budget allocation method under the protection of LDP.

- Effective technology is required when a combined privacy approach deals with mobile crowdsourcing because the data is very large in mobile crowdsourcing.

Bibliography

[1] Chi-Yin Chow and Mohamed F Mokbel. Trajectory privacy in location-based services and data publication. *ACM Sigkdd Explorations Newsletter*, 13(1):19–29, 2011.

[2] Chrysanthi Papoutsi, Julie E Reed, Cicely Marston, Ruth Lewis, Azeem Majeed, and Derek Bell. Patient and public views about the security and privacy of electronic health records (ehrs) in the uk: results from a mixed methods study. *BMC medical informatics and decision making*, 15(1):86, 2015.

[3] Francesco Bonchi. Privacy preserving publication of moving object data. In *Privacy in Location-Based Applications*, pages 190–215. Springer, 2009.

[4] D Malathi, R Logesh, V Subramaniyaswamy, V Vijayakumar, and Arun Kumar Sangaiah. Hybrid reasoning-based privacy-aware disease prediction support system. *Computers & Electrical Engineering*, 73:114–127, 2019.

[5] Benjamin Fung, Ming Cao, Bipin C Desai, and Heng Xu. Privacy protection for rfid data. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1528–1535. ACM, 2009.

[6] Rui Chen, Benjamin CM Fung, Noman Mohammed, Bipin C Desai, and Ke Wang. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, 231:83–97, 2013.

[7] Osman Abul, Francesco Bonchi, and Mirco Nanni. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910, 2010.

[8] Samaneh Mahdavifar, Mahdi Abadi, Mohsen Kahani, and Hassan Mahdikhani. A clustering-based approach for personalized privacy preserving publication of moving object trajectory data. In *International Conference on Network and System Security*, pages 149–165. Springer, 2012.

[9] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[10] Thông T Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053*, 2016.

[11] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In *Advances in neural information processing systems*, pages 2879–2887, 2014.

[12] Zhan Qin, Yin Yang, Ting Yu, Issa Khalil, Xiaokui Xiao, and Kui Ren. Heavy hitter estimation over set-valued data with local differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 192–203, 2016.

[13] Ashwin Machanavajjhala, Johannes Gehrke, and Michaela Götz. Data publishing against realistic adversaries. *Proceedings of the VLDB Endowment*, 2(1):790–801, 2009.

[14] Kai Zhao, Zhen Tu, Fengli Xu, Yong Li, Pengyu Zhang, Dan Pei, Li Su, and Depeng Jin. Walking without friends: publishing anonymized trajectory dataset without leaking social relationships. *IEEE Transactions on Network and Service Management*, 16(3):1212–1225, 2019.

[15] Elahe Ghasemi Komishani, Mahdi Abadi, and Fatemeh Deldar. Pptd: preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. *Knowledge-Based Systems*, 94:43–59, 2016.

[16] Liyue Fan, Li Xiong, and Vaidy Sunderam. Fast: differentially private real-time aggregate monitor with filtering and adaptive sampling. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 1065–1068. ACM, 2013.

[17] Georgios Kellaris, Stavros Papadopoulos, Xiaokui Xiao, and Dimitris Papadias. Differentially private event sequences over infinite streams. *Proceedings of the VLDB Endowment*, 7(12):1155–1166, 2014.

[18] Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao, and Li Xiong. Quantifying differential privacy in continuous data release under temporal correlations. *IEEE transactions on knowledge and data engineering*, 31(7):1281–1295, 2018.

[19] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 715–724, 2010.

[20] Liyue Fan and Li Xiong. Real-time aggregate monitoring with differential privacy. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2169–2173. ACM, 2012.

[21] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4):1–53, 2010.

[22] Benjamin CM Fung, Ke Wang, Ada Wai-Chee Fu, and S Yu Philip. *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, 2010.

[23] Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In *Annual International Cryptology Conference*, pages 36–54. Springer, 2000.

[24] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 439–450, 2000.

[25] Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57, 2004.

[26] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[27] Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 376–385. Ieee, 2008.

[28] Anna Monreale, Gennady L Andrienko, Natalia V Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. Movement data anonymity through generalization. *Trans. Data Privacy*, 3(2):91–121, 2010.

[29] Mehmet Ercan Nergiz, Maurizio Atzori, and Yucel Saygin. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, pages 52–61. ACM, 2008.

[30] Rolando Trujillo-Rasua and Josep Domingo-Ferrer. On the privacy offered by (k, δ)-anonymity. *Information Systems*, 38(4):491–494, 2013.

[31] Noman Mohammed, Benjamin Fung, and Mourad Debbabi. Preserving privacy and utility in rfid data publishing. 2010.

[32] Manolis Terrovitis and Nikos Mamoulis. Privacy preservation in the publication of trajectories. In *Mobile Data Management, 2008. MDM'08. 9th International Conference on*, pages 65–72. IEEE, 2008.

[33] Khalil Al-Hussaeni, Benjamin CM Fung, and William K Cheung. Privacy-preserving trajectory stream publishing. *Data & Knowledge Engineering*, 94:89–109, 2014.

[34] Manolis Terrovitis, Giorgos Poulis, Nikos Mamoulis, and Spiros Skiadopoulos. Local suppression and splitting techniques for privacy preserving publication of trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[35] Xiangwen Liu, Liangmin Wang, and Yuquan Zhu. Slat: Sub-trajectory linkage attack tolerance framework for privacy-preserving trajectory publishing. In *2018 International Conference on Networking and Network Applications (NaNA)*, pages 298–303. IEEE, 2018.

[36] Moein Ghasemzadeh, Benjamin CM Fung, Rui Chen, and Anjali Awasthi. Anonymizing trajectory data for passenger flow analysis. *Transportation research part C: emerging technologies*, 39:63–79, 2014.

[37] Lin Yao, Xinyu Wang, Xin Wang, Haibo Hu, and Guowei Wu. Publishing sensitive trajectory data under enhanced l-diversity model. In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pages 160–169. IEEE, 2019.

[38] Xinxin Liu, Kaikai Liu, Linke Guo, Xiaolin Li, and Yuguang Fang. A game-theoretic approach for achieving k-anonymity in location based services. In *2013 Proceedings IEEE INFOCOM*, pages 2985–2993. IEEE, 2013.

[39] A Ercument Cicek, Mehmet Ercan Nergiz, and Yucel Saygin. Ensuring location diversity in privacy-preserving spatio-temporal data publishing. *The VLDB Journal*, 23(4):609–625, 2014.

[40] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[41] Zhuo Ma, Tian Zhang, Ximeng Liu, Xinghua Li, and Kui Ren. Real-time privacy-preserving data release over vehicle trajectory. *IEEE Transactions on Vehicular Technology*, 68(8):8091–8102, 2019.

[42] Zhibo Wang, Xiaoyi Pang, Yahong Chen, Huajie Shao, Qian Wang, Libing Wu, Honglong Chen, and Hairong Qi. Privacy-preserving crowd-sourced statistical data publishing with an untrusted server. *IEEE Transactions on Mobile Computing*, 18(6):1356–1367, 2018.

[43] Anna Monreale, Wendy Hui Wang, Francesca Pratesi, Salvatore Rinzivillo, Dino Pedreschi, Gennady Andrienko, and Natalia Andrienko. Privacy-preserving distributed movement data aggregation. In *Geographic Information Science at the Heart of Europe*, pages 225–245. Springer, 2013.

[44] Cynthia Dwork. Differential privacy in new settings. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 174–183. SIAM, 2010.

[45] Meng Li, Liehuang Zhu, Zijian Zhang, and Rixin Xu. Achieving differential privacy of trajectory data publishing in participatory sensing. *Information Sciences*, 400:1–13, 2017.

[46] Daiyong Quan, Lihua Yin, and Yunchuan Guo. Enhancing the trajectory privacy with laplace mechanism. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 1218–1223. IEEE, 2015.

[47] Yang Cao and Masatoshi Yoshikawa. Differentially private real-time data release over infinite trajectory streams. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 2, pages 68–73. IEEE, 2015.

[48] Daniel Kifer and Ashwin Machanavajjhala. A rigorous and customizable framework for privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, pages 77–88, 2012.

[49] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1291–1306, 2017.

[50] Bin Yang, Issei Sato, and Hiroshi Nakagawa. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD international conference on Management of Data*, pages 747–762, 2015.

[51] Tianqing Zhu, Ping Xiong, Gang Li, and Wanlei Zhou. Correlated differential privacy: Hiding information in non-iid data set. *IEEE Transactions on Information Forensics and Security*, 10(2):229–242, 2014.

[52] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *NDSS*, volume 16, pages 21–24, 2016.

[53] Xiaotong Wu, Taotao Wu, Maqbool Khan, Qiang Ni, and Wanchun Dou. Game theory based correlated privacy preserving analysis in big data. *IEEE Transactions on Big Data*, 2017.

[54] Rui Chen, Gergely Acs, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 638–649, 2012.

[55] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M Procopiuc, and Divesh Srivastava. Dpt: differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015.

[56] Shuo Wang and Richard O Sinnott. Protecting personal trajectories of social media users through differential privacy. *Computers & Security*, 67:142–163, 2017.

[57] Qian Wang, Yan Zhang, Xiao Lu, Zhibo Wang, Zhan Qin, and Kui Ren. Rescuedp: Real-time spatio-temporal crowd-sourced data publishing with differential privacy. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9. IEEE, 2016.

[58] Qian Wang, Yan Zhang, Xiao Lu, Zhibo Wang, Zhan Qin, and Kui Ren. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 15(4):591–606, 2016.

[59] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309, 2015.

[60] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[61] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[62] Joe Near. Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*, 2018.

[63] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

[64] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30:3571–3580, 2017.

[65] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[66] Tianhao Wang, Ninghui Li, and Somesh Jha. Locally differentially private frequent itemset mining. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 127–143. IEEE, 2018.

[67] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*, pages 729–745, 2017.

[68] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Marginal release under local differential privacy. In *Proceedings of the 2018 International Conference on Management of Data*, pages 131–146, 2018.

[69] Zhikun Zhang, Tianhao Wang, Ninghui Li, Shibo He, and Jiming Chen. Calm: Consistent adaptive local marginal for marginal release under local differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 212–229, 2018.

[70] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 441–459, 2017.

[71] Matthew Joseph, Aaron Roth, Jonathan Ullman, and Bo Waggoner. Local differential privacy for evolving data. *Advances in Neural Information Processing Systems*, 31:2375–2384, 2018.

[72] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[73] Rui Chen, Haoran Li, A Kai Qin, Shiva Prasad Kasiviswanathan, and Hongxia Jin. Private spatial data aggregation in the local setting. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 289–300. IEEE, 2016.

[74] Xingxing Xiong, Shubo Liu, Dan Li, Zhaohui Cai, and Xiaoguang Niu. Real-time and private spatio-temporal data aggregation with local differential privacy. *Journal of Information Security and Applications*, 55:102633, 2020.

[75] Yang Cao and Masatoshi Yoshikawa. Differentially private real-time data publishing over infinite trajectory streams. *IEICE TRANSACTIONS on Information and Systems*, 99(1):163–175, 2016.

[76] ASM Hasan, Qiang Qu, Chengming Li, Lifei Chen, and Qingshan Jiang. An effective privacy architecture to preserve user trajectories in reward-based lbs applications. *ISPRS International Journal of Geo-Information*, 7(2):53, 2018.

[77] Emre Kaplan, Thomas B Pedersen, Erkay Savaş, and Yücel Saygın. Discovering private trajectories using background information. *Data & Knowledge Engineering*, 69(7):723–736, 2010.

[78] Hongtao Li, Jianfeng Ma, and Shuai Fu. Analyzing mechanism-based attacks in privacy-preserving data publishing. *Optik-International Journal for Light and Electron Optics*, 124(24):6939–6945, 2013.

[79] David J Martin, Daniel Kifer, Ashwin Machanavajjhala, Johannes Gehrke, and Joseph Y Halpern. Worst-case background knowledge for privacy-preserving data publishing. *arXiv preprint arXiv:0705.2787*, 2007.

[80] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, technical report, SRI International, 1998.

[81] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. \ell-diversity: Privacy beyond\kappa-anonymity. In *null*, page 24. IEEE, 2006.

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

[83] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 229–240. ACM, 2006.

[84] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[85] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. T-drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):220–232, 2011.

[86] Cynthia Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.

[87] Haoran Li, Li Xiong, Xiaoqian Jiang, and Jinfei Liu. Differentially private histogram publication for dynamic datasets: an adaptive sampling approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1001–1010. ACM, 2015.

[88] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

[89] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 316–324. ACM, 2011.

[90] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, pages 99–108. ACM, 2010.

[91] Annalisa Cocchia. Smart and digital city: A systematic literature review. In *Smart city*, pages 13–43. Springer, 2014.

[92] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[93] Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A McCann, and S Yu Philip. Lopub high-dimensional crowdsourced data publication with local differential privacy. *IEEE Transactions on Information Forensics and Security*, 13(9):2151–2166, 2018.

[94] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*, volume 10, 2011.

[95] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM conference on embedded networked sensor systems*, pages 85–98, 2009.

[96] Saikat Guha, Alexey Reznichenko, Kevin Tang, Hamed Haddadi, and Paul Francis. Serving ads from localhost for performance, privacy, and profit. In *HotNets*, 2009.

[97] Ke Gu, Lihao Yang, and Bo Yin. Location data record privacy protection based on differential privacy mechanism. *Information Technology and Control*, 47(4):639–654, 2018.

[98] Tao Wang, Zhigao Zheng, and Mohamed Elhoseny. Equivalent mechanism: Releasing location data with errors through differential privacy. *Future Generation Computer Systems*, 98:600–608, 2019.

[99] D Hemkumar, S Ravichandra, and DVLN Somayajulu. Impact of prior knowledge on privacy leakage in trajectory data publishing. *Engineering Science and Technology, an International Journal*, 23(6):1291–1300, 2020.

[100] Cynthia Dwork. Differential privacy in new settings. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 174–183. SIAM, 2010.

[101] D Hemkumar, S Ravichandra, and DVLN Somayajulu. Impact of data correlation on privacy budget allocation in continuous publication of location statistics. *Peer-to-Peer Networking and Applications*, pages 1–16, 2021.

List of Publications

Publications from the thesis

Journal papers:

1. **Hemkumar D**, S Ravichandra and DVLN Somayajulu, ”Impact of prior knowledge on privacy leakage in trajectory data publishing”, *Engineering Science and Technology, an International Journal, Elsevier*, 23(6), pp.1291-1300, 2020, DOI: https://doi.org/10.1016/j.jestch.2020.06.002

2. **Hemkumar D**, S Ravichandra and DVLN Somayajulu, ”Impact of data correlation on privacy budget allocation in continuous publication of location statistics”, *Peer-to-Peer Networking and Applications, Springer*, pp.1-16, 2021, DOI: https://doi.org/10.1007/s12083-021-01078-6

3. **Hemkumar D**, S Ravichandra and DVLN Somayajulu, ”Impact of Data Correlation on Privacy Budget Allocation in Local Differential Privacy for Continuous Data Release Settings”, *Performance Evaluation, Elsevier*, (*Under review*).

4. **Hemkumar D**, S Ravichandra and DVLN Somayajulu, ”Compare the Impacts of Data Correlation on Privacy Leakage in a Combined Privacy Preserving Approach”, *International Journal of Security and Networks, Inderscience*, (*Under review*).