Georgia State University

ScholarWorks @ Georgia State University

Computer Science Dissertations | Department of Computer Science

Fall 8-8-2023

User-centric privacy preservation in Internet of Things Networks

Akshita Maradapu Vera Venkata Sai

Follow this and additional works at: https://scholarworks.gsu.edu/cs_diss

Recommended Citation

Maradapu Vera Venkata Sai, Akshita, "User-centric privacy preservation in Internet of Things Networks." Dissertation, Georgia State University, 2023.
https://scholarworks.gsu.edu/cs_diss/203

This Dissertation is brought to you for free and open access by the Department of Computer Science at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Computer Science Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

User-centric privacy preservation in Internet of Things Networks

by

Akshita Maradapu Vera Venkata Sai

Under the supervision of Yingshu Li, Ph.D. and Zhipeng Cai, Ph.D.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

Recent trends show how the Internet of Things (IoT) and its services are becoming more omnipresent and popular. The end-to-end IoT services that are extensively used include everything from neighborhood discovery to smart home security systems, wearable health monitors, and connected appliances and vehicles. IoT leverages different kinds of networks like Location-based social networks, Mobile edge systems, Digital Twin Networks, and many more to realize these services.

Many of these services rely on a constant feed of user information. Depending on the network being used, how this data is processed can vary significantly. The key thing to note is that so much data is collected, and users have little to no control over how extensively their data is used and what information is being used. This causes many privacy concerns, especially for a naïve user who does not know the implications and consequences of severe privacy breaches.

When designing privacy policies, we need to understand the different user data types used in these networks. This includes user profile information, information from their queries used to get services (communication privacy), and location information which is much needed in many on-the-go services. Based on the context of the application, and the service being provided, the user data at risk and the risks themselves vary. First, we dive deep into the networks and understand the different aspects of privacy for user data and the issues faced in each such aspect. We then propose different privacy policies for these networks and focus on two main aspects of designing privacy mechanisms: The quality of service the user expects

and the private information from the user's perspective. The novel contribution here is to focus on what the user thinks and needs instead of fixating on designing privacy policies that only satisfy the third-party applications' requirement of quality of service.

INDEX WORDS: User privacy, Location privacy, User check-ins, User motivation, Location Based Services, Location-Based Social Networks, Mobile Social Networks

Copyright by
Akshita Maradapu Vera Venkata Sai
2023

User-centric privacy preservation in Internet of Things Networks

by

Akshita Maradapu Vera Venkata Sai

Committee Chair:	Yingshu Li
Committee:	Zhipeng Cai
	Wei Li
	Jun Kong

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2023

DEDICATION

I'd like to dedicate this dissertation to my parents Maradapu Venkata Ashok Vardhan and Maradapu Venkata Avanija. They have played a pivotal role in my early years and always encouraged me to build my own path. Without their support and innumerable sacrifices, I would have never been the person I am today.

My brother, Maradapu Abhay Kumar, was a source of invaluable strength during my research journey. He has been there, always offering encouragement and being a sounding board to provide fresh perspectives. His relentless optimism and belief in me were the fuel that kept me going.

Last but not least, my husband, Nisheeth Bandaru. He was my biggest support system, ally and calm in the middle of chaos. He has supported me in my decision to apply for this Ph.D. program and has been my biggest cheerleader ever since. His dedication to my dreams, patience, and unwavering faith in me has made this dissertation possible. I am truly grateful to have him in my life.

ACKNOWLEDGMENTS

I am thankful to Georgia State University and feel blessed to have been given an opportunity to pursue my graduate studies here. The past several years at this university have shaped my life and given me many memories I will cherish for the rest of my life. This dissertation wouldn't have been possible without the guidance of my advisors and the support and encouragement from seniors and friends.

I would like to thank my advisors, Dr. Yingshu Li and Dr. Zhipeng Cai. Dr. Li has been one of the most compassionate people I ever met. She has always been very patient and has gone above and beyond to provide me with every opportunity that would aid my career. Her guidance was not just limited to research but to all the facets of my life, and she always focused on giving me complete education centered around overall development. I am grateful to have Dr. Cai as one of my advisors. He has always encouraged me to collaborate with other researchers, build connections, and push me to reach my full potential. His commitment to excellence and dedication to imparting knowledge has profoundly shaped my intellectual journey.

I would also like to thank my committee members, Dr. Wei Li, for her attention to detail and valuable guidance, and Dr. Jun Kong, for his time, support, and feedback that contributed to my dissertation.

I express my gratitude to Dr. Rajshekar Sunderraman, for his guidance during my entire graduate studies and Dr. Daniel Takabi, for his support during the final years of my Ph.D.

program.

I am blessed to be a part of a huge research group, which has allowed me to learn from my seniors, Dr. Madhuri Siddula, Dr. Yan Huang, and Dr. Saide Zhu. I thank my teammates Zuobin Xiong, Chenyu Wang, Kainan Zhang, Prajwal Panzade, Euiseong Ko, Honghui Xu, Jinkun Han, Shatha Alharty, and Qasim Zia. Group research meetings with them have always been a highlight of my week, and these interactions have contributed to exciting discussions and collaborations and provided a fresh perspective on my research.

Last but not the least, I thank Tammie Dudley, Jamie Hayes and all the members in the department, for your help and support in this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- MSN - Mobile Social Networks
- LBSN - Location Based Social Networks
- LBS - Location Based Services
- IoT - Internet of Things
- IIoT - Industrial Internet of Things
- MES - Mobile Edge Systems
- DTN - Digital Twin Networks
- FL - Federated Learning

CHAPTER 1
Introduction

1.1 Background and Motivation

The past few decades have seen a boom in the development and production of mobile and smart devices and communication technologies. This has led to a new world of connected devices called the Internet of Things (IoT). IoT aims to provide users with end-to-end service, continuous connectivity, lower latency, and an overall better quality of service [1, 2, 3, 4, 5].

Since IoT is so widespread, it needs to leverage several different networks [6, 7] to provide each service as shown in Figure. 1.1. For example, if Alice wants to let people know where she is headed while traveling in real-time, she uses location-based social networks and Mobile Edge Systems (MES) [8]. Similarly, if Alice uses a navigation system to find out the best way to overcome traffic and reach work soon, she will use just the Mobile Edge System [9]. Therefore, one can say that it is tough to clearly tell where one network ends, and the other begins in terms of the workflow.

The fuel for all these networks and services is user-generated data. This can be his location information or his profile information, or even the requests that he generates to get some results from a search engine [10, 11]. With networks requiring so much user data constantly, a user, in most cases, is entirely unaware of what data is being used and how it is being used to provide these said services. This raises many privacy concerns for the user, especially when they do not understand how a breach of their information may adversely

Figure 1.1 Different IoT applications

affect their life. Also, as mentioned earlier, since IoT is an amalgamation of several networks, each modeled quite differently, it poses many challenges in developing comprehensive privacy policies [12, 13, 14, 15, 16]. A lack of privacy-preserving architectures allows adversaries to access all of the user's sensitive data. Following are a few possible privacy threats to users:

1. **Identity theft:** This includes exploiting a user's personal information to create impersonation profiles on social networks or to steal monetary possessions. One such incident was the data breach occurred in Equifax in 2016 [17], that led to a leak of user's sensitive credit card information like their social security number, name, credit information and much more. This data breach led to a lot of counterfeited transactions that brought upon losses to both users and the company. In [18], Sweeney, showed how a released anonymized social data can still be analyzed to trace back to a specific individual.

2. **Location tracking:** With a lot of location-based applications that store user location

information and share them publicly, outsiders can obtain information about a user's daily activities and trajectories easily. This can then be further analyzed to profile users and launch many other attacks [19].

3. **Fake profiles:** This is a very common occurrence when it comes to platforms like Facebook or Instagram where a lot of fake profiles are created with the intention to contact people and lure them into giving away sensitive information [20]. These profiles sometimes leave back viruses and trojans on the devices that are used to open messages from the said profile.

4. **Malicious links:** These links are usually shared through messages and emails, which leads the recipients to phishing websites and trick them into giving away sensitive information like passwords or credit card information, or automatically download a pretend software onto the device which might launch certain background attacks. Most times the users are completely unaware of these attacks.

5. **Hacking:** This is the most common and popular issue in social networks, where adversaries, also known as "Black Hats" [21] can make their way into a user's account and procure complete control over it.

1.2 Types of Privacy based on User data

The best way to completely address the different privacy issues is to first completely understand the underlying user data. In all IoT networks, user's data and their corresponding privacy can be broadly classified into three categories:

1. **Location Privacy**: User's location is extremely important for many services provided by an MSN. Preserving user's location information without affecting the quality of service is crucial, and these location privacy-preserving solutions are highly dependent on the application [22, 23, 24]. A few applications like share rides require a single location while other applications like navigation systems, require continuous location inputs. Solutions for location privacy-preservation include perturbation [25], obfuscation [26], k-anonymity [27], spatial cloaking [28] and temporal cloaking [29] to name a few.

2. **User Attribute privacy**: Users in an MSN can be considered as nodes of an MSN graph. Therefore, user information includes both node information and the information regarding the links they establish within the network [30, 31]. The goal of user privacy-preservation solutions is to preserve both node and link information connected to the user [32]. Node and edge perturbation [33], and anonymization [34] are a few common methods used in designing user privacy-preserving solutions [35].

3. **Communication privacy**: This mainly focuses on preserving the content and the context of the information being exchanged in a network or uploaded to the server and other third-party applications. This information may include message contents, a few attributes of user's profile information, queries made to the server that may include sensitive information like userID or user location, and many more. Most communication privacy-preserving mechanisms use session key agreements and digital signatures [36], authentication and verification schemes [37], and different encryption methods [38, 39].

1.3 Organization

The rest of the dissertation is organized as follows: Chapter 2 provides the network model. Chapter 3 covers the different aspects of user data that can be vulnerable to different privacy attacks when published online and a literature review of existing research to address these privacy concerns. Chapter 4 investigates the privacy concerns of check-in data in Location Based Social Networks and how to address them. Chapter 5 introduces the future research directions for this work and finally in Chapter 6, we will summarize the proposed research.

CHAPTER 2

Network Model

This chapter provides a clear understanding of the different components of a Mobile Social Network (MSN) and how to mathematically formulate these components and the network features

2.1 MSN Components and Architecture

2.1.1 Components in MSN

A *MSN* is a dynamic environment, therefore, its components are constantly going through changes in states and the connections they form. Thus, an MSN at any given point can have several types of components like users, groups and communities, different mobile infrastructures like Bluetooth, cellular data, or the basic internet. The main components of an MSN can be narrowed down to the following:

1. Nodes or Mobile Devices: These include all devices of mobile nature that the end-users might use to connect to an MSN. It can be a mobile phone, a tablet, laptop, or any device that can establish a Bluetooth connection and is embedded with sensors.

2. Network: This includes the infrastructure or the framework that is used to establish connections and serve as a communication medium among end-users and between end-users and servers. The infrastructure might include servers, routers, access points, cellular base stations, and also cloud infrastructure if the services being provided are hosted on the cloud.

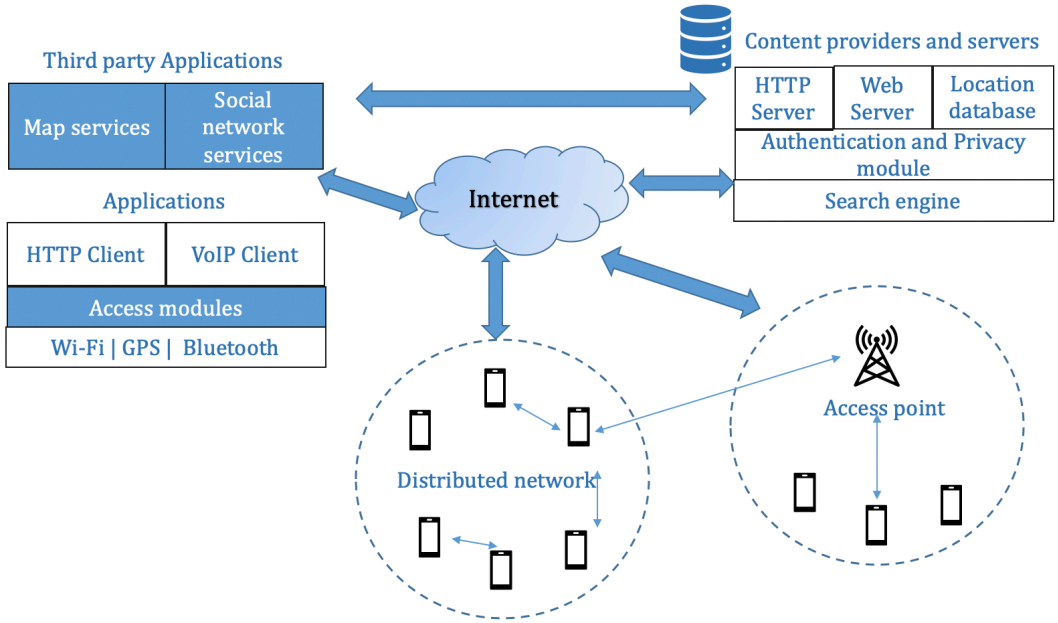


Figure 2.1 MSN Architecture

3. Content Providers and Servers: Content providers include all the servers that the users may directly or indirectly send requests to and also other third-party servers that make content available to end-users.

2.1.2 Architecture

Depending on the way the users communicate with each other and with the server, MSNs can either have a centralized, decentralized, or hybrid architecture. Most MSNs today follow a hybrid architecture as shown in Fig. 2.1, because they are either mobile versions of older web applications or they want to exploit the benefits of the hybrid architecture. The centralized aspect of hybrid architecture provides benefits like simple implementation, reduced hardware, low maintenance costs, and high efficiency while the decentralized or distributed aspect provides benefits of low reliability and stability requirements, better traffic load balancing,

lower latencies and presence of alternate sources of data in case of a server failure.

2.2 MSN as a Graph and Social Features

In order to resolve the different privacy issues, one should clearly understand the structure of MSN and it's social aspects from a technical stance. This can be done by formulating MSN as a graph. For simplicity let's assume that the graph is undirected, and is denoted as $G = (K, W)$. The users of an MSN are denoted by nodes $\mathbf{K} = \{1,2,\ldots, i,\ldots, k\}$. The edges represents connections between two nodes and the weight of each edge can be obtained as follows,

$$w_{ij} = \frac{F_{ij}}{T} \tag{2.1}$$

where F_{ij} is the frequency of communications between two nodes i and j within a given time period T. In order to derive the other features, we need an adjacency matrix. This adjacency matrix will help us describe the nature interactions among users in the network. As we have an undirected graph, the adjacency matrix $\mathbf{A}$, will be symmetric and it can be obtained as follows,

$$A_{ij} = \begin{cases} 1, & \text{if } w_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

We will now define and formulate the social features of an MSN.

2.2.1 Node Degree

The degree of a node tells us the number of connections the current node (user) has with other nodes (users) within the graph G (network). The **degree** of a node i can be obtained as follows,

$$d_i = \sum_{j=1}^{k} A_{ij} \tag{2.3}$$

2.2.2 Node Strength

Node Strength gives information about the frequency of communications between a node (user) and the other nodes (users) in the graph G (network). The **node strength** of a node i can be obtained as follows,

$$s_i = \sum_{j=1}^{k} A_{ij} w_{ij} \tag{2.4}$$

2.2.3 Clustering Coefficient

This feature of the graph sheds light on how tightly social groups and communities are connected. Higher the clustering coefficient, greater the aggregation relationship between a node and its surrounding nodes. This feature further helps us define social groups and virtual communities. If d_i is the node degree of node i then the number of edges (N_i), between node i and the other nodes is given by,

$$N_i = \frac{d_i(d_i - 1)}{2} \tag{2.5}$$

The **local clustering coefficient** of node i can be obtained as follows,

$$C_i = \frac{\sum_{j=1}^{k} w_{ij}}{N_i} \tag{2.6}$$

The **global clustering coefficient** can be calculated as an average of local clustering coefficients of all nodes and can be obtained as follows,

$$C = \frac{\sum_{i=1}^{k} C_i}{|K|} \tag{2.7}$$

2.2.4 Betweeness Centrality

In a network, there is always exchange of information between a source **S** and destination **D**. In the absence of a direct path between **S** and **D**, the information passes through some node i. A few nodes always act as intermediate nodes and few would never be an intermediate node. This feature helps us compare the importance of a few nodes over the others in an MSN. A greater betweeness centrality implies that more available paths between any source and destination are passing through node i. The betweeness centrality of a node i can be obtained as follows,

$$b_i = \frac{\sigma_i}{2\sigma_{S,D}} \tag{2.8}$$

where σ_i is the number of paths that pass through i and $\sigma_{S,D}$ is the number of paths between source S and destination D.

2.2.5 Closeness Centrality

This feature indicates the amount of interaction of a node with the other connected nodes in the network. A greater value indicates higher number of interactions and more frequent communications among the nodes. For node i, the **closeness centrality** can be calculated as follows,

$$C_c(i) = \frac{k-1}{\sum_{j=1}^{k} d(ij)} \tag{2.9}$$

where $d(ij)$ is the length of the path between node i and node j.

CHAPTER 3

Analysis of User Privacy issues in MSN

In this chapter we will go over the threats and the privacy preserving solutions for different user data mentioned in Section 1.2.

3.1 Known Threats to User data in MSN

Advancements in MSN technology have made more resources available to adversaries to design attacks that pose threats to different aspects of an MSN. It is vital to understand the nature of these attacks and their underlying mechanism to design privacy-preserving solutions. In this section, we categorize the threats to location, user, and communication privacy in MSNs.

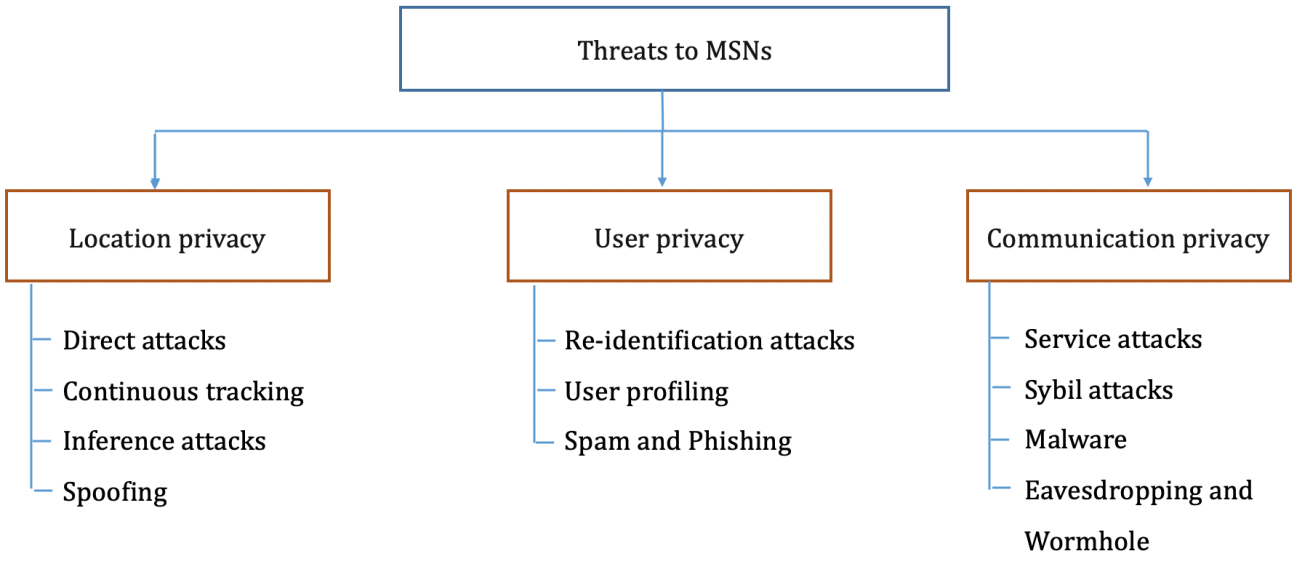


Figure 3.1 Threats to privacy in MSNs

3.1.1 Threats to Location data

1. **Direct sharing attacks:** These attacks happen when users share their location directly on social platforms like Facebook and Foursquare as check-ins or Geo-tagging. On these platforms, attackers have direct access to user's check-in history, therefore, they need not build a model to extract users' location information. For this reason, these attacks are passive in nature. Due to the availability of original location data, the attack models, in these cases are easy to build and are very effective. [40][41] describe extracting location information and launching location-based attacks.

2. **Continuous tracking:** This is a more active approach compared to direct sharing attacks. These attacks are designed to geo-locate a user without any prior information and they are based on indirectly shared locations like the obfuscated location[42, 43]. Sharing obfuscated locations is very popular in Facebook Marketplace and Wechat [41]. As these are popular applications, the amount of obfuscation used is publicly available. This allows attackers to create a model to reverse engineer the exact location and study the user's movement patterns. These are known as **spatial knowledge attacks** [44]. Another type of continuous tracking happens in neighborhood-discovery services, where the users might want to explore places or events in the neighborhood. Here, the user's location is sent to the LBS (Location based server) as part of a search query. This location can then be inferred either by **direct query sampling** [45] or by sending **multiple queries** to identify the exact location of the user [19].

3. **Inference attacks:** Inference attacks are defined as the analysis of data to gain

knowledge of the subject unlawfully. These attacks are more common as most mobile applications have location sharing capabilities and do not impose strong privacy mechanisms to protect this information [46]. In [47, 48], the authors were able to perform an inference attack on GPS data gathered from their volunteers. They were able to exactly identify and locate the homes of these volunteers by analyzing the GPS data and segmenting it into discrete trips. They then implemented four heuristic algorithms to identify the exact location of the subject's home.

4. **Spoofing:** One of the most common types of spoofing is **geo-location spoofing** [49], where a user fakes his location with malicious intent and is primarily done by either using VPNs or a DNS Proxy. **GPS spoofing** is another type of spoofing that aims at deceiving users by broadcasting incorrect GPS signals, delaying GPS signals, or re-sending these signals in a different zone.

3.1.2 Threats to User attribute data

1. **Re-identification attacks:** These attacks use multiple data sources and common "quasi-identifiers" in these sources to uniquely identify a user [50]. Several studies show how public anonymized medical data can be used to uniquely identify patients [51]. For example, in [18], the author was able to uniquely identify about 87% of users by combining the medical and voter registration data and just three identifiers. Similarly, authors in [52] were able to identify users from four spatio-temporal points. In [53], the authors designed a re-identification attack which uses points of a user's mobility trace obtained from a trace dataset like GeoLife and Cabspotting, to form a

heat-map structure. From the GeoLife trace set, they were able to uniquely identify about 80% of the users.

2. **User profiling:** In MSNs, users are either matched to other users or events based on the attributes they share, like user location or interests. This matching process sometimes requires users to publish their information, which may include sensitive data. The attackers can also study user behaviors online, from their check-in data, posts, and many more [54]. All the leaked information combined with the availability of powerful data mining and analysis techniques allows attackers to profile users or gather their information [55, 16]. The most common way of misusing user information is the creation of fake profiles, impersonation and identity theft, which by extension can contribute towards launching targeted attacks [56].

3. **Spam and Phishing attacks:** Spam and phishing attacks work together and are a common way of obtaining user's confidential information. These attacks work by sending legitimate looking emails or messages that lead users to a website with some embedded logic to extract all the information entered. While spam directs users to a third-party website, phishing attacks directs users to an almost perfect looking website. These attacks have become increasingly common in MSNs like WhatsApp, LinkedIn, Facebook, Twitter, and many more. On platforms like Facebook and Twitter, attackers impersonate famous brands, celebrities, and sometimes customer support. All of these pages lead users to a legitimate looking official website and either ask users to answer a few questions or to enter their information in a form on the website [57]. Another

famous spam was on WhatsApp, where a message was circulated about 'offering a free pair of Adidas trainers to celebrate their 93rd anniversary'. Similar to the previous case, the message had a link that directed the users to legitimate looking Adidas webpage where they were asked to enter their credit card information [58][59].

3.1.3 Threats to Communication data

1. **Eavesdropping and wormholes:** Eavesdropping is when an attacker intercepts the communication between two parties without their knowledge. In wired communications this attack happens over the network but, when it comes to MSNs to launch this attack, a malicious code is embedded into the application. There are several social networks and other applications on App Store and Google Play like Facebook that is programmed to listen to audio through the mobile device's microphone. These apps listen to the audio from television shows or ads and mine the audio data to send targeted advertisements to the users [60][61]. Sometimes, social networks like Facebook also listen to users' conversations as part of social media monitoring and send targeted advertisements to them [62].

   A wormhole attack, on the other hand, is where an attacker gets packets from one point, tunnels them to another point and either replays them later or directs them to a different location. These attacks make use of users' network IDs and can further lead other attacks like Man-in-the-middle (MITM), where the recorded packet information may be changed and forwarded or to replay attacks. These attacks are designed to disrupt routing protocols or other security protocols being used in MSNs [63]

2. **Service attacks:** These attacks aim at making a resource unavailable or disrupting the service. Denial of Service (DoS) and Replay Attacks are examples of such attacks. DoS or DDoS (Distributed Denial of Service) is a cyber-attack that infects multiple devices by injecting malware to attack servers. Several MSN applications like Facebook, LinkedIn, Uber, Airbnb, banking applications, and e-commerce applications are vulnerable to this attack due to the ease with which an attacker can profile a user. A Dos attack on a mobile phone happens through an application that is downloaded onto the device. This application either directly performs a DDoS attack or opens up a security loophole in the way that the attacker has complete control over the device. This attack primarily reduces revenue to companies by blocking network traffic and incurs additional costs to them to overcome the issue. WireX botnet [64] is one such application that was dubbed as an "Android Clicker" and, affected over 120,000 Android devices and conducted massive DDoS attacks in the application layer. Another application, Mirai botnet affected a lot of social networks like Amazon and Twitter [65]. A replay attack is a network attack where valid communication information is collected and then replayed or delayed. It is a version of the Man-in-the-middle attack.

3. **Malware:** In these attacks, users are directed to a website that either automatically downloads malicious code or requests users to download a supporting media player or an application or enable access to cookies to continue viewing the page. This in fact is some malicious code that when downloaded and/or installed allows the attackers to control mouse and keyboard activities of the infected device. The distribution of

malware in MSNs happens through fake profiles [66, 67] or broadcasted messages that show up directly in the user's inbox [68].

4. **Sybil Attacks:** In this type of attack, a malicious user also known as Sybil may create several fake identities to gather information about users and attack the communication network itself. These attacks are particularly prevalent in MSNs due to its open and distributed architecture. Sybils are then used to launch phishing and DoS attacks to distribute malware. One of the major attacks by sybils is on routing protocols, where the sybils place themselves in a way, such that several individual paths between different sources and destination pass through them [69]. In another attack, Sybils establish connections with other Sybils and honest nodes and then start disseminating spam, advertisements, and malware to violate user privacy. Additionally, sybils can generate different reviews to favor their services or undermine other services, and this is done by focusing on some specific behaviors and repeating them at high frequency [70][71, 72].

3.2 Proposed Solutions

Extensive research has been conducted in recent years to address and resolve privacy issues in MSNs Ṫhese solutions have a few similarities with respect to concept, technique or the feature of MSN being preserved. To carry out further research in this field, it is imperative to have a clear understanding of the work done so far. Therefore, in this section, we provide an elaborate classification of privacy-preserving solutions as shown in Fig. 3.2 and summarize them.

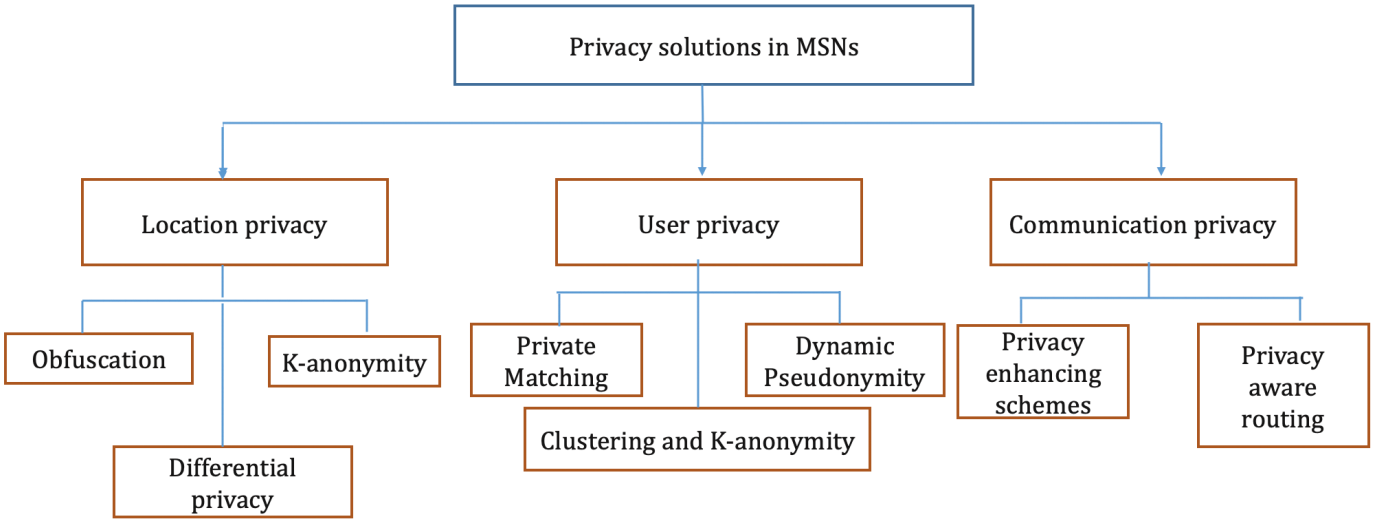


Figure 3.2 Privacy Solutions in MSNs

3.2.1 Location Privacy

Based on the technique used in location privacy-preserving mechanism, the solutions can be categorized into :

1. *K-anonymity based schemes*
2. *Obfuscation based schemes*
3. *Differential privacy based schemes*

3.2.1.1 K-anonymity based schemes

K-anonymity is a popular technique that is used in several privacy solutions and was first proposed in [18]. According to [18], a scheme is k-anonymous if the probability of uniquely identifying a particular entity from k entities is at most $1/k$. Several solutions in this area that use k-anonymity have been proposed and they either considered Online Social Networks

(OSNs) or involved trusted third parties (TTP). Also, most MSNs store the user information and location information on separate servers, which then requires some sort of encryption to safely link the two servers and provide accurate query results. This type of distributed architecture increases the risk of information leakage [73].

In [74] authors proposed a mechanism called **CenLocShare** to address the above mentioned issues. Firstly, to overcome the issues caused by having two different servers was solved by combining the Social Network Server (SNS), and Location Based Server (LBS), into a single server called Location Storing Social Network Server (LSSNS). Then, the authors identify scenarios where the user might send his location data to LSSNS For each scenario, the user submits a query to LSSNS, during which he sends his location along with (k-1) dummy locations, thus using k-anonymity. The main contributions of this work are; it provides a centralized scheme by using a single server, which reduces the risk of information leakage, it designed a scenario-specific LBS query processing instead of a generalized solution, uses "Sequence ID" in queries to prevent replay and tampering attacks and finally, reduces the storage requirement and the time needed to process queries compared to other mechanisms. Although the scheme provides several benefits, the centralized approach increases the computation complexity as the network increases. Also, the scheme only preserves the privacy of a single location. If the user submits continuous queries it might be relatively easy for an attacker to map the trajectory of the user, as the radius within which the dummy locations are generated is fixed.

As mentioned in the previous solution, queries sent to the server may increase the risk

of information leakage. Therefore, in situations where the user sends continuous queries to the server, there is an increased risk to the location information. To address this, a solution called **Collaborative trajectory privacy-preserving scheme** was proposed in [75]. The basic idea of the scheme is to preserve location privacy by reducing the number of queries being made to the LBS by exploiting the caching ability of the user devices in the network. The scheme contains two algorithms :

1. **Multi-hop caching aware cloaking**: This allows the user to communicate within a H_{max} (maximum hop distance) by sending a collaboration request. The users who respond to this request share their cached information. Based on this information, users can create a k-anonymous cloaking region and also locally obtain query results for what should be the next location, instead of sending a query to the LBS. The scheme provides different versions of this algorithm for the requestor and receiver of the "collaboration request".

2. **Collaborative privacy-preserving querying**: This allows users to obtain information locally from the data cached and shared by other users, or it can send a query to a remote LSP. In order to obtain accurate results, the algorithm checks for the freshness of cached data. It also allows users to send fake queries to the LBS with the farthest location in its cloaking region, to introduce confusion.

The scheme can be used in both static and continuous querying scenarios. The k-anonymity in creating the cloaking regions and the confusion introduced by the fake queries provide two-fold privacy preservation. But, the major disadvantage of this scheme is that

the algorithms are computationally intensive, and may not be feasible for a mobile device in MSNs due to their limited computation capacities.

In [27], a solution similar to [75] called Privacy Preserving System (PPS) was proposed. This solution overcomes the computation challenges faced in [75]. To minimize queries being made to the LBS, PPS is used to maintain a single cache instead of multiple caches, with all the frequent location requests and their corresponding results. When a user makes an LBS request, the PPS checks if the location in the query meets either of the following conditions and then returns the result from the cache :

- A request is being made within a small acceptable distance or,
- A request is being made from a place that is a subarea of already cached locations.

If none of the above conditions are met by the location, the PPS directs the request to LSP by obfuscating the user's location. The obfuscation radius is selected in a way that the region is k-anonymous. If the region has sparse users, then dummy users are added to make it k-anonymous. Compared to all the other solutions, this scheme, firstly, provides privacy to continuous queries, thus preserving the trajectory of the user. Secondly, it overcomes the computation issues presented by the previous two schemes, by using a single cache to store all the results. This not only reduces the amount of data that needs protection but also reduces the number of communication exchanges and requests between users, resulting in less power consumption. Finally, the solution provided is more practical as it considers real-world user distributions, and how it is not always possible to have enough users to provide k-anonymity and proposes a way to overcome this obstacle as well.

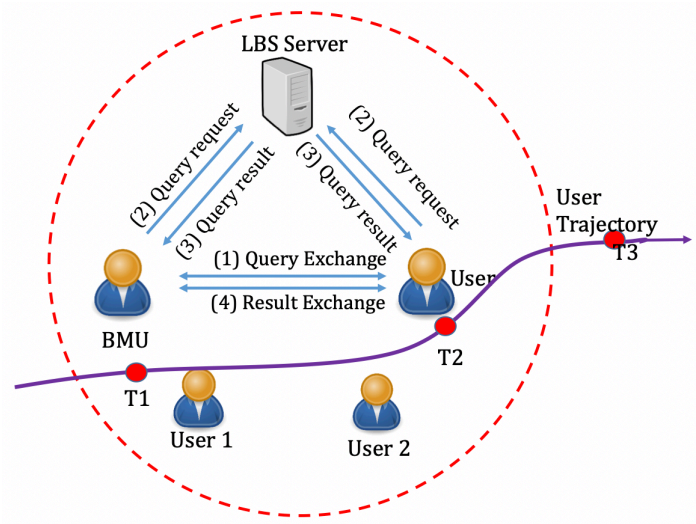


Figure 3.3 DQE Mechanism

In the recent years, the number of users using MSNs has increased, therefore, when solutions based on k-anonymity are implemented in practical scenarios, if the cloaking region is small, then the tendency of disclosing location information is more. Therefore, we need solutions based on other techniques like **obfuscation** which are not subject to user distribution.

3.2.1.2 Obfuscation based schemes:

In [76], the authors proposed a privacy preserving scheme called **Deviation based Query Exchange (DQE)** to preserve the user's trajectory data. The scheme preserves privacy at the user level and has the following steps.

1. **Step 1:** Finding the Best Matching User (BMU).

2. **Step 2:** Deviation based Query Exchange (DQE).

In **Step 1** user **U** is matched with other users and the best match possible is identified. During this step, to ensure user privacy, private matching [Section 3.2.2.2], is performed by introducing confusion to the user's original information $(x, y, movement\ direction)$. A similarity value is calculated to see how similar a user is to others, and the user who is the least similar to **U** is selected as the **BMU**. In **Step 2** shown in Fig. 3.3, **U** exchanges his ID with the **BMU**. The BMU will then forward U's location query to the LSP by obfuscating the location and the obtained results are exchanged later. This solution is extremely well rounded and thought through because, firstly, the obfuscation happens at the user level, instead of at a central point, thus avoiding having a single point of failure/attack. Secondly, even if an attacker gets the ID, it is difficult to link a user to the ID because the BMUs keep changing as the user's moves. Thirdly, the solution provides an additional level of privacy by obfuscating the location before submitting the query to the LSP. Finally, the query results are encrypted using asymmetric encryption making it resistant to eavesdropping attacks. Fig. 3.4 shows how **DQE** preserves a user's location trajectory. But, the authors do not consider the varying speeds of the users, making it difficult to apply this solution to a more practical scenario, and they do not discuss how often the scheme finds BMUs. Finally, as the solution is carried out at the user level, these calculations and multiple communication exchanges may drain the limited power capacity of the mobile devices.

To overcome the drawbacks of the DQE scheme, [77] proposed a model called **Smart-MASK** which machine learning to build a fine-grained location privacy system. The model works as follows :

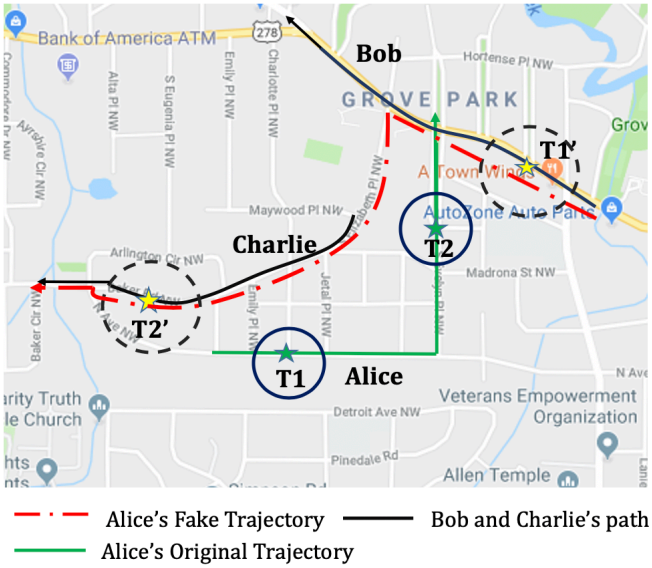


Figure 3.4 BMU Selection in DQE

1. A clustering algorithm is used to generate the user's location profile by using the user's mobility history and location profiles.

2. Each check-in in mobility history has a different sensitivity level. Based on this, a trained Classification and Regression Tree (CART) model assigns a privacy level (low, medium or high) to the check-in.

3. The users choose their location sharing preference (coarse or fine-grained location).

4. Based obfuscation level and the user's preference, the obfuscation engine performs a hybrid obfuscation technique that includes the obfuscation operators: radius enlargement, radius reduction, and center shifting. When the predicted privacy level is "high", a simple cloaking is applied along with hybrid obfuscation.

Unlike the previous solution, **SmartMASK** is centralized and thus, more capable of han-

dling intense computations due to more available resources. Only privacy-level prediction and application of the obfuscation level is done for each new location which is much less intensive. But, as the model performs hybrid obfuscation on the location, the utility of location data may be decreased considerably.

To reduce the utility loss from obfuscation, authors in [78] propose an ML-based model to learn the user motivation behind a location check-in. The proposed method, firstly, takes the location check-ins and already available user motivation to train a model that predicts future motivation. Then, users then provide information on the effect of different obfuscation level on their check-in utility. Based on these responses and the predicted motivation labels, a cost-sensitive decision tree model (J48) is trained to predict the user's perceived privacy level. This solution is the first of its kind as it considers user-specific utility while designing the model that does not use differential privacy. Firstly, it addresses the effect of obfuscation on utility and specifically trains models to predict a privacy level that retains the highest amount of utility. Secondly, designing such intelligent models relieves users from making sensitive and critical privacy decisions. The major drawback here would be the unavailability of these types datasets for future researchers.

3.2.1.3 Differential Privacy based schemes:

Differential privacy (DP) has become the gold standard in privacy. Unlike most other privacy-preserving solutions, differential privacy based solutions' main aim to retain data utility, they also assume that the attacker has complete knowledge of the users and the network, and finally these schemes can also quantify the level of privacy they provide [79].

Differential privacy based solutions can be successfully applied in places where aggregate information is published [80], and DP would require that the changes made one location of a user should have a negligible impact on the final output making it impossible to send useful information to the LBS.

To address this issue, Dewri proposed a method that uses both k-anonymity and differential privacy to preserve location information [81]. In this method, the author fixes an anonymity set consisting of k locations where the probability of reporting the same obfuscated location x from these k locations is the same. To achieve this, the author adds Laplacian noise [82] to each Cartesian coordinate of the location. Though the choice of Laplacian noise is better to retain utility and has been extensively proved in the work, one of the major issues with this scheme is the selection of anonymity set that greatly affects privacy.

Another differential privacy based solution for preserving location privacy is **LPT-DPk** [83]. In this work, Yin et.al. focus on persevering frequent location patterns. The authors first create a frequent pattern tree called the Location Information tree based on the frequency of location check-ins. Once the tree is generated, the top-k frequent patterns are selected by using weighted selection based on Exponential mechanism [84] and a Laplacian noise [82] is added to the top-k frequent location patterns set to preserve privacy. The method is then evaluated against another top-k mechanism for the utility of the location data. The LPT-DPk Scheme retains more utility of the data and has a relatively low and stable error compared to the other previously proposed DP based schemes, but, it does not

discuss how the initial frequent patterns are derived, as this greatly impacts the effectiveness of the mechanism.

In [85], the authors propose a novel DP method that implements Reinforcement learning (RL) to preserve a node's semantic trajectory. The scheme is designed using game model as well, with the nodes and adversary as players. RL is implemented to selected the optimal privacy budget ϵ for the DP scheme. The obtained optimal budget is used to generate the "gamma noise" which will then be added to the location. This obfuscated location is then forwarded to the LBS and the results obtained as returned to the user. The implementation of RL in Location privacy is a relatively new concept and this paper effectively makes use of both RL and game model. Also, the optimal budget (strategy) is selected in a dynamic environment, unlike most other DP schemes, which on static location instances.

3.2.2 User Attribute Privacy

User privacy aims at preserving a user's profile information while communicating with other users or servers in an MSN [86, 87]. User privacy-preserving solutions can be broadly categorised into :

1. *Clustering and K-anonymity*

2. *Private Matching*

3. *Dynamic Pseudonymity*

3.2.2.1 Clustering and K-anonymity

In these schemes similar users are grouped together and privacy is provided at a group level to reduce information loss that might occur when techniques like Naive Anonymization are applied. **SANGEERA** [88] is one such clustering algorithm with the following procedure:

1. Nodes are partitioned into different clusters based on their quasi-identifiers and neighborhood information.

2. The quasi-identifier attributes are anonymized for each cluster to achieve k-anonymity.

3. All users in a cluster are collapsed to one node to prevent the revelation of intra-cluster nodes and edges.

4. Multiple relationships (edges) between two clusters are collapsed into one edge, to provide anonymity.

Previous anonymization solutions resulted in a significant information loss, whereas in SANGEERA, as edge generalization is adopted over perturbation, the structural information loss is reduced to a great extent. But, the major drawback of this scheme occurs when the clusters are very small or dense. In this situation even if the quasi-identifiers are anonymized the adversary can easily identify users based on other sources of user information. The algorithm does not consider the mobility and the dynamic nature of MSNs, where neighborhoods and communities change constantly making it inapplicable for more real-world MSNs.

To address drawbacks in solutions like SANGEERA, a new algorithm called **Equicardinal Clustering** was proposed in [89]. In this work, the user information is preserved at the network level. The algorithm can be summarized as follows:

1. Similar users are clustered using k-means.
2. To achieve k-anonymity in each cluster, the users are reclustered as follows :
   (a) Distance between the users and each cluster centroid is measured.
   (b) Based on the distances, users are assigned new clusters.
   (c) This is repeated until there are no more than n/k users are present in each cluster.
3. Users in a single cluster are represented by a cluster head.
4. All the links between two clusters are replaced by a single weighted edge. This weight represents the number of links between the two clusters.

This solution firstly reduces the information loss greatly compared to other schemes that use traditional clustering algorithms. As the neighborhood of the user is not considered for clustering, the privacy provided is not subject to user location, making the solution applicable to both OSNs and MSNs.

As most clustering algorithms used in these solutions are NP-hard, the solutions only provide sub-optimal results, therefore, cluster independent schemes like Private Matching need to be considered that utilize social features of an MSN to design a solution.

3.2.2.2 Private Matching

Friend Matching is a core feature of MSNs. MSNs allow users to connect with people in their neighborhood or network based on shared interests. To match users, they have to share sensitive information with the network which poses a serious threat to user privacy.

To solve the above issue a privacy-preserving profile matching scheme for MSNs called **FindU** was proposed in [90]. In **FindU** three privacy levels are implemented for private matching between P_1 and P_i, where $2 \leq$ i $\leq$ N:

1. PL-1 : P_1 and P_i will know the common attribute set.

2. PL-2 : P_1 and P_i will only know the size $m_{1,i}$ of the the common attribute set.

3. PL-3 : P_1 and P_i will only know the rank of each value $m_{1,i}$.

These privacy levels can be personalized by users, and the adversary can only obtain the output and the private inputs, thus, decreasing the amount of information he can obtain with each increasing level of privacy. In order to attain privacy levels, they designed two schemes :

1. Basic Scheme: This is defined to realize PL-1. In this scheme, a technique called Private Set Intersection (PSI) is used, where the attributes are encoded using hashing before the users share them.

2. Advanced Scheme: This is defined to realize PL-2 and PL-3. In this scheme, they use both PSI with BP (Blind-and-permute), where the user's attributes are encoded and

the shared attribute sets are permuted so the link between the ranks and the attributes is broken. In this scheme, each sharing is then encoded using homomorphic encryption.

Once the protocol ends, P_1, ranks $m_{1,i}$ locally to identify its best match, and then sends a connection request. This scheme takes into consideration every possible step at which the adversary may try to obtain information, and then designs schemes to make sure that the matching is resistant to active attacks. Also, the information of the user is being encrypted instead of being generalized, thus retaining the complete utility of the data. Though FindU provides several benefits compared to state-of-the-art privacy schemes in MSN like FNP and FC-10, it has higher communication costs as the encryptions are done on each communication between P_1 and P_i at every step in the matching process.

In order to overcome the issues in solutions like FindU, a privacy-preserving profile matching mechanism called **POSTER**, was proposed in [91]. In POSTER, the secure matching is done by using perturbation.

1. All user attributes are converted to binary values to create profile vectors.

2. Mixed vectors for secure sharing are generated by adding noise to the profile vectors and performing a secure dot product of profile vectors of A and B who have to be matched.

After these initial steps are carried out, the authors created the following two schemes :

Basic Scheme: The secure dot product computations occur on the receiver end and in the presence of other users called Helpers. If the helper has both noise and the mixed vector, he

might be able to obtain the user's private information, thus this is collusion.

Collusion Resistant Scheme: To avoid collusions, the scheme divides noise or perturbation into chunks and send it to multiple helpers. They compute the dot product and sends it to B. B, then computes the final dot product to check its similarity with A.

This paper firstly does both secure friend matching as well as authentication by using a Verification scheme, to make sure that valid users are exchanging information. Also, as the mechanism does not use the computation heavy operations like Homomorphic encryption, it reduces the computation complexity and communication cost. The major drawback is that it does not focus on "helpers" selection which makes it vulnerable to Sybil attacks because when an adversary can act as multiple users, a few such profiles can be selected as helpers for the same communication allowing the adversary to obtain information even in the Collusion Resistant scheme of POSTER.

To avoid depending on other users, as in POSTER for private matching, Li et.al in [20] proposed a scheme called **Match-MORE**. This scheme was designed for users to securely match with friend-of-friends. The complete mechanism lies in the **Matching degree function** which uses Katz Score and Dice similarity coefficient to calculate the social strength of two users, and the similarity score between two users, respectively. The matching happens in two phases :

1. **Friend Discovery Phase**: In this phase, A discovers new friends by broadcasting a connection request. Each responder sends their similarity score as a reply. A then selects the responder with the best score as the 'friend'.

2. **Friend Recommendation Phase**: In this phase, the new friend goes through his friend list and calculates the similarity between A and all his friends, and then recommends a friend with the best similarity score to A.

In both phases, the similarity calculation is done using the Matching degree function and all the communications implement bloom filters. Unlike the schemes mentioned earlier, this paper quantifies the privacy provided by the scheme by using Shannon entropy and theoretically proves the accuracy, effectiveness, and efficiency of the scheme. Also, Match-More is lightweight as it completely avoids the use of Homomorphic encryption. Finally, unlike other schemes that share actual attributes (true or perturbed) for matching, this scheme avoids that and just does matching using scores, therefore, reducing the risk of information leakage to the minimum.

3.2.2.3 Dynamic Pseudonymity:

Dynamic Pseudonymity Mechanism (DPP) proposed in [37] aims at providing both user and location information privacy. To protect the user's identity, the scheme uses multiple anonymizers. The DPP scheme divides the user's LBS queries into chunks, where each chunk is forwarded to a different anonymizer. While forwarding the query chunks belonging to the same user are assigned different pseudo-identities while interacting with different anonymizers. This makes sure that the adversary can not link user information to a query result or obtain a true User ID from any one of the anonymizers. The anonymizers also have a k-anonymity function to ensure user's location privacy in the query. The paper discusses two threat models, "weak adversary" and "strong adversary" based on the locations of attack

which are: 1) the wireless channel between the user and the LBS and 2) the anonymizer itself. The scheme was then designed to address these threats. The major advantages of this scheme are the fact the authors considered different levels of threats making the solution well-rounded. The use of hash trees and a single k-anonymity operation greatly reduces the computation time compared to other dynamic pseudonymity solutions. Also, the privacy that this scheme provides is two-fold privacy as it ensures both user and location privacy.

Another Dynamic Pseudonymity based scheme was proposed in [92] which focuses on understanding the context behind an LBS request to ensure user privacy. To impede the adversary from linking user ID to his true location, an identity management system is used. To make the system more secure, a hashing over pseudoID is performed, where the hash key is combined. The method proposed in paper [92] enhances user privacy at two levels: 1) The identity management system replaces user identities with pseudo identities and this pseudoID is retained until the service request is fulfilled, thus for each service request, a new pseudoID is assigned to the user. 2) Hashing over user's pseudo IDs by using userID + Service time as the hash key, to securely share the pseudo IDs. One of the main benefits of this scheme is that it has multiple levels at which privacy is ensured. Also, as the pseudo ID is hashed, this not provides user data privacy but also ensures communication privacy to a certain extent, as the pseudo ID is a part of the query packet being forwarded to the LBS. But, as the scheme generates new IDs and performs hashing for every query, and the computation is performed on the user-end, it might deplete the limited resources available on the mobile device [93].

3.2.3 Communication Privacy

Depending on how a communication privacy-preserving mechanism is carried out, the solutions can be classified into :

1. *Privacy enhancing schemes*

2. *Privacy-Aware Routing mechanisms*

3.2.3.1 Privacy enhancing schemes

These schemes are designed to work alongside already implemented security protocols in MSNs and they heavily depend on Digital Signatures, Key Agreements, and certificates.

PRIF, proposed in [94] is an improved version of the forwarding scheme that is prevalent in MSNs today. The forwarding scheme proposed revolves around the concept of common interest-based communities formed in MSNs and, it preserves communication privacy by ensuring the privacy of the interacting parties. This is done by hiding the user's interests and other information before he joins a community or interacts with anyone from the community. This privacy-preserving authentication protocol uses Schroff signatures and Group certificates handled by a central trusted authority TA. Though the scheme uses strong authentication protocols to ensure the privacy of the communicating entities, the use of central authority for the token generation is not ideal as it becomes a single point of failure for this scheme. Instead, if the scheme provides a way to improve trust between the communicating entities, it eliminates the need for central authorities and also adds a distributed aspect to the scheme.

Privacy Preserving Authentication Scheme (PPAS) [36] is a scheme based on group signatures. Group signatures are primarily used to provide user anonymity and unlinkability. This scheme ensures user legitimacy by generating an unlinkability token by using the group's public key parameters. To ensure integrity and confidentiality during communications the scheme uses signing and verification algorithms and session key agreement for every communication. This mechanism reduces the overload that the above-mentioned scheme suffers from, by reducing the number of tokens generated and considering group signatures instead. Unlike other schemes, it also considers the mobility of the user as part of the scheme, instead of a snapshot of the MSN. But, similar to the previous scheme, this solution also depends on a single trusted authority to carry out the token generation.

As most privacy-enhancing schemes rely on trusted authorities to function which may be vulnerable to serious breaches, it is necessary to consider routing protocols that enhance privacy because they are more widespread in terms of the aspects they include and effect.

3.2.3.2 Privacy-Aware Routing mechanisms

Onion routing [95] is one of the first strategies proposed to preserve communication privacy. This strategy ensures data integrity in both connection and connection-less systems. It uses mixers or onions, which store, encrypt, and forward data to the next node in random order. Generally, multiple mixers are used to ensure that communication is protected against traffic analysis. Though this strategy ensures anonymous communication, the major drawback is the uni-directional feature of mixers or onions. This means that the mixers can only carry operations for one-way communication. To make it bi-directional a set of reply onions

needs to be deployed. The second drawback of onion routing is that it focuses on a single communication. An improvement over this method is proposed in [96]. This method extends onion routing to a multi-casting scenario and uses "Bloom filters" to enhance communication anonymity by obscuring the routing list of communication packets. It is one of the first works to use the concept of bloom filters in a privacy setup, and it also overcomes the disadvantages of the one-way privacy enhancement present in older solutions.

3PR [97], is a communication privacy-preserving scheme that uses machine learning techniques to learn user's mobility patterns to predict their future routes and uses those predicted routes to route message. This is done by calculating the maximum likelihood of a node encountering the destination, and then, these likelihood values are hidden from other nodes within and outside the community. As part of the scheme, privacy-preserving functions like "max probability" and "partial sum" that make use of random number generators are proposed. This work uses the idea of "route-recommendations" in a communication setup which is novel and first of its kind. One major drawback of this solution is that preserves only the information related to the packet's possible destination and fails to preserve the privacy of the packet's content (message), which may hold sensitive information.

3.3 Online resources and Databases

In this section, we will be listing and summarizing a few data sets and data generating tools available online that are used extensively in research on privacy in MSNs and LBSNs.

The most popular datasets are the **Facebook** [98] and **Twitter**[99] that available as part

Name	Domain	Dataset information
Facebook	user profiles and friend lists	Ego-networks:10 Nodes: 4,039 Edges: 88,234
Twitter	user friend lists and ego-networks	Ego-networks:973 Nodes: 81,306 Edges: 768,149
Deezer		
Romania		Nodes: 41,773 Edges: 125,826
Croatia	friend network with liked music and genres	Nodes: 54,573 Edges: 498,202
Hungary		Nodes: 47,538 Edges: 222,887
Brightkite dataset	check-ins and friendship network	Check-ins: 4,491,143 Nodes: 58,228 Edges: 214,078
Gowalla	user profiles, location profiles and location check-ins	Check-ins: 36,001,959 Users: 319,063 Locations: 2,844,076
Weeplace	check-ins, profiles and location information	Check-ins: 7,658,368 Users: 15,799 Locations: 971,309
Foursquare	location based friendship network	Nodes: 106,218 Edges: 3,473,834
GeoLife	user GPS trajectory	Users: 182 Trajectories: 17,621 Timespan: 3 years
LifeMap	user location monitoring, user trajectories	Users: 8 Nodes: 9681 Edges: 1717 Wi-Fi APIS: 52,510
T-drive	taxi GPS traceset	Users: 10,000 Timespan: 1 week
Cabspotting	taxi GPS raceset	Users: 500 Timespan: 30 days
Manhattan Taxi	taxi GPS traceset	Trajectories: 1000 Timespan: 1 year

Table 3.1 Summary of MSN and LBSN datasets

of SNAP (Stanford Network Analysis Project) by Stanford University [100]. The Facebook dataset [98] consists of users, their friend lists and ego networks. It was collected from the Facebook apps of survey participants and has 10 ego networks with a total of 4039 nodes and 88234 edges. Each user is represented by 25 features like location, education degree, job start date, and end date, employer information, workplace, and several others. These networks are represented by undirected graphs. The Twitter dataset [99] like the Facebook dataset, consists of friendship circles (lists) and ego networks, but the network here is represented as a directed graph with about 973 ego networks and a total of 81306 nodes and 768149 edges.

Another friendship dataset from SNAP [100] is **Deezer**[101, 102]. The dataset has friendship networks of users from 3 different European countries and contains three sub-networks

represented by directed graphs for Romania, Croatia, and Hungary. The number of nodes and edges in each of these sub-networks are mentioned in Table 3.1. The dataset also provides user's preferred genres preferred which have been compiled based on the songs users liked in a music network.

Brightkite [103] is an LBSN dataset that consists of user's location check-ins and their friendship relationships. The data-set includes user check-ins and friendship networks of users within the social network. It has more than 4 million check-ins with 58,228 nodes (users) and 214,078 edges. Each check-in includes user id, check-in time stamp, latitude and longitude readings, and location id. This dataset has been specifically used to study user friendships and mobility patterns in MSNs [104]. This particular dataset is sparser than other mobility datasets because we only have places at which users checked-in deliberately.

Gowalla [105] is a popular LBSN dataset collected using Gowalla API. For each user, the dataset has user profiles, their friendships, and location check-in history and each location an attached location profile. Based on these profiles, the locations are categorized into 7 subcategories like community, nightlife, entertainment and many more. Over the years, the data collected from the API comprises 36 million check-ins with 319,063 users and over 2,844,076 locations. This dataset has been mainly used in location recommendation systems [106], but can also be used to generate privacy models based on the location predictions.

The **Foursquare** dataset [107] is another LBSN dataset that is a part of the Social Computing Data Repository at Arizona State University and provides a friendship network among users of the Foursquare network. The social network was available to users with

GPS enabled mobile devices and the data was collected through software installed on their respective devices. This dataset consists of 106,218 nodes and 3,473,834 edges.

Weeplaces [108] is another dataset obtained from a website. The website that was used to collect this dataset is now integrated into other social network applications like Facebook Places, Foursquare and Gowalla. The website was used to visualize user check-ins and the dataset generated includes user's friends who use Weeplaces, location check-ins and additional information about the locations. This dataset contains 7,658,368 check-ins generated by 15,799 users over 971,309 locations. It is similar to the Brightkite and Gowalla datasets but provides additional information about the location like locationID, city, and location category.

GeoLife GPS Trajectories [109, 110] is a trajectory dataset collected by Microsoft Research Asia, Geolife project. It contains GPS trajectories of 182 users collected over three years (from April 2007 to August 2012). It has a total of 17,621 trajectories with location co-ordinates and the altitude of users' locations and the locations were updated every 1 to 5 seconds. Datasets like these can be used to understand user's mobility patterns [111] and design privacy models based on it. GeoLife has been used extensively to provide location privacy in Mobile Crowd Sensing (MCS) systems.

LifeMap Mobility data [112] is a dataset generated by a mobility monitoring system called LifeMap at Yonsei University in Seoul. The dataset contains fine-grained mobility data of 8 users collected over two months in Seoul, Korea. The dataset includes location coordinates, Wi-Fi fingerprints and user-defined places. The locations were collected by the

system every 2 to 5 minutes. The data were collected from users' mobile devices and has 9861 nodes on 1717 paths and 52510 Wi-Fi APIs. This dataset was used in research on mobility learning and movement predictions of an MSN user [113, 114, 115].

T-drive [116], **Cabspotting** [117], and **Manhattan Taxi** trajectory [118][119] are a few taxi trajectory datasets. Manhattan taxi trajectory has 1000 taxi trajectories collected over one year in the city of Manhattan [120]. Cabspotting is a traceset of mobility data of taxi cabs in San Francisco. It contains GPS coordinates of 500 taxi cabs collected over 30 days. T-Drive is a dataset collected as part of Microsoft Research, featuring taxi drivers in Beijing and has over 10,000 users with data collected over one week.

Apart from the datasets mentioned so far, there are also ways to generate synthetic data using online data generators. These tools provide users with several options and fields to generate custom datasets. Mackaroo [121] is one such website that lets us generate mock data with user profiles and a variety of other fields like location, occupation and so on. Generatedata [122] is another online data generator similar to Mackaroo. It lets us generate random users with a variety of fields along with their location data. The data can be generated in a plethora of formats like Excel, HTML, JSON, SQL, and XML. DTM Generator [123] is another popular data generator that produces data rows and schema objects and also other optional schema objects like views, triggers and many more. It is highly compatible with most popular database systems like MySQL, Oracle and Microsoft SQL server. Several other data generators have been mentioned and summarized in [123].

CHAPTER 4

User Privacy in LBSNs based on User Motivation

4.1 Motivation

In recent years, mobile technology has seen a great deal of development on both hardware and communication fronts, and better internet availability has made mobile devices omnipresent. This evolution encouraged several web-based applications to migrate to their mobile versions to provide better reach and experience to their users [124]. Moving to a mobile platform has opened up several opportunities for these applications to provide different location-based services to their user base. For example, Facebook alone has several services that were either improved or introduced after moving to an almost complete mobile operation of their application. One such service is check-ins, which was an already existing feature on Facebook. Now, users can post places they visit on the go and attach pictures or maps pointing to their exact location with their check-in. Another such service is **Facebook Marketplace** [125], which is relatively new to the platform and allows people to use their location for advertising items for sale or discover sales nearby, find apartments, and many more. **Facebook Places** [126], is another new service, which is similar to Foursquare [107], allowing people to use their location to explore their neighborhood. This shows that many social networks today use user's location information in most of their services, thus qualifying them as **Location-based Social Networks (LBSNs)**.

Given the popularity of these networks, it is expected that we have more users signing up for these platforms and taking advantage of their services and features. The most commonly

used feature on these platforms is ‘posts’, also popularly known as ‘check-ins’. In these posts, the users share locations that they are visiting with their friends. This is done to get some recommendations about the place or make themselves perceive as interesting, thus helping them make better connections with their social circle [127].

While checking in on LBSNs, the users release a lot of information like the geographical coordinates of the location, the location type (restaurant, stadium, movie theatre), time, if they are already at the place, or are heading towards the location and several other things as shown in Figure 4.1. Therefore, a simple check-in might release a lot more information than the user has intended to. The released information, combined with other sources, can be used to devise and launch strong inference attacks [128, 129, 130, 131**?**]. Also, if the user check-ins are frequent, the attacker can collect all such check-ins and launch re-identification attacks (to infer places like Home and Workplace of the user) [132], profile users’ daily activities or identify commonly taken routes [133]. Therefore, there is a high privacy risk associated with location check-ins, irrespective of the frequency of check-ins for a particular user.

For example, let us assume that Alice goes to Georgia State University, and on most days, she shares a Facebook post in the morning saying she has reached the university. On another day, she posts about a basketball game she is attending at the university. This particular post also has a few friends tagged in it, and many others were seen making similar posts around the same time. In the former case, the check-ins happen more regularly or consistently in the morning, so one might infer that she is heading to “Work”. However, in

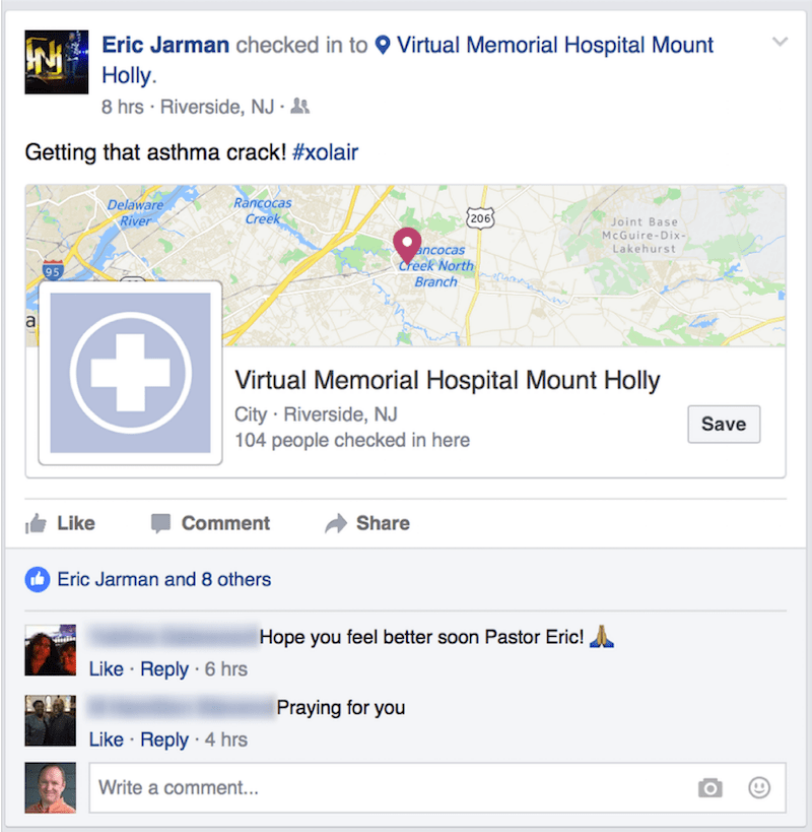


Figure 4.1 Example check-in on an LBSN

the latter case, though the check-in location is the same as the former check-in's location, it has a social aspect to it, with tagged friends and many others making similar check-ins simultaneously. The first type of check-in has a more personal intention, like keeping track of her university visits. As this check-in lacks social nature and is more regular, releasing this information over a prolonged period will lead to the attacker inferring her activity or the type of location as 'Work'; therefore, it makes more sense to provide stronger privacy for all such check-ins. The second kind of check-in has a social intention behind it, given the number of people making similar check-ins simultaneously, the check-in time, and the tagged friends. In this case, if we provide strong privacy, we will have a higher information loss and

may not meet Alice's social needs any longer. Therefore, we can relax the privacy policy (apply a lesser amount of privacy) to retain the 'motivation' behind the check-in [134].

Several privacy policies are available on social networking applications or proposed as part of research to prevent inference attacks and sensitive information disclosure while releasing the said social network data to third-party applications. The major drawback with the former is that the settings are either so deeply nested in the application that the user might not be able to navigate to them [135] or are complicated for a naive user to understand. Another thing to note is the *motivation* behind the check-ins. Users check-in with some form of social intention, in which case, applying any privacy policy might lead to information (intention) loss and therefore reducing the service quality when the obfuscated information is released. Stemming from the lack of a clear understanding of privacy and privacy settings and the need to satisfy their social needs, the users either opt for complete disclosure (complete location information) or complete non-disclosure (no location information). These extreme settings either have very high privacy risks or very low utility. In privacy-preserving solutions like[136, 137, 138, 139], all the locations are treated similarly, and the same level of privacy is provided. In these cases, with sufficient knowledge of the user's connections and the network itself, the attacker can back-engineer the policy to obtain the different parameters of the model [140]. Therefore, there is a need for privacy models that take the user's intention (motivation) behind a check-in into consideration to meet both their social and privacy needs and introduce some inconsistency, making it difficult for the attacker to back-engineer the policy.

In this chapter, we provide a user motivation-based privacy policy to bridge the gap between user psychology and privacy policies. The model preserves location check-ins based on the motivation (intention) behind a particular check-in. To our knowledge, this is the first work that has designed a privacy policy centered on user motivation. We consider each check-in individually; therefore, the same user's check-ins might have different policies applied to preserve the information, thus introducing a lot more variation than most other works.

4.2 Problem Statement

4.2.1 Components

Definition 1. Location based social network (LBSN): LBSN can be defined as an undirected graph $G = \{V, E, C\}$, consisting of a set of LBSN users $V = \{v_1, v_2, \ldots, v_n\}$. The friendship relations among the users are represented by the edge set **E**, of the social network, where $e(v_i, v_j) \in E$, indicates that a friendship relation exists between users v_i and v_j, and $v_i, v_j \in V$. **C** represents the set of check-ins made by the users on the LBSN.

Definition 2. check-ins: A set of check-ins made on a LBSN can be represented as $C = \{c_1, c_2, \ldots, c_j\}$. Each check-in c_i consists of user identifier, location information and the time stamp and can be denoted as $c_i =< v_i, l_i, t_i >$, and all the locations belong to a location universe $L = \{l_1, l_2, \ldots l_p\}$. Each location l_i can be represented as $l_i = \{lid_i, latitude_i, longitude_i, type_i\}$

4.2.2 Characterizing User Motivation

It is crucial to obtain more information about a check-in and understand it better before designing the privacy policies. We can extract meta attributes, otherwise known as context features, by using both the check-in features and the social network. In this work, the following context features were considered: weekday, time of day, user check-in frequency, location check-in frequency, and co-location.

Weekday: The "weekday" context-feature tells us if a particular check-in has been made during the weekdays (Monday to Friday) or on weekends (Saturday and Sunday). We use the 'timestamp' of the check-in to obtain this information. This is a binary feature, which can be represented as follows:

$$weekday = \begin{cases} 1, & \text{if } day \in \{Saturday, Sunday\} \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

Time of day: This particular feature tells us if a check-in was made in the morning, afternoon or evening. The check-in's timestamp is used to compute this feature. It can be denoted as follows:

$$time\_of\_day = \begin{cases} 0, & \text{if } Morning \\ 1, & \text{if } Afternoon \\ 2, & \text{otherwise} \end{cases} \tag{4.2}$$

Location frequency: This particular context feature provides an insight into the activity of a particular location. We compute it by calculating the frequency of a location in all

the check-ins made on the system. It can be denoted using *Iversion bracket notation* [141] as follows :

$$location\ frequency(l) = \sum_{i=1}^{|C|} [l_i = l] \tag{4.3}$$

User frequency: This particular context feature provides us an insight into a particular user's activity. We compute this by calculating the frequency of occurrences of a user v, in all the check-ins made on the system. It can be denoted using *Iversion bracket notation* as :

$$user\ frequency(v) = \sum_{i=1}^{|C|} [v_i = v] \tag{4.4}$$

where C_i is the set of all the check-ins made by user v_i.

Figure 4.2 Structure of UMPP model

Co-location: Consider two check-ins $c_i =< v_i, l_i, t_i >$ and $c_j =< v_j, l_j, t_j >$. If $l_i = l_j$ (same location), $|t_i - t_j| <= \tau$ (check-ins occurring within a threshold), where τ is the temporal threshold and $e(v_i, v_j)$ exists (v_i and v_j are friends), it is called a co-location. For each check-in, we consider the total number of such co-locations. Algorithm 1, provides the steps for calculating the co-location.

Algorithm 1 Computing co-location of the check-ins

Input: $G = (V, E, C)$: the location based social network

C: the set of check-ins

τ: user-defined time difference

Output: the co-locations for all check-ins

1: **for each** check-in $c_i \in C$ **do**
2: Initial $c_i[co - location] = 0$
3: **for each** check-in $c_j \in C$ $(i \neq j)$ **do**
4: m = $c_i[v_i]$, n = $c_j[v_j]$
5: **if** $|c_i[timestamp] - c_j[timestamp]| \leq \tau$ **and** $c_i[l_i] == c_j[l_j]$ **and** $e_{mn} \in E$ **then**
6: $c_i[co - location] = c_i[co - location]+1$
7: **end if**
8: **end for**
9: Save $c_i[co - location]$ as the co-location for check-in c_i
10: **end for**
11: **return** the co-locations for all check-ins

4.2.3 User Motivation

As explained in Section 4.1, every check-in made on an LBSN has an intention or user motivation associated with it. To identify the motivation behind a check-in, we first compute all the context features mentioned in Section. 4.2.2, and then cluster the check-ins based on this data. In [142], it is explained how certain features of the check-in can indicate the

intention behind a check-in; therefore, we apply this idea in our labeling. For each cluster, we analyze trends of the computed context features and proceed with labeling the cluster with either of the following two labels:

- Social motivation: A check-in is said to have social motivation if the check-in is made with an intention to communicate a person's whereabouts with others in the network. We apply this label to the check-ins where the location is very active (has high location frequency), the user is active (high user frequency), the check-in has high co-location values (user's friends have also checked in into the same location in the same time), when the check-ins happen in the evenings and/or weekends and finally, when the type of the location is a public place like restaurant, cinema, museum, etc.

- Private motivation: A check-in is said to have a private motivation if the user makes the check-ins for a personal reason. As mentioned in [143], sometimes users check-in only to keep track of their activities. In such cases, the check-ins are more frequent and happen in similar locations over time. We apply this label to check-ins that have less active (less location frequency), the user is active (high user frequency), the check-ins has a less or no co-location and happen on mornings and evenings and/or weekdays and finally when the type of the location indicates home or office.

4.2.4 Problem Definition

Given a LBSN network G, as defined in Definition. 1, location check-ins C, as defined in Definition. 2, and user motivation um of the check-in. This paper aims to preserve the privacy of the check-ins in C, based on user motivation um, with the following objectives:

- Minimize re-identification of user motivation behind check-ins C.
- Minimize the information loss for *social check-ins*, while maximizing privacy.

4.3 User Motivation Based privacy preservation

The basic structure of the UMPP model is shown in Figure 4.2. Each location check-in has three different types of obfuscation applied to it.

- Timestamp obfuscation
- Semantic Location context obfuscation
- Semantic Location obfuscation

Each of these obfuscation techniques effect multiple context features, thus providing a much better chance against re-identification. Following is the detailed explanation of these obfuscation techniques.

4.3.1 Timestamp obfuscation

Timestamp obfuscation is the most commonly used technique used in privacy policies to prevent reidentification and user profiling and tracking. In our model, we adopt a *Reverse kNN* approach to obfuscate the timestamp. As shown in Lines 2 - 5 of Algorithm 2, we take an individual check-in c_i and generate a nearest neighbor check-in set C_j, containing k nearest check-ins with respect to the timestamp t_i of check-in c_i. We then randomly select one of the nearest neighbor check-ins (c_{j_i}) and use the timestamp of that check-in as the new timestamp t'_i of c_i.

Algorithm 2 User Motivation based Privacy Preservation

Input: $C = \{c_1, c_2, \ldots, c_i\}$: original check-ins set
$c_i =< v_i, l_i, t_i >$: the original check-in
v_i: user identifier for check-in c_i
$l_i = \{lid_i, latitude_i, longitude_i, type_i\}$: the location information of c_i with id, latitude, longitude, and location type
t_i: the timestamp of c_i
um_i: the user motivation label of c_i
Output: $C' = \{c'_1, c'_2, \ldots, c'_i\}$: obfuscated check-ins set
$c'_i =< v_i, l'_i, t'_i >$: obfuscated check-in
$l'_i = \{lid_i, latitude'_i, longitude'_i, type'_i\}$: obfuscated location information with processed latitude, longitude, and location type
t'_i: obfuscated timestamp

```
for each check-in c_i ∈ C do
    TIMESTAMP OBFUSCATION:
    Using k-nearest-neighbor algorithm with t_i to obtain a check-ins set C_j = {c_j1, c_j2, ..., c_jk} of closest neighbors
    Randomly select check-in c_ji from C_j
    t'_i = t_ji (the timestamp of c_ji)

    SEMANTIC LOCATION CONTEXT OBFUSCATION:
    Generate the lexical location context tree CT_i of l_i over type_i
    if um_i == social then
        type'_i= 1st level ancestor of type_i in CT_i
    else if um_i == private then
        type'_i= 2nd level ancestor of type_i in CT_i
    end if

    SEMANTIC LOCATION OBFUSCATION:
    Generate semantic location tree LT_i of l_i over latitude_i, longitude_i
    if um_i == social then
        latitude'_i= latitude of 1st level semantic location in LT_i
        longitude'_i= longitude of 1st level semantic location in LT_i
    else if um_i == private then
        latitude'_i= latitude of 2nd level semantic location in LT_i
        longitude'_i= longitude of 2nd level semantic location in LT_i
    end if
end for
return obfuscated check-ins C' = {c'_1, c'_2, ..., c'_i}
```

4.3.2 Semantic Location Context Obfuscation

A location context is the 'type' or 'nature' of the location like a restaurant, a religious place, cafe, airport, and many more. In most LBSNs, this location type or context information is either directly posted or hidden as metadata when a user checks in; therefore, it can be easily extracted. Given that the context of a location is easily available, even if a privacy policy is applied to the geographical data, one can mine the exact location by checking out how many places within the area share the same context and/or obtain the data from other tagged users. Therefore, it is crucial to consider preserving this information as well in check-ins. In the UMPP model, we preserve this data by applying semantic obfuscation at the lexical level.

For example, let us consider a location which is a "Sushi bar". This is a very specific context for a location. Now if we want to preserve this information, we can move one level up on the lexical tree as shown in Figure. 4.3, and generalize the location to a "Japanese Restaurant". Furthermore, if we want to preserve more location context information, we generalize it further to "Food/Restaurant". As shown in Lines 7-13 of Algorithm 2, in our model, we move up to different levels in the lexical context tree of location type $type_i$, based on the user motivation um_i of check-in c_i.

4.3.3 Semantic Location Obfuscation

Semantic Location Obfuscation is a way of generalizing the level of a location's address. As we move up the semantic obfuscation scale, more and more parts of the address are omitted,

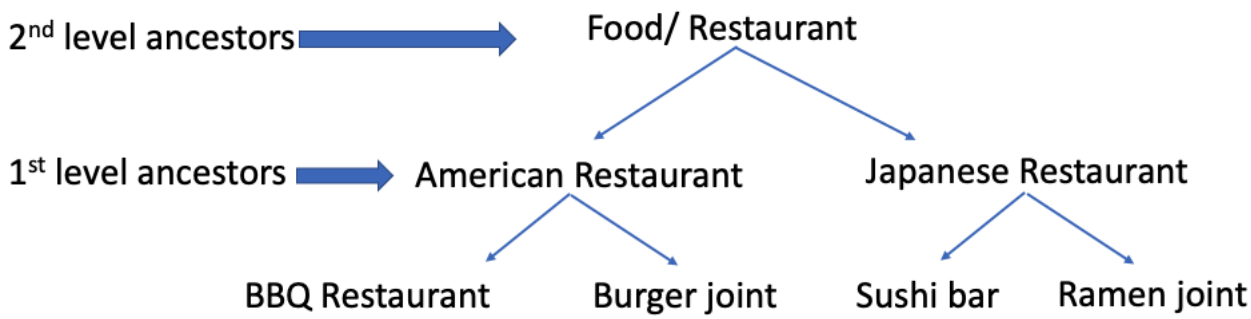


Figure 4.3 Lexical tree of location context or types

thus greatly reducing the granularity of the location, and this technique has been found to provide much better preserving preservation compared to simple geographical obfuscation [144]. Figure. 4.4 shows the levels of semantic information of a sample address format. Therefore, if one needs to preserve more privacy, a higher-level semantic obfuscation can be applied to the location address. In our model, we move up to different levels in the semantics of the address based on the user motivation um_i of check-in c_i, as shown in Lines 15-23 of Algorithm 2, and then we obtain the co-ordinates of that semantic level component (either city or state) as the new co-ordinates of c_i.

Location Address = < Street Number and Name, City, State, Country>

Decreasing granularity

3rd level	Co-ordinates of Country
2nd level	Co-ordinates of State
1st level	Co-ordinates of City
Exact location	Co-ordinates of Street Number and Name

Increasing Privacy level

Figure 4.4 Semantic levels of location information

4.4 Performance Validation

4.4.1 Datasets

The proposed model is evaluated two real-world datasets: Gowalla [105] and Brightkite [104]. Both of these are Location-based social networks that allow users to share their locations in the check-ins. In both datasets, we use the check-ins made in the United States over a time frame of 30 days. Table 4.1 shows the details of the selected check-ins in both the datatsets.

	Gowalla	Brightkite
Check-ins	35,000	30,000
Users	1900	1794
Locations	498	347

Table 4.1 Dataset Statistics

4.4.2 Evaluation Metrics

To evaluate the performance of our proposed model, we measure the performance based on two metrics: Re-identification accuracy and Information loss.

4.4.2.1 Re-identification accuracy (RAC)

It is essential to note that a check-in's motivation should not be accurately identified after privacy preservation. To measure the **re-identification accuracy**, we first train a classification model that can accurately predict the user motivation. Then, we measure how accurately the classification model can predict the user motivation of the obfuscated check-ins. The lower the re-identification accuracy, the higher is the privacy provided. The

re-identification accuracy can be calculated as :

$$RAC = \frac{correctly\ predicted\ motivation\ labels}{Total\ number\ of\ check-ins} \tag{4.5}$$

4.4.2.2 Information Loss (IL)

To calculate the information loss, we use the Average of Sum of Squared Errors (Avg.SSE) metric. The error is calculated between the original check-in c_i and the obfuscated check-in c'_i. Avg.SSE can be calculated as follows:

$$Avg.SSE = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \left(c_{ij} - c'_{ij}\right)^2}{n} \tag{4.6}$$

where,

n is the number of check-ins, and

m is the number of features in the check-in

4.4.3 Results

We evaluate our model's performance against another location check-in privacy preserving model **PrivCheck** [145]. We implement the historical check-in privacy model of PrivCheck for comparison. This model takes user-specified private data and then clusters the check-ins based on the activity/ location type and then applies privacy. To apply the model to the scenario presented in our work, we take all the check-in features as user-specified private features. The results provided are averages over 10 runs of the experiments.

4.4.3.1 Affect on Social Motivation check-ins

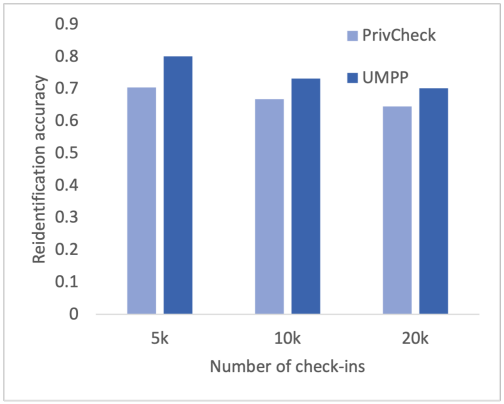
The model aims to preserve the social motivation check-ins to some extent while minimizing information loss. Figure. 4.5 shows that the social motivation check-ins preserved using the UMPP model can be reidentified more than the check-ins preserved using PrivCheck. Furthermore, the information loss of the UMPP model, as shown in Figure. 4.6, is much lesser than the PrivCheck model. Though the UMPP model provides 9% less privacy over both datasets than PrivCheck, the information loss is almost 20% lower than the PrivCheck model. Therefore, our goal to reduce information loss in "social motivation" check-ins is met by a much higher margin when compared to the baseline, at a very small privacy price.

	Gowalla		Brightkite	
number of check-ins	RAC	IL (10^2)	RAC	IL (10^2)
5k	0.800	3.248	0.817	2.940
10k	0.732	3.903	0.769	3.438
20k	0.701	4.532	0.756	4.041

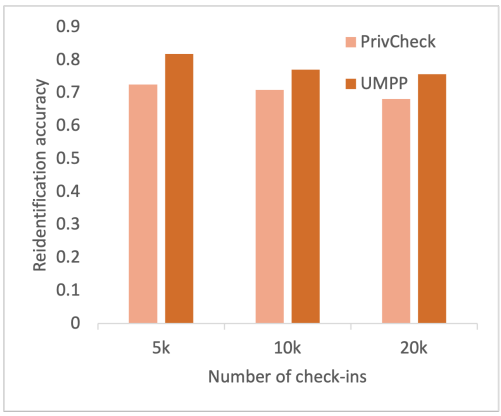
Table 4.2 UMPP on Social motivation check-ins

	Gowalla		Brightkite	
number of check-ins	RAC	IL (10^2)	RAC	IL (10^2)
5k	0.703	3.893	0.725	3.386
10k	0.667	4.621	0.708	4.076
20k	0.646	5.384	0.681	4.790

Table 4.3 PrivCheck on Social motivation check-ins

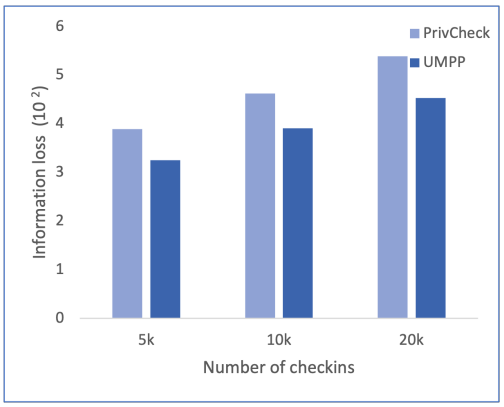


(a) Gowalla

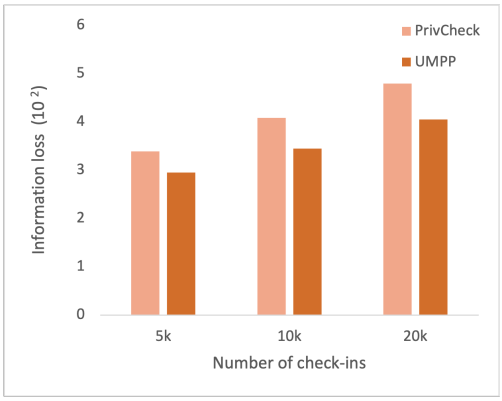


(b) Brightkite

Figure 4.5 RAC on Social Motivation check-ins



(a) Gowalla



(b) Brightkite

Figure 4.6 IL on Social Motivation check-ins

4.4.3.2 Affect on Private Motivation check-ins

Our model aims to provide more privacy for private motivation check-ins, which is clearly shown in Figure. 4.7. The UMPP model reduces the reidentification of the private motivation labels by 6% compared to the PrivCheck model. If we observe the information loss on the private motivation check-ins in both models on the datasets in Figure. 4.8, we can see that the information loss is almost the same. Therefore, the UMPP model provides an average

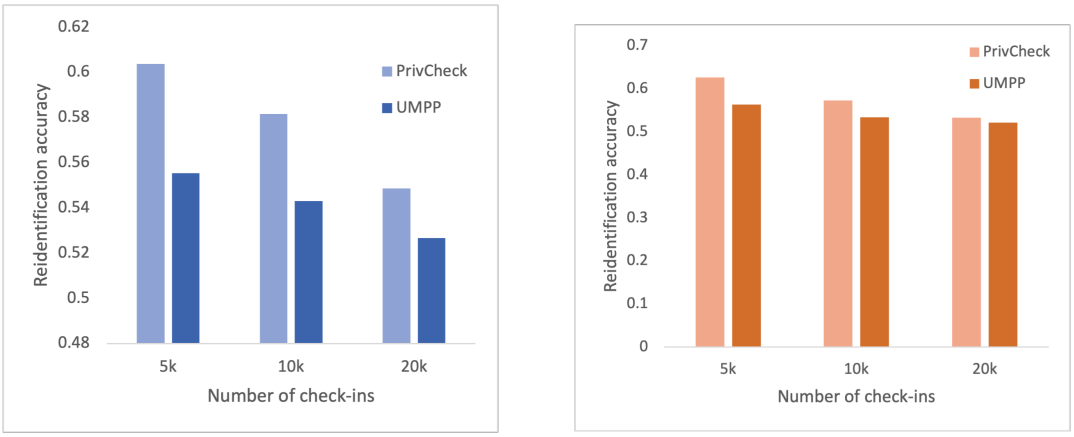
of 6% more privacy than the PrivCheck model for the same amount of information loss on both datasets.

	Gowalla		Brightkite	
number of check-ins	RAC	IL (10^2)	RAC	IL (10^2)
5k	0.555	6.953	0.563	6.516
10k	0.542	7.396	0.533	7.176
20k	0.526	7.827	0.521	7.335

Table 4.4 UMPP on Private motivation check-ins

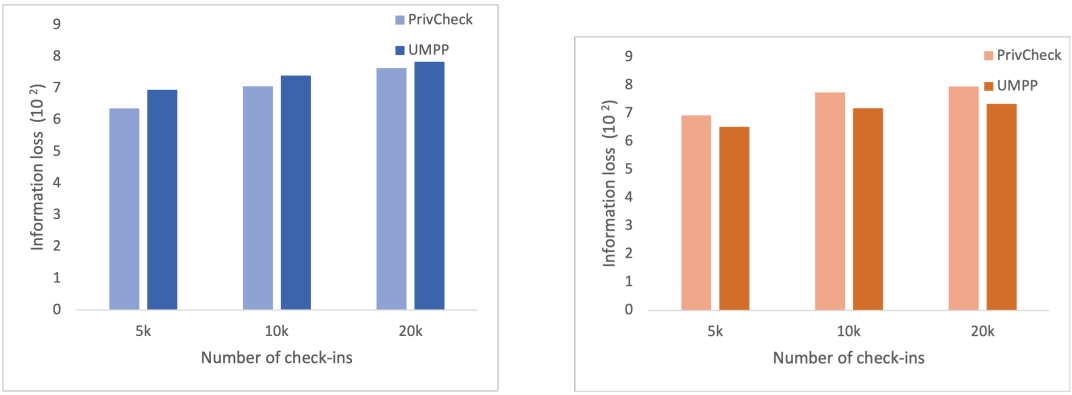
	Gowalla		Brightkite	
number of check-ins	RAC	IL (10^2)	RAC	IL (10^2)
5k	0.603	6.365	0.626	6.921
10k	0.581	7.065	0.573	7.743
20k	0.548	7.633	0.532	7.951

Table 4.5 PrivCheck on Private motivation check-ins



(a) Gowalla (b) Brightkite

Figure 4.7 RAC on Private Motivation check-ins



(a) Gowalla (b) Brightkite

Figure 4.8 IL on Private Motivation check-ins

4.5 Summary

Mobile devices and their usage have become the norm in today's world. To cater to the users of these mobile devices, many applications in use today have some or all services that LBSNs provide. Therefore, each user is a part of several different LBSNs at any given time. Given the universal nature of LBSNs, the user's data is constantly being used to provide better services. Therefore, information leakage in LBSNs is a major threat to the user. In our paper, we focus on one such service on LBSNs called check-ins. In most current applications, when a user's check-in is preserved, it does not consider the intention or the motivation behind the check-in. This results in losing the meaning of the check-in and, by extension, the utility. Therefore, there is a need to develop privacy policies that take the user's motivation behind a particular check-in into consideration. This work proposes a model that divides the check-ins into two categories and applies different privacy policies based on the categories' requirements.

Experimental results show that the model effectively reduces the information loss by

about 20% for the 'social motivation' check-ins, at a very small privacy price, as compared to the baseline model. The results also indicate that for the 'private motivation' check-ins, the model provides 6% privacy than the baseline model for almost the same amount of information loss. Therefore, achieving the goals of retaining more information for social check-ins and providing more privacy for private check-ins.

CHAPTER 5

Future Work

5.1 Future Work 1: User-centric privacy of continuous queries in Mobile Edge Systems

5.1.1 Introduction

In autonomous driving applications like that in Tesla, when a driver chooses to use autodriving mode, the vehicle needs to process different information like lane condition, traffic, signs and make decisions like changing lanes, slowing down, changing routes, and many more within a very short time.A few of these operations like initial image processing might happen on the vehicle, but other intensive operations need to be forwarded to a server. If only one server is responsible for all the vehicles, it might cause delays in responses, which might fail the autodriving application. To avoid this situation, smaller servers called Edge Servers (ES) are deployed on the edge of the network, as shown in Figure 1. These servers only handle the cars (or other mobile devices) that are geographically closer to them, thus taking off much load of the core server and providing sooner results and better services.

As the ESs only handle vehicles/nodes within their range when the car moves. In this case, the private information might be available to multiple ESs simultaneously during communication exchanges and handovers [146, 147, 13]. Also, one must consider that several other MES mobile devices like navigation systems and mobile phones can simultaneously communicate with these servers. Coming up with privacy solutions in this scenario will be difficult as:

1. Different devices use different communication technologies. Therefore, it is challenging to develop an efficient and well-rounded privacy policy that handles discrepancies in multiple communication formats.

2. The presence of several ESs and the possibility of private information being available on multiples ESs at the same time might attract attackers to eavesdrop and cause breaches at the edge server level.

In [148], Mao et al. provide a very detailed account of the different privacy issues that arise from the heterogeneity of MES. They also list the different possible attacks on the various user information at the user level, the edge server level, and at the core server. The authors in [149] discuss the privacy concerns for each kind of application and the effect on the party involved. They also focus on other security concerns like Authentication issues, Denial of Service for different MES applications. Zhang et al. [150], provide a detailed account of the security threats to each component of an MES, the current research on handling these threats. They also emphasize that the privacy and security requirements of an MES keep evolving, and to this end, they list gaps in the current privacy preservation research and weak links of MES. In [151], the authors provide a simple noise addition model to preserve location privacy in an MES. They consider the scenarios of single LBS query and continuous LBS query (trajectory). This work evaluates the proposed model only in terms of the privacy reserved and the computation costs and overlooks utility which is crucial to be retained for any service. In [152], the authors propose a differential privacy-based method to preserve the location information of location data streams (continuous requests). Though differential

privacy methods have been proven to provide good utility [153], the query that is changed the most in a cluster of queries might result in temporary deviation of results provided, which might be undesirable in many services.

5.1.2 Goal

As evident from the literature, there are definitely some major gaps in location privacy research in MES, and it is necessary to address these issues to provide complete and balanced privacy to user's location. The goal of this project is to propose a privacy policy that addresses these gaps.

To address the privacy gaps, we will be considering multiple communication scenarios among different components of an MES and design separate mechanisms for each such scenario. This makes sure that the same privacy is not applied to the queries, which will avoid an attacker from back engineering the policy, based on the final results.

5.2 Future Work 2: Comprehensive analysis of user privacy issues in Digital Twin Networks

5.2.1 Introduction

Digital twins and digital twin networks have been gaining a lot of traction in the last few years from both industry and academia [154, 155]. Do the idea of digital twins might seem like something new and more recent, it was first introduced by Michael Greives in 2002 [156] at the University of Michigan. He defined the digital twin as "As a set of virtual information constructs that fully describes a potential or natural physical manufactured product from

the micro atomic level to the macro geometric level". Therefore, digital twins can be as descriptive as possible or as high level as possible based on the application scenario.

Most research focuses on DTs and pays less attention to the network that actually helps realize these digital twins. Digital Twin Networks (DTNs) provide the backbone architecture to establish and maintain DTs and provide DT services. It provides a communication network to gather new physical twin data, allowing DT to evolve. DTNs also allow network operators to design network optimization solutions, perform troubleshooting, or plan network upgrades taking into account the network's expected user growth. The key thing to notice is that all the DTN communications happen over ubiquitous networks, and due to the kind of services in use today,these communications are more frequent (frequent sampling of the PT to maintain DT's freshness) and involve a lot of rather sensitive information (to provide specific services). Therefore, many privacy concerns come into the picture as DTs and DTNs become more widespread, mainly because, in most cases, the data is directly collected from the end- users, who are naive and might not understand the nature and implications of privacy breaches. For example, a breach in the data collected by a fitness tracker or smartwatch might reveal a user's personal information like gender, daily activities, and health data. Similarly, a breach in the supply chain DTNs can disrupt the entire supply chain by allowing the attackers to directly manipulate the state of the respective physical components. Currently, there are not many research works that cover digital twin networks as a whole and even fewer that actually talk about the privacy issues that are introduced with the widespread implementation of DTNs

5.2.2 Goal

The goal is to address this gap in literature. Following are the things we plan on handling as part of this survey:

1. Provide a general formulation of DTNs, that can be adapted for most DTN applications, irrespective of the applied use case, and provide commonly used metrics in DTNs.

2. Investigate DTN applications in vehicular and aviation networks, 6G networks, healthcare, and manufacturing and supply chain management.

3. Emphasize on the privacy issues introduced by DTNs in the applications mentioned earlier and also how they mitigate pre-existing privacy issues.

4. Finally, suggest techniques and tools like federated learning [157] and blockchain [158], that can help overcome the said privacy issues in the DTN setting.

5.3 Future Work 3: User privacy in Federated Learning aided DTN

5.3.1 Introduction

With the development of Industrial IoT (IIoT) applications like IoV, smart cities, smart grids, and many more, companies are turning towards crowdsourcing to incentivize global communities to work for a common goal [159, 160, 161]. This enables support for advanced collaboration among smart products, services, users, and service providers. Since these services have complex hierarchies and deal with multimodal data types like queries, and

sensor readings spanning long periods, the operations require many computation resources [162]. DTNs have become an important way of realizing these services, as one can tap into digital resources, train accurate models, and aid in the co-evolution of the physical and virtual space. One major disadvantage of DTNs is that they rely on constant data streams to support the DT mapping, the level of interaction, and the amount of user information involved, which discourages users due to the privacy implications [163, 164]. A Federated Learning (FL) platform is a data science system developed for dispersed and non-centralized data. FL approaches enable enterprises to utilize their data together to cooperatively train models without explicitly sharing or centralizing their data, most introduced and used by Google in 2016; subsequently, it has been widely used in different research fields. Several research articles have already studied and addressed privacy issues introduced by a Federated learning approach in different applications [165]; therefore, a Federated learning added DTN has been gaining much traction for IIoT applications recently.

5.3.2 Goal

The goal of our research is to devise comprehensive solutions that focus on addressing user privacy at the following two levels:

1. Intra-twin : Intra twin communications refers to a communication between a physical twin and its corresponding digital twin. Data breaches in this communication impacts the safe operation of the DT. The adversary can manipulate the information, to maliciously influence the workflow.

2. Inter-twin communication : This includes communications among the DTs. Since there is a lot mutual information sharing among DTs, there is a chance of exposing a sensitive user information. As DT reside in the virtual space, the security and privacy of DTs, heavily relies on the cloud security. Cloud platforms are known to be vulnerable to attacks like Sybil attack, where the adversary can manipulate the DT to provide a user's location for example, and also DDoS attacks, resulting in service paralysis.

CHAPTER 6

Conclusion

This dissertation conducts research on the problem of user privacy in different IoT networks. Concerning this problem, we focus on the different kinds of user data being used in the respective application services and its utilization and what are the privacy threats that are posed by this kind of usage.

First, we identify or categorize the most common types of user data that a lot of the IoT services utilize, and then we conduct an in-depth analysis of the different privacy threats that are associated with each of these categories. We then provide a very extensive categorization of the different kinds of solutions in already existing research that address the said privacy concerns. The aim is to provide a comprehensive understanding of general user privacy in IoT networks, which is crucial to designing sound privacy mechanisms.

In this process, we identified a gap in research where privacy mechanisms only consider satisfying the third-party application and not necessarily the network users. To address this issue, we propose a privacy model that is based on user motivation for preserving the user check-in data on social networks. As part of this model, our primary focus was to mathematically characterize what user motivation is concerning check-ins and then create an adaptive privacy policy that considers each check-in as its own and applies different privacy to each check-in instead of considering all the check-ins of a user with similar level of importance, which is the norm in most user check-in privacy models. The comprehensive experimental results show that our model outperforms the other baseline models in terms of

user satisfaction-based utility metrics and, at the same time, preserves enough privacy for those check-ins that the user might seem highly private.

The proposed models are evaluated on real-world datasets and against other baselines models. We also discuss some future directions for privacy analysis in other IoT networks like Mobile Edge Systems and Digital Twin Networks. This dissertation, on the whole, provides a very comprehensive understanding on existing privacy issues, concerning user data, and some solutions to address them from a unique user perspective.

ACKNOWLEDGMENTS

This dissertation is partly supported by the National Science Foundation (NSF) under grant NOs. 1912753, 1704287, 1829674, 1741277, 2011845.

REFERENCES

[1] Y. Kabalci, E. Kabalci, S. Padmanaban, J. B. Holm-Nielsen, and F. Blaabjerg, "Internet of things applications as energy internet in smart grids and smart environments," *Electronics*, vol. 8, no. 9, p. 972, 2019.

[2] Z. Cai and T. Shi, "Distributed query processing in the edge-assisted iot data monitoring system," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 679–12 693, 2020.

[3] X. Zheng, Z. Cai, and Y. Li, "Data linkage in smart internet of things systems: a consideration from a privacy perspective," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 55–61, 2018.

[4] M. Wang, Y. Wang, A. M. V. V. Sai, Z. Liu, Y. Gao, X. Tong, and Z. Cai, "Task assignment for hybrid scenarios in spatial crowdsourcing: A q-learning-based approach," *Applied Soft Computing*, vol. 131, p. 109749, 2022.

[5] Q. Zhang, Y. Wang, G. Yin, X. Tong, A. M. V. V. Sai, and Z. Cai, "Two-stage bilateral online priority assignment in spatio-temporal crowdsourcing," *IEEE Transactions on Services Computing*, 2022.

[6] K. Li, L. Tian, W. Li, G. Luo, and Z. Cai, "Incorporating social interaction into three-party game towards privacy protection in iot," *Computer Networks*, vol. 150, pp. 90–101, 2019.

[7] C. Jang, J. Han, A. M. V. V. Sai, Y. Li, and O. Yi, "A study on scalar multiplication

parallel processing for x25519 decryption of 5g core network sidf function for mmtc iot environment," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[8] S. Ojagh, M. R. Malek, S. Saeedi, and S. Liang, "A location-based orientation-aware recommender system using iot smart devices and social networks," *Future Generation Computer Systems*, vol. 108, pp. 97–118, 2020.

[9] X. Xu, X. Zhang, H. Gao, Y. Xue, L. Qi, and W. Dou, "Become: Blockchain-enabled computation offloading for iot in mobile edge computing," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4187–4195, 2019.

[10] N. Tariq, M. Asim, F. Al-Obeidat, M. Zubair Farooqi, T. Baker, M. Hammoudeh, and I. Ghafir, "The security of big data in fog-enabled iot applications including blockchain: A survey," *Sensors*, vol. 19, no. 8, p. 1788, 2019.

[11] Q. Luo, D. Yu, A. M. V. V. Sai, Z. Cai, and X. Cheng, "A survey of structural representation learning for social networks," *Neurocomputing*, vol. 496, pp. 56–71, 2022.

[12] Z. B. Celik, E. Fernandes, E. Pauley, G. Tan, and P. McDaniel, "Program analysis of commodity iot applications for security and privacy: Challenges and opportunities," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–30, 2019.

[13] H. Xu, W. Li, and Z. Cai, "Analysis on methods to effectively improve transfer learning performance," *Theoretical Computer Science*, vol. 940, pp. 90–107, 2023.

[14] Z. Xiong and W. Li, "Federated generative model on multi-source heterogeneous data in iot," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[15] X. Zheng, L. Tian, G. Luo, and Z. Cai, "A collaborative mechanism for private data

publication in smart cities," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 7883–7891, 2020.

[16] Z. Cai and Z. He, "Trading private range counting over big iot data," in *2019 IEEE 39th international conference on distributed computing systems (ICDCS)*. IEEE, 2019, pp. 144–153.

[17] A. John, "Equifax data breach: What consumers need to know." [Online]. Available: shorturl.at/IJPSZ

[18] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, 2000.

[19] N. Talukder and S. I. Ahamed, "Preventing multi-query attack in location-based services," in *Proceedings of the third ACM conference on Wireless network security*, 2010, pp. 25–36.

[20] F. Li, Y. He, B. Niu, H. Li, and H. Wang, "Match-more: An efficient private matching scheme using friends-of-friends' recommendation," in *2016 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2016, pp. 1–6.

[21] K. H. Kwon and J. Shakarian, "Black-hat hackers' crisis information processing in the darknet: A case study of cyber underground market shutdowns," in *Networks, Hacking, and Media–CITA MS@ 30: Now and Then and Tomorrow*. Emerald Publishing Limited, 2018.

[22] Z. Xiong, Z. Cai, Q. Han, A. Alrawais, and W. Li, "Adgan: Protect your location privacy in camera data of auto-driving vehicles," *IEEE Transactions on Industrial*

Informatics, vol. 17, no. 9, pp. 6200–6210, 2020.

[23] J. Wang, Z. Cai, Y. Li, D. Yang, J. Li, and H. Gao, "Protecting query privacy with differentially private k-anonymity in location-based services," *Personal and Ubiquitous Computing*, vol. 22, pp. 453–469, 2018.

[24] Y. Wang, Z. Cai, X. Tong, Y. Gao, and G. Yin, "Truthful incentive mechanism with location privacy-preserving for mobile crowdsourcing systems," *Computer Networks*, vol. 135, pp. 32–43, 2018.

[25] C.-Y. Chow, "Cloaking algorithms for location privacy." 2008.

[26] P. Wightman, W. Coronell, D. Jabba, M. Jimeno, and M. Labrador, "Evaluation of location obfuscation techniques for privacy in location based information systems," in *2011 IEEE Third Latin-American Conference on Communications*. IEEE, 2011, pp. 1–6.

[27] T. N. Phan, T. K. Dang, T. A. Truong, and T. H. Lam, "A context-aware privacy-preserving solution for location-based services," in *2018 International Conference on Advanced Computing and Applications (ACOMP)*. IEEE, 2018, pp. 132–139.

[28] C.-Y. Chow, M. F. Mokbel, and X. Liu, "Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments," *GeoInformatica*, vol. 15, no. 2, pp. 351–380, 2011.

[29] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proceedings of the 1st international conference on Mobile systems, applications and services*, 2003, pp. 31–42.

[30] X. Zheng, G. Luo, and Z. Cai, "A fair mechanism for private data publication in online social networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 880–891, 2018.

[31] Z. Cai, X. Zheng, J. Wang, and Z. He, "Private data trading towards range counting queries in internet of things," *IEEE Transactions on Mobile Computing*, 2022.

[32] X. Zheng, Z. Cai, J. Yu, C. Wang, and Y. Li, "Follow but no track: Privacy preserved profile publishing in cyber-physical social systems," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1868–1878, 2017.

[33] N. Mattani, J. S. Kumar, A. Prabakaran, and N. Maheswari, "Privacy preservation in social network analysis using edge weight perturbation," *Indian Journal of Science and Technology*, vol. 9, no. 37, pp. 1–10, 2016.

[34] A. Campan and T. M. Truta, "Data and structural k-anonymity in social networks," in *International Workshop on Privacy, Security, and Trust in KDD*. Springer, 2008, pp. 33–54.

[35] X. Zheng, G. Luo, and Z. Cai, "A fair mechanism for private data publication in online social networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 880–891, 2018.

[36] M. Hazazi, Y. Tian, and M. Al-Rodhaan, "Privacy-preserving authentication scheme for wireless networks," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*. IEEE, 2018, pp. 1–6.

[37] S. Zhang, G. Wang, M. Z. A. Bhuiyan, and Q. Liu, "A dual privacy preserving scheme

in continuous location-based services," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 4191–4200, 2018.

[38] S. Jahid, P. Mittal, and N. Borisov, "Easier: Encryption-based access control in social networks with efficient revocation," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, 2011, pp. 411–415.

[39] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2018.

[40] M. Li, H. Zhu, Z. Gao, S. Chen, L. Yu, S. Hu, and K. Ren, "All your location are belong to us: Breaking mobile social networks for automated user location tracking," in *Proceedings of the 15th ACM international symposium on Mobile ad hoc networking and computing*, 2014, pp. 43–52.

[41] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 646–660, 2016.

[42] Z. Cai, X. Zheng, and J. Yu, "A differential-private framework for urban traffic flows estimation via taxi companies," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.

[43] Z. Xiong, H. Xu, W. Li, and Z. Cai, "Multi-source adversarial sample attack on autonomous vehicles," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2822–2835, 2021.

[44] B. Lee, J. Oh, H. Yu, and J. Kim, "Protecting location privacy using location semantics," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1289–1297.

[45] C.-Y. Chow and M. F. Mokbel, "Enabling private continuous queries for revealed user locations," in *International Symposium on Spatial and Temporal Databases*. Springer, 2007, pp. 258–275.

[46] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li, "Deep learning based inference of private information using embedded sensors in smart devices," *IEEE Network*, vol. 32, no. 4, pp. 8–14, 2018.

[47] J. Krumm, "Inference attacks on location tracks," in *International Conference on Pervasive Computing*. Springer, 2007, pp. 127–143.

[48] Z. Xiong, Z. Cai, C. Hu, D. Takabi, and W. Li, "Towards neural network-based communication system: Attack and defense," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–14, 2022.

[49] M. Gill, "Geo-spoofing explained," Jan. 2020. [Online]. Available: https://www.comparitech.com/blog/vpn-privacy/geospoofing/

[50] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2016.

[51] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PloS one*, vol. 6, no. 12, p. e28071, 2011.

[52] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, no. 1, pp. 1–5, 2013.

[53] M. Maouche, S. B. Mokhtar, and S. Bouchenak, "Ap-attack: a novel user re-identification attack on mobility datasets," in *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2017, pp. 48–57.

[54] X. Zheng, Z. Cai, and Y. Li, "Data linkage in smart internet of things systems: a consideration from a privacy perspective," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 55–61, 2018.

[55] G. Vasanthakumar, K. Sunithamma, P. Deepa Shenoy, and K. Venugopal, "An overview on user profiling in online social networks," *Int. J. Appl. Inf. Syst*, vol. 11, no. 8, pp. 25–42, 2017.

[56] I. Illascu, "Targeted attacks : User profiling," Sept. 2014. [Online]. Available: https://news.softpedia.com/news/User-Profiling-Services-Could-Be-Compromised-for-More-Targeted-Attacks-460284.shtml

[57] E. Tannam, "Phishing ad claiming to be Twitter verification service," May. 2018. [Online]. Available: https://www.siliconrepublic.com/enterprise/twitter-ad-verification

[58] "The rising threat of social media phishing attacks," Oct. 2018. [Online]. Available: https://corrata.com/the-rising-threat-of-social-media-phishing-attacks/

[59] R. Dumont, "Adidas "prize" used as bait in attempt to lure people into biting," Jun. 2018. [Online]. Available: https://www.welivesecurity.com/2018/06/14/phishing-anniversary-free-50-month-subscription/

[60] E. Limer, "Eavesdrop through Phone Microphones to target ads," Jan. 2018. [Online]. Available: https://www.popularmechanics.com/technology/security/a14533262/alphonso-audio-ad-targeting/

[61] A. Langone, "How Facebook or any other App could use your phone's microphone to gather data," Mar. 2018. [Online]. Available: http://money.com/money/5219041/how-to-turn-off-phone-microphone-facebook-spying/

[62] T. Kohut, "Our phones are listening to us," Feb. 2018. [Online]. Available: https://globalnews.ca/news/4039276/smart-devices-facebook-listening/

[63] Y.-C. Hu, A. Perrig, and D. B. Johnson, "Wormhole attacks in wireless networks," *IEEE journal on selected areas in communications*, vol. 24, no. 2, pp. 370–380, 2006.

[64] M. Kumar, "WireX DDoS Botnet"," Aug. 2017. [Online]. Available: https://thehackernews.com/2017/08/android-ddos-botnet.html

[65] A. Davies, "Akamai's verdict on Mirai," Nov. 2016. [Online]. Available: https://disruptive.asia/akamai-assesses-mirai-bad-worse/

[66] S. Gallagher, "Syrian rebels lured into malware honeypot sites," Feb. 2015. [Online]. Available: https://arstechnica.com/information-technology/2015/02/syrian-rebels-lured-into-malware-honeypot-sites-through-sexy-online-chats/

[67] Z. He, Y. Lin, Y. Liang, X. Wang, A. M. Vera Venkata Sai, and Z. Cai, "Modeling

malware propagation dynamics and developing prevention methods in wireless sensor networks," *Nonlinear Combinatorial Optimization*, pp. 231–250, 2019.

[68] "Social Media Scams Based on Current Events," Aug. 2014. [Online]. Available: https://community.norton.com/blogs/norton-protection-blog/social-media-scams-based-current-events

[69] R. John, J. P. Cherian, and J. J. Kizhakkethottam, "A survey of techniques to prevent sybil attacks," in *2015 International Conference on Soft-Computing and Networks Security (ICSNS)*. IEEE, 2015, pp. 1–6.

[70] K. Zhang, X. Liang, R. Lu, and X. Shen, "Sybil attacks and their defenses in the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 5, pp. 372–383, 2014.

[71] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection," in *22nd USENIX Security Symposium (USENIX Security 13)*, 2013, pp. 241–256.

[72] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial iots," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.

[73] J. Wang, Z. Cai, and J. Yu, "Achieving personalized k-anonymity-based content privacy for autonomous vehicles in cps," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4242–4251, 2019.

[74] X. Xiao, C. Chen, A. K. Sangaiah, G. Hu, R. Ye, and Y. Jiang, "Cenlocshare: A centralized privacy-preserving location-sharing system for mobile online social networks,"

Future Generation Computer Systems, vol. 86, pp. 863–872, 2018.

[75] T. Peng, Q. Liu, D. Meng, and G. Wang, "Collaborative trajectory privacy preserving scheme in location-based services," *Information Sciences*, vol. 387, pp. 165–179, 2017.

[76] S. Zhang, G. Wang, Q. Liu, and J. H. Abawajy, "A trajectory privacy-preserving scheme based on query exchange in mobile social networks," *Soft Computing*, vol. 22, no. 18, pp. 6121–6133, 2018.

[77] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: Attacks and defense," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 646–660, 2016.

[78] K. Huguenin, I. Bilogrevic, J. S. Machado, S. Mihaila, R. Shokri, I. Dacosta, and J.-P. Hubaux, "A predictive model for user motivation and utility implications of privacy-protection mechanisms in location check-ins," *IEEE Transactions on Mobile Computing*, vol. 17, no. 4, pp. 760–774, 2017.

[79] G. Li, Z. Cai, G. Yin, Z. He, and M. Siddula, "Differentially private recommendation system based on community detection in social network applications." *Security & Communication Networks*, 2018.

[80] Z. He, A. M. V. V. Sai, Y. Huang, D. Seo, H. Zhang, and Q. Han, "Differentially private approximate aggregation based on feature selection," *Journal of Combinatorial Optimization*, vol. 41, pp. 318–327, 2021.

[81] R. Dewri, "Local differential perturbations: Location privacy under approximate knowledge attackers," *IEEE Transactions on Mobile Computing*, vol. 12, no. 12, pp.

2360–2372, 2012.

[82] "Laplace distribution," in *Encyclopedia of Mathematics*, Aug. 2014. [Online]. Available: http://encyclopediaofmath.org/index.php?title=Laplace_distribution&oldid=33035

[83] C. Yin, J. Xi, R. Sun, and J. Wang, "Location privacy protection based on differential privacy strategy for big data in industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3628–3636, 2017.

[84] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 2007, pp. 94–103.

[85] W. Wang, M. Min, L. Xiao, Y. Chen, and H. Dai, "Protecting semantic trajectory privacy for vanet with reinforcement learning," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–5.

[86] X. Zheng, Z. Cai, G. Luo, L. Tian, and X. Bai, "Privacy-preserved community discovery in online social networks," *Future Generation Computer Systems*, vol. 93, pp. 1002–1009, 2019.

[87] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li, "Deep learning based inference of private information using embedded sensors in smart devices," *IEEE Network*, vol. 32, no. 4, pp. 8–14, 2018.

[88] T. A. Campan and T. Truta, "A clustering approach for data and structural anonymity," in *In Proceedings of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'08), in Conjunction with KDD*, vol. 8,

2008.

[89] M. Siddula, Y. Li, X. Cheng, Z. Tian, and Z. Cai, "Anonymization in online social networks based on enhanced equi-cardinal clustering," vol. 6, no. 4, 2019, pp. 809–820.

[90] M. Li, N. Cao, S. Yu, and W. Lou, "Findu: Privacy-preserving personal profile matching in mobile social networks," in *2011 Proceedings IEEE INFOCOM*. IEEE, 2011, pp. 2435–2443.

[91] R. Li, H. Li, X. Cheng, X. Zhou, K. Li, S. Wang, and R. Bie, "Perturbation-based private profile matching in social networks," *IEEE Access*, vol. 5, pp. 19 720–19 732, 2017.

[92] A. Pingley, W. Yu, N. Zhang, X. Fu, and W. Zhao, "A context-aware scheme for privacy-preserving location-based services," *Computer Networks*, vol. 56, no. 11, pp. 2551–2568, 2012.

[93] Y. Wang, X. Tao, F. Zhao, B. Tian, and A. M. Vera Venkata Sai, "Sla-aware resource scheduling algorithm for cloud storage," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, pp. 1–10, 2020.

[94] L. Zhu, C. Zhang, C. Xu, X. Du, R. Xu, K. Sharif, and M. Guizani, "Prif: A privacy-preserving interest-based forwarding scheme for social internet of vehicles," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2457–2466, 2018.

[95] P. Syverson, D. Goldschlag, and M. Reed, "Onion routing for anonymous and private internet connections," *Communications of the ACM*, vol. 42, no. 2, p. 5, 1999.

[96] I. Aad, C. Castelluccia, and J.-P. Hubaux, "Packet coding for strong anonymity in ad

hoc networks," in *2006 Securecomm and Workshops*. IEEE, 2006, pp. 1–10.

[97] O. Hasan, J. Miao, S. B. Mokhtar, and L. Brunie, "A privacy preserving prediction-based routing protocol for mobile delay tolerant networks," in *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2013, pp. 546–553.

[98] "Facebook dataset." [Online]. Available: https://snap.stanford.edu/data/ego-Facebook.html

[99] "Twitter dataset." [Online]. Available: https://snap.stanford.edu/data/ego-Twitter.html

[100] J. Leskovec and A. Krevl, "Snap datasets: Stanford large network dataset collection," 2014. [Online]. Available: http://snap.stanford.edu/data

[101] R. S. B. Rozemberczki, R. Davis and C. Sutton, "Gemsec: Graph embedding with self clustering," 2018. [Online]. Available: https://arxiv.org/abs/1802.03997

[102] "Deezer dataset." [Online]. Available: https://snap.stanford.edu/data/gemsec-Deezer.html

[103] E. Cho, S. A. Myers, and J. Leskovec, "Brightkite dataset." [Online]. Available: https://snap.stanford.edu/data/loc-Brightkite.html

[104] "Brighkite dataset," [Online]. Available: https://snap.stanford.edu/data/loc-Brightkite.html.

[105] "Gowalla dataset." [Online]. Available: http://www.yongliu.org/datasets/index.html

[106] Y. Liu, W. Wei, A. Sun, and C. Miao, "Exploiting geographical neighborhood

characteristics for location recommendation." New York, NY, USA: Association for Computing Machinery, 2014. [Online]. Available: https://doi.org/10.1145/2661829.2662002

[107] R. Zafarani and H. Liu, "Social computing data repository at ASU," 2009. [Online]. Available: http://socialcomputing.asu.edu

[108] Y. Liu, "Weeplaces dataset." [Online]. Available: https://www.yongliu.org/datasets.html

[109] "Geocities dataset." [Online]. Available: https://www.microsoft.com/en-us/download/details.aspx?id=52367

[110] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 791–800.

[111] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 312–321.

[112] H. S. Y. Chon, E. Talipov and H. Cha, "Crawdad dataset yonsei/lifemap," https://crawdad.org/yonsei/lifemap/20120103 and https://doi.org/10.15783/C7RW26.

[113] Y. Chon, E. Talipov, H. Shin, and H. Cha, "Mobility prediction-based smartphone energy optimization for everyday location monitoring," in *Proceedings of the 9th ACM conference on embedded networked sensor systems*, 2011, pp. 82–95.

[114] J. Chon and H. Cha, "Lifemap: A smartphone-based context provider for location-

based services," *IEEE Pervasive Computing*, vol. 10, no. 2, pp. 58–67, 2011.

[115] Y. Chon, H. Shin, E. Talipov, and H. Cha, "Evaluating mobility models for temporal prediction with high-granularity mobility data," in *2012 IEEE International Conference on Pervasive Computing and Communications*. IEEE, 2012, pp. 206–212.

[116] "T-drive dataset." [Online]. Available: https://privamov.github.io/accio/docs/datasets.html#t-drive

[117] N.-D. M. Piorkowski and M.Grossglauser, "Crawdad data set epfl/mobility," 2009. [Online]. Available: https://crawdad.org/epfl/mobility/20090224/cab

[118] "Manhattan taxi trajectory dataset." [Online]. Available: https://www.cs.cornell.edu/~arb/data/Manhattan-taxi-trajectories/index.html

[119] C. Whong, "Foiling nyc's taxi trip data." [Online]. Available: https://chriswhong.com/open-data/foil_nyc_taxi/

[120] A. R. Benson, D. F. Gleich, and L.-H. Lim, "The spacey random walk: A stochastic process for higher-order data," *SIAM Review*, vol. 59, no. 2, pp. 321–345, 2017.

[121] "Mackaroo - online data generator." [Online]. Available: https://mockaroo.com

[122] "Generatedata - online data generator." [Online]. Available: http://www.generatedata.com

[123] "DTMGenerator - online data generator." [Online]. Available: https://www.rankred.com/test-data-generation-tools/

[124] N. Deshdeep, "App or website? 10 reasons why apps are better," accessed: May. 24, 2021. [Online]. Available: https://vwo.com/blog/10-reasons-mobile-apps-are-better/.

[125] "Facebook market place," [Online]. Available: https://www.facebook.com/marketplace/.

[126] DCI, "Facebook places for business: Location-based networking at a new level," accessed: Sep. 07, 2010. [Online]. Available: https://www.dotcominfoway.com/facebook-places-for-business-location-based-networking-at-a-new-level/#gref.

[127] A. Carman, "Why do you share your location?" accessed: Dec. 19, 2017. [Online]. Available: https://www.theverge.com/2017/12/19/16792336/location-sharing-apps-privacy-whyd-you-push-that-button-podcast.

[128] A. R. Shahid, N. Pissinou, S. S. Iyengar, and K. Makki, "Check-ins and photos: Spatiotemporal correlation-based location inference attack and defense in location-based social networks," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, 2018, pp. 1852–1857.

[129] A. M. V. Venkata Sai and Y. Li, "A survey on privacy issues in mobile social networks," *IEEE Access*, vol. 8, pp. 130 906–130 921, 2020.

[130] Z. He, Z. Cai, J. Yu, X. Wang, Y. Sun, and Y. Li, "Cost-efficient strategies for restraining rumor spreading in mobile social networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2789–2800, 2016.

[131] M. Siddula, Y. Li, X. Cheng, Z. Tian, and Z. Cai, "Privacy-enhancing preferential lbs query for mobile social network users," *Wireless Communications and Mobile Computing*, vol. 2020, 2020.

[132] T. Anwar, K. Liao, A. Goyal, T. Sellis, A. Kayes, and H. Shen, "Inferring location types with geo-social-temporal pattern mining," *IEEE Access*, vol. 8, pp. 154 789–154 799, 2020.

[133] A. Noulas, C. Mascolo, and E. Frias-Martinez, "Exploiting foursquare and cellular data to infer user activity in urban environments," in *2013 IEEE 14th International Conference on Mobile Data Management*, vol. 1. IEEE, 2013, pp. 167–176.

[134] A. M. V. Venkata Sai, K. Zhang, and Y. Li, "User motivation based privacy preservation in location based social networks," in *2021 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*, 2021, pp. 471–478.

[135] M. Madejski, M. Johnson, and S. M. Bellovin, "A study of privacy settings errors in an online social network," in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 2012, pp. 340–345.

[136] Y. Yao, J. Zhu, J. Liu, and N. N. Xiong, "Verifiable and privacy-preserving check-ins for geo-social networks," *Journal of Internet Technology*, vol. 19, no. 4, pp. 969–980, 2018. [Online]. Available: https://jit.ndhu.edu.tw/article/view/1716

[137] A. R. Shahid, N. Pissinou, S. Iyengar, and K. Makki, "Check-ins and photos: Spatiotemporal correlation-based location inference attack and defense in location-based social networks," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On*

Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2018, pp. 1852–1857.

[138] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2016.

[139] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 665–673, 2017.

[140] L. Yu, S. M. Motipalli, D. Lee, P. Liu, H. Xu, Q. Liu, J. Tan, and B. Luo, "My friend leaks my privacy: Modeling and analyzing privacy in social networks," in *Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies*, 2018, pp. 93–104.

[141] "Inversion bracket notation," [Online]. Available: https://en.wikipedia.org/wiki/Iverson_bracket.

[142] F. Wang, G. Wang, and P. S. Yu, "Why checkins: Exploring user motivation on location based social networks," in *2014 IEEE International Conference on Data Mining Workshop*, 2014, pp. 27–34.

[143] T. Spiliotopoulos and I. Oakley, "Understanding motivations for facebook use: Usage metrics, network structure, and privacy," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 3287–3296.

[144] I. Bilogrevic, K. Huguenin, S. Mihaila, R. Shokri, and J.-P. Hubaux, "Predicting users'

motivations behind location check-ins and utility implications of privacy protection mechanisms," in *22nd Network and Distributed System Security Symposium (NDSS)*, 2015.

[145] D. Yang, D. Zhang, B. Qu, and P. Cudré-Mauroux, "Privcheck: Privacy-preserving check-in data publishing for personalized location based services," ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 545–556. [Online]. Available: https://doi.org/10.1145/2971648.2971685

[146] H. Xu, Z. Cai, D. Takabi, and W. Li, "Audio-visual autoencoding for privacy-preserving video streaming," *IEEE Internet of Things Journal*, vol. 9, no. 3, pp. 1749–1761, 2021.

[147] H. Xu, Z. Cai, and W. Li, "Privacy-preserving mechanisms for multi-label image recognition," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 4, pp. 1–21, 2022.

[148] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE communications surveys & tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[149] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2017.

[150] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: Survey and open issues," *IEEE access*, vol. 6, pp. 18 209–18 237, 2018.

[151] Y. Wang, Z. Tian, S. Su, Y. Sun, and C. Zhu, "Preserving location privacy in mobile

edge computing," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[152] V. Stephanie, M. Chamikara, I. Khalil, and M. Atiquzzaman, "Privacy-preserving location data stream clustering on mobile edge computing and cloud," *Information Systems*, vol. 107, p. 101728, 2022.

[153] D. J. Mir, *Differential privacy: an exploration of the privacy-utility landscape*. Rutgers The State University of New Jersey-New Brunswick, 2013.

[154] C. Wang and Y. Li, "Digital-twin-aided product design framework for iot platforms," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9290–9300, 2021.

[155] C. Wang, Z. Cai, and Y. Li, "Sustainable blockchain-based digital twin management architecture for iot devices," *IEEE Internet of Things Journal*, 2022.

[156] C. Miskinis. (2019) The history and creation of the digital twin concept. [Online]. Available: https://www.challenge.org/insights/digital-twin-history/

[157] F. Zhao, Y. Huang, A. M. V. V. Sai, and Y. Wu, "A cluster-based solution to achieve fairness in federated learning," in *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2020, pp. 875–882.

[158] S. Zhu, W. Li, H. Li, L. Tian, G. Luo, and Z. Cai, "Coin hopping attack in blockchain-based iot," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4614–4626, 2018.

[159] Y. Lin, Z. Cai, X. Wang, F. Hao, L. Wang, and A. M. V. V. Sai, "Multi-round

incentive mechanism for cold start-enabled mobile crowdsensing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 993–1007, 2021.

[160] K. Yan, G. Luo, X. Zheng, L. Tian, and A. M. V. V. Sai, "A comprehensive location-privacy-awareness task selection mechanism in mobile crowd-sensing," *IEEE access*, vol. 7, pp. 77 541–77 554, 2019.

[161] K. Yan, G. Lu, G. Luo, X. Zheng, L. Tian, and A. M. V. V. Sai, "Location privacy-aware task bidding and assignment for mobile crowd-sensing," *IEEE Access*, vol. 7, pp. 131 929–131 943, 2019.

[162] J. Li, A. M. V. V. Sai, X. Cheng, W. Cheng, Z. Tian, and Y. Li, "Sampling-based approximate skyline query in sensor equipped iot networks," *Tsinghua Science and Technology*, vol. 26, no. 2, pp. 219–229, 2020.

[163] Y. Huang, Y. J. Li, and Z. Cai, "Security and privacy in metaverse: A comprehensive survey," *Big Data Mining and Analytics*, vol. 6, no. 2, pp. 234–247, 2023.

[164] Z. Zhan, Y. Wang, P. Duan, A. M. V. V. Sai, Z. Liu, C. Xiang, X. Tong, W. Wang, and Z. Cai, "Efficient recruitment strategy for collaborative mobile crowd sensing based on gcn trustworthiness prediction," *arXiv preprint arXiv:2306.04366*, 2023.

[165] S. Jamil, M. Rahman *et al.*, "A comprehensive survey of digital twins and federated learning for industrial internet of things (iiot), internet of vehicles (iov) and internet of drones (iod)," *Applied System Innovation*, vol. 5, no. 3, p. 56, 2022.