A Geospatial Approach to Preserving Location Privacy

by

Aditya Tadakaluru

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 06, 2023

Keywords: location privacy, dummy locations, geospatial, building footprint entropy, semantic balance, temporal elimination

Approved by

Xiao Qin, Professor of Computer Science and Software Engineering
Gerry Dozier, Professor of Computer Science and Software Engineering
Farah Kandah, Associate Professor of Computer Science and Software Engineering
Akond Rahman, Assistant Professor of Computer Science and Software Engineering
Elvan Ceyhan, Associate Professor of Mathematics and Statistics

Abstract

Sharing true locations of users has become a basic requirement for accessing Location-based services (LBS) on a wide range of web and mobile applications. LBS require users to provide their current location for service delivery and customization. The potential for misuse of true location information by LBS providers and security risks associated with location information falling into wrong hands warrant a pressing need to protect online users' location privacy. Location privacy protection techniques address concerns associated with the potential mishandling of location information submitted to the LBS provider. Location accuracy has a direct impact on the quality of service (QoS), where higher location accuracy results in better QoS. In general, the main goal of any location privacy technique is to achieve maximum QoS while providing minimum or no location information if possible and using dummy locations is one such location privacy technique. However, most of the existing methods for generating dummy locations have problems addressing scenarios where the true location is part of a large parcel area or if the true location is in a remote area with no building structures nearby.

In the first part of this dissertation, we propose a novel context-optimized and spatial-aware (COSA) dummy locations generation framework for location privacy, built and evaluated on real-world geospatial data. We evaluated the proposed solution using real-city parcel data and outlined and geo-visualized the results at each step. In the second part of this dissertation, we propose a novel enhanced parcel-based location privacy framework - PLP+ - to construct spatially similar dummy locations anchored on the real-world spatial context of locations such as parcels, building footprints, and road proximity. Our results unveil that PLP+ successfully addresses the map elimination attack in the location set with up to 50 dummy locations by not placing the locations in vacant parcels.Also,there were no dummy

locations sharing the same parcel as their true locations out of the 500 dummy locations generated by PLP+, indicating the effectiveness of PLP+ against location homogeneity attack. We develop a novel parameter estimator algorithm for density-based clustering to identify spatial privacy zones within a city. The new algorithm is capable of curtailing the target search area for parcel similarity search from an entire city dataset of 123,848 parcels to a smaller privacy area of 31,412 parcels, with no statistically significant difference in search results. We devise a novel strategy to quantify location privacy by the virtue of building footprint entropy, and we demonstrate that dummy locations generated by PLP+ are consistently higher in footprint entropy offering better location privacy.

In the third part of this dissertation, we introduced a temporal constraint attack whereby an adversary can exploit the temporal constraints associated with the semantic category of locations to eliminate dummy locations and identify the true location. We demonstrated how an adversary can devise a temporal constraint attack to breach the location privacy of a residential location. We addressed this major limitation of current dummy approaches with a novel Voronoi-based semantically balanced framework (VSBDG) capable of generating dummy locations that can withstand a temporal constraint attack. Built based on real-world geospatial datasets, VSBDG framework leverages parcel-based similarity, spatial relationships, and operations. Our results show a high physical dispersion cosine similarity of 0.99 between the semantic categories even with larger location set sizes. This indicates a strong and scalable semantic balance for each semantic category within the VSBDG's output location set. The VSBDG algorithm is capable of producing location sets with high average minimum dispersion distance values of 5861.89 meters for residential locations and 6258.05 meters for POI locations. The findings demonstrate that the locations within each semantic category are scattered farther apart, entailing optimized location privacy.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Location privacy

Location-based services (LBS) are a ubiquitous phenomenon in our current times, where the user is provided with services, information, or both based on their current location [28]. It has become a standard functionality on mobile phones and websites requesting access to the user's current location so that geographically customized information or service options can be presented to the user. In many instances, the user's location information collected by these third-party apps is being stored and used for future use, including advertising and sales. Although LBS provides great benefits to the user through geographically tailored services, it also raises concerns regarding location privacy by requiring the user to reveal their true location to the service provider. In some cases, a compromised LBS server can pose a security risk where the user's true location can be exploited by an adversary.

The current approaches for achieving location privacy in an LBS environment can be classified into four main categories: (1) anonymization, (2) obfuscation, (3) encryption, and (4) dummy locations [23]. In both anonymization and obfuscation techniques, the true location information of the user is modified before sending it to the service provider [18] and may result in lower quality of service [23]. The encryption-based techniques levy substantial setup and storage overhead costs on the user, making it a less optimal choice. The dummy locations technique involves the user sending the dummy locations along with the true location to the LBS service provider, making it difficult for the service provider or the adversary to determine the true location of the user and thus guaranteeing location privacy [23].

1.2 Dummy Locations for Location privacy

The possibility for exploitation of location data by an LBS provider or data breach by an adversary opens the door for location privacy concerns. Although location sharing has become a necessity for receiving high-quality service in an LBS scenario, it is critical to achieving the maximum quality of service without sacrificing the location privacy of the users. Sending dummy locations alongside the true location to the LBS server is a well-explored approach for location privacy. The LBS server would not be able to - under a perfect scenario - distinguish true from dummy locations, so the server processes all locations and sends results back to the users to achieve location privacy. The dummy locations approach receives highly accurate LBS query results because users' true locations are delivered, thereby offering a better quality of service compared to other location privacy techniques [23].

The presumption that several dummy locations sent alongside the true location will make it difficult to discern the true location of the user from the dummy ones relies heavily on the process used to generate dummy locations and their relationship to the true location. Lu *et al.* [29] proposed generating dummies within rectangular or circular regions to avoid location identification by spatial elimination. Zhang *et al.* [60] proposed techniques to avoid identification using the commonalities within geographic semantics, such as generating all dummy locations within a hospital complex. Parmar and Rao [36] also discussed the risk of location identification based on the dissimilarities within geographic semantics; for instance, if the user location is in a building structure and generated dummy locations in natural features such as mountains and rivers. Based on the literature, the two key limitations of the existing approaches are outlined in Table 1.1.

The semantic type of the location can have a significant impact on the effectiveness of the dummy locations in protecting a true location from identification [10]. The conventional approach is to use semantic location diversity to identify dummy locations that are semantically different from each other and thus make it harder to distinguish between true and dummy locations. This approach doesn't work in scenarios such as when the true location

Table 1.1: Two key limitations and related location privacy scenarios

Location privacy scenario	Key limitation
User location is in a large parcel area such as a hospital, university, or corporate work campus	The dummy locations generated are usually within the same parcel or building compound, making them vulnerable to a "location homogeneity attack" [36], where true location can be easily identified because of the lack of diversity between true and dummy locations [36]
User location is a parcel or building structure in an area with few or no building structures in close proximity	The dummy locations generated are usually non-building structures and natural features such as rivers and green areas, making them vulnerable to a "map-matching attack" [36], where dummies can be spatially eliminated to identify the true location

is residential since the residential locations are inherently different from the non-residential POI locations such as the purpose of use and hours of operation. A dummy approach must handle these intrinsic differences between various semantic types to successfully protect the location privacy of the user. Otherwise, this could result in generating dummy locations that are susceptible to temporal constraint attacks. To the best of our knowledge, there is no known solution to this problem since the existing dummy approaches either do not acknowledge the semantic differences associated with handling a true location such as residence versus POI or do not factor in the semantic type of the location for dummy generation altogether.

1.3 Research Questions and Objectives

The main goal of this dissertation research is to develop a geospatial-driven dummy-based location privacy framework for preserving the location privacy of the user. This dissertation is composed of three studies namely Context optimized and spatial aware dummy locations generation framework for location privacy [1](COSA), An Enhanced Parcel-based Location Privacy Framework using - Building Footprint Entropy for Spatially Similar Dummy Locations (PLP+), and A Voronoi-based Semantically Balanced Dummy Generation Framework for Location Privacy [2] (VSBDG). Each study is discussed in detail in chapters 3, 4,

[1]First published in [Journal of Geovisualization and Spatial Analysis, Volume 6, Article 27, 2022] by Springer Nature [48]

[2]First published in [Analytics, Volume 2, Pages 246–264, 2023] [49]

and 5 respectively. The research questions and objectives pertaining to each of the aforementioned studies are presented separately in the following sub-sections.

1.3.1 Research Questions and Objectives of COSA

The COSA research study is aimed to answer the following research questions in the field of location privacy-

- Is it possible to generate dummy locations based on similarity in spatial context using real-world geospatial datasets?
- What are the impacts of using spatially similar dummy locations on location privacy preservation?

The primary objective of COSA is to build a location privacy framework to generate spatially similar dummies that are capable of withstanding location homogeneity and map-matching attacks. The secondary objective is to spatially enable the framework to preserve the indistinguishability of true locations by implementing COSA driven by real-world geospatial datasets. The third objective of this study is to forge COSA as an open reusable framework that allows plugging in similarity search criteria as needed to calibrate the dummy selection process for generating context-sensitive dummies.

1.3.2 Research Questions and Objectives of PLP+

The PLP+ research study is aimed to answer the following research questions in the field of location privacy -

- Is it possible to simultaneously address both map-matching and location-homogeneity attacks using dummy locations identified based on parcel-based similarity?
- Can spatially similar parcels be found by searching within a smaller area of interest instead of an entire county or city without a reduction in the similarity quality of the output parcels?

- What is an approach to quantify total location privacy achieved by a location set based on the geographical context of locations?

Our overarching goal in the PLP+ study is to develop a location privacy dummy location framework that generates dummy locations anchored on their geographical feature similarity to real locations. Our first objective is to devise a dummy approach to simultaneously addressing map-matching and location-homogeneity attacks, with which current dummy techniques fail in dealing. Secondly, since searching for spatially similar dummy locations is a computationally intensive task, our second objective is to bring forth a novel privacy-area generation approach, aiming to facilitate our search for contextually similar dummy locations in a fast and efficient manner.

Quantifying location privacy plays a key role in developing and gauging the efficacy of a location privacy approach. Currently, there is a limited number of studies addressing the quantification of dummy-aided location privacy. The COSA proposed in [48] is the first among its peers to propose a dummy generation framework based on spatial context and real-world geospatial data sets. As a consequence of this, there are no other studies that address quantifying dummy-aided location privacy that hinges on spatial context. The PLP+'s third design objective is to construct a location privacy quantification module that is slated to accurately measure the total location privacy achieved by a set of dummy locations and a true location. This goal is achieved through a metric computed through the location's spatial context.

1.3.3 Research Objectives of VSBDG

The VSBDG research study is aimed to answer the following research questions in the field of location privacy -

- What is the impact of using dummies that are unaware of the differences in temporal constraints between semantic types on the user's location privacy?

- For a given true location, how to generate a location set with dummy locations that are capable of withstanding time constraint attack?

The current dummy approaches overlook the distinction between various semantic categories and their intrinsic temporal constraints such as residential versus non-residential locations. Due to these limitations, the dummy locations produced are susceptible to temporal constraint attack by an adversary in scenarios where the true location of the user is residential. This erosion of location privacy protection for users or IoT devices whose legitimate locations are residential locations urges the need for a solution that can originate dummy locations capable of thwarting temporal constraint attacks by an adversary. In this paper, a novel approach to generate dummy locations that are effective against temporal constraint attack and preserves the location privacy of residential users is proposed and developed.

The main research objective of the VSBDG study is to devise a novel dummy generation framework to produce semantically balanced dummy locations that can withstand any temporal constraint attacks imposed by adversaries. Although the focus of this part of the dissertation study is on addressing temporal constraint attacks in the case of users whose legitimate locations are a residence, the VSBDG framework is capable of furnishing comprehensive location privacy for true locations of both residential and non-residential POI semantic types. To the best of our knowledge, VSBDG is the first among its peers to keep temporal constraint attacks at bay with a simple yet effective solution.

1.4 Dissertation Organization

The rest of the dissertation is organized as follows. Chapter 2 describes the related work. Chapter 3 presents the COSA framework mainly covering the proposed system architecture, results, and discussion. Similarly, Chapters 4 and 5 present the PLP+ and VSBDG frameworks respectively. Finally, we state the conclusions and future directions in Section 6.

Chapter 2

Related Work

A typical LBS system operation involves the user's device transmitting its current true location to an LBS server, which is processed, and results are then sent back to the user. LBS systems can be categorized into two main groups based on their query architecture, i.e., snapshot and continuous [23]. In a snapshot LBS, the location transmittal to the LBS server is a singleton event with no subsequent tracking of location involved, for example, searching for points of interest (POI) such as restaurants in a location-based application. In a continuous LBS, the location information is continuously transmitted to the LBS server at application-specific intervals such that a historical location record of the client's true location is stored on the LBS server. A good example of a continuous LBS system is using online maps for driving directions where the user's true location is continuously transmitted to the LBS server for processing. The purpose of this application is to guide the user with information such as real-time traffic, weather conditions, and any re-routing in case of oncoming traffic congestions [58]. The discussions on location privacy and related solutions presented in this paper are geared mainly toward the snapshot LBS system.

In a dummy locations approach for location privacy, several dummy locations are sent along with the true location to the LBS server for query processing. The assumption that the LBS server does not know the true location but the client, i.e., the users' device knows the true location, lays the foundation for the core idea of using dummies for location privacy in a snapshot LBS scenario. In other words, preserving location privacy by sending the dummy locations along with the true location guarantees high-quality of services in an LBS snapshot query scenario since the user's true or unmodified location is used for service customization [25]. This approach also does not require a third-party anonymizer [29].

A handful of early studies spearheaded the design of dummy generation techniques. For example, Kido *et al.* [25] and Lu *et al.* [29] generated dummy locations without consideration of their similarity to true locations. Niu *et al.* [35] proposed V-circle and V-grid algorithms where final dummy locations are chosen based on their similarity in query probability to a genuine location. Niu *et al.* [34] proposed a subsequent solution that introduced an enhanced-DLS algorithm that not only generates dummy locations anchored on their similarity in location query probability but also maximizes the physical dispersion of the dummy locations to ensure good location privacy protection. Nisha *et al.* [33] devised a proxy-based approach in which dummies are identified from within a privacy area calculated using a proxy of the true location. The proxy instead of a true location along with the dummies is sent to the LBS server for processing. Legitimate locations are later extracted from the results received for the proxy location from the LBS server. Despite the additional privacy achieved by not disclosing true locations, the client-side extraction of true results may not be possible in real-world scenarios due to the resource constraints on the client devices. Tadakaluru [48] proposed an unique approach that uses land parcel features for finding dummy locations that are spatially similar to the true location.

The non-dummy approaches such as anonymization, involving methods such as spatial cloaking and k-anonymity fail to perform efficiently in edge cases with distinct properties distributed in a non-uniform fashion along the geographic distance such as densely populated urban areas where the population is high in the center and decrease as we move further from an urban area toward the rural area. The conventional non-location privacy algorithms based on k-anonymity do not consider distance as an important variable, thereby failing to perform adequately in location privacy scenarios where distance is considered a fundamental variable [29]. Location privacy solutions discussed and proposed in this paper assume that the main purpose of these algorithms is to safeguard the location of the user from external parties and not for securing the identity of the user.

The dummy locations approach to location privacy is widely studied and several techniques have been proposed to generate dummies for a given true location in an LBS scenario. The key underlying goal for using dummies is to provide anonymity to users' true location and minimize its probability of identification by an LBS server or an adversary [59]. Hence the selection approach used in generating dummies has a significant impact on the total location privacy realized by those dummies for the given true location. The research and discussions in this study are specifically targeted toward generating dummies in a snapshot LBS environment, which is the focal point of our review of current literature in this section.

The approach to using dummies for location privacy was first proposed by Kido *et al.* [25] where the dummies are generated based on the neighborhood region they belong. Here, the dummy locations are generated randomly within the regions without considering the context of the dummies and their similarity to the true location. The privacy area-aware dummies-based location privacy (PAD) by Lu *et al.* [29] proposed two area-aware circular and grid-based solutions for generating non-random dummies. The two area-aware PAD algorithms, i.e., CirDummy and GridDummy, are successful in resolving the issues posed by the earlier versions of the dummy generation algorithms that relied heavily on spatial cloaking, where random dummies are generated mainly based on the mobile user population. These do not account for scenarios where clustering of dummies could possibly compromise location privacy.

Although these PAD algorithms can generate dummies uniformly distributed within a given area independent of the mobile user distribution, they fail in situations where the geographical composition of the area is significantly homogenous. Examples include if the user location is in an area with few or no dwellings nearby or a large parcel such as a university campus where the user's location can be easily identified by simple map elimination techniques [36]. The approaches such as Lu *et al.* [29] can attain efficiencies in communication by transmitting the configuration data of the grid instead of sending the dummies within

the grid. However, they fail to address the lack of efficiency in processing these dummies on the LBS server [35].

Niu *et al.* [35] addressed the limitations of Lu *et al.* [29] in their improved versions of the PAD algorithms [29], namely V-circle-based and V-grid-based algorithms. These algorithms include an additional step of blurring the dummy locations to new positions with query probabilities that are close to that of the true location with a goal of achieving higher entropy, resulting in a more effective level of location privacy. The query probabilities are used to calculate an entropy-based metric to quantify location privacy and help with the blurring process where the location with higher entropy is considered to provide better location privacy. Anamala and Subramanian [31] proposed a similar approach using historical query probabilities of the location with the semantic classification of the location to identify dummy locations that are contextually similar to the true location.

Despite the approach based on location historical query probability values to generate dummies that are similar in the query probability context of true location, it is not practical in a real-world scenario because of the complications associated with the following: (1) obtaining trusted historical query probability data for every location within the given privacy area and (2) leveraging the data in a computationally efficient manner. One of the main assumptions of the approaches described by Niu *et al.* [35] and Anamala and Subramanian [31] is the ready availability of a given users' query probability values for both dummy and true locations. In order to generate the optimal dummy locations, these algorithms require query probability values of all possible locations in the user's location space to be stored and retrieved on demand, which can result in computing and communication overhead.

Nisha *et al.* [33] proposed a location privacy approach where a proxy of the true location is used to calculate an initial target area which is further divided into smaller zones to assist with generating dummy locations. This approach does not consider the spatial context of the areas where dummy locations are generated, thus making it highly prone to two key limitations in location privacy scenarios discussed in Table 1.1. Shi *et al.* [43] use

semantically similarity in calculating dummy locations for location privacy, making it less vulnerable to one of the two limitations, called a "map-matching attack" [36]. However, the semantic similarity does not address the possibility of most of the dummy locations lying within the same larger input parcel since all the dummy locations are semantically similar. Consequently, the approach by Shi *et al.* [43] is prone to a "location homogeneity attack" [36]. For example, the initial target area generated can be within a large parcel area of a hospital, making both proxy and dummy locations all located within the same parcel and easy to identify the true location of the user.

The approaches for generating dummy locations based on location similarities such as those reported by Nisha *et al.* [33] and Wu *et al.* [52] involve traversing through a larger area and further dividing it into several subsections. Knowing what a given area is contextually representative of is both crucial for calculating and comparing the similarities between these areas and a key step in identifying the optimal dummy locations. The current approaches rely on attribute data associated with an area for evaluating the context of a given area. The attribute data associated with a geographical area is not always readily available but also can be inaccurate since the size of the smaller areas is generally chosen randomly without any specific criteria, such as the use of 100×100 sub-regions and 1000×1000 location cells in [52].

Nisha *et al.* [33] promise greater location privacy by sending only the proxy and the dummy locations to LBS servers and not the true location for processing. The main assumption is the result for the user's true location that is embedded within a large set of results received from the LBS server for the proxy and dummy locations. In the end, it is up to the querying user's client program or a machine to extract these results associated with the true location. Wu *et al.* [52] propose a similar approach where the client is responsible for separating and filtering the results for true location from the results of dummy location. This can result in a severe computational overhead on the client machine in cases where the true and dummy locations are in a densely populated area with larger query results.

Moreover, this approach might not be practical in scenarios involving clients with limited computing and storage capabilities since they might not be able to filter and extract the results associated with the true location from the larger set of results received from LBS servers.

The common drawback of the current approaches [25, 35, 29, 31, 33] discussed above is that they do not consider the geographical features of a location in computing the dummy locations for a given true location. This lack of consideration for spatial context makes these approaches vulnerable to several types of de-anonymization attacks by an adversary, and unfeasible for implementation in real-world applications. Such dummies can be easily eliminated by an adversary by employing a "map-matching attack" [36] where the simple overlay of these dummy locations on a map would give away their dummy identity and increase the odds of identifying the true location for an adversary. Also, these approaches do not have built-in mechanisms to avoid dummies being generated on natural features such as rivers, lakes, and non-building features such as roads.

Studies such as Hara *et al.* [19], proposed a dummy generation approach in a continuous LBS scenario where spatial context involving geographical features such as roads are leveraged in omitting dummy locations that are vulnerable to de-anonymization attack by an adversary. Although Hara *et al.* [19] emphasize the importance of spatial context in generating the dummies for location privacy, it is built in the context of a continuous trajectory of user locations which is different from a discrete specific location associated with a snapshot LBS application. Apart from the natural features, the dummies can also be generated in green areas within both urban and rural areas that can be easily identified by an adversary using a "map-matching attack" [36].

The dummy generation methods for a trajectory of locations in a continuous LBS scenario involve both spatial and temporal context and are not suitable for dummy generation for a specific location in a snapshot LBS scenario with a spatial context only [59]. In other words, the application of spatial context can vary significantly between the trajectory of

user locations (continuous LBS) versus single user location (snapshot LBS). In this study, we focus on dummy generation methods for location privacy in a snapshot LBS environment where dummies are generated for a single user's true location that is located within a land parcel area.

Wu et. al. [52] and Shi et. al. [43] proposed dummy generation approaches based on semantic similarity where the dummies are generated in locations that are semantically like the true location. Through the inclusion of semantical context of the locations in evaluating locations for dummies, these approaches eliminated the possibility of placing dummies in natural features such as rivers, and mountains which would make them easy to eliminate by an adversary. The dummies based on semantic similarity are prone to "location-homogeneity attack" [36] where the true location is in a large parcel such as a hospital or university campus and the dummies generated are within the same parcel area. In this case, an adversary can easily determine the general true location of the user as the large parcel area such as hospital or university campus.

Chen and Shen [10] developed MaxMinDistDS and Simp-MaxMinDistDS that determines the dummy locations using maximum semantic diversity and physical dispersion. Zhang *et al.* [60] presented a similar approach applying maximum semantic diversity and physical dispersion as a benchmark for selecting dummy locations. Anamala and Subramanian [31] brought forth an approach to determining dummy locations using maximum semantic diversity and historical location query probabilities as criteria Shi *et al.* [43] designed a dummy generation solution using semantic similarity between locations as a principle. In this approach, a semantic location is defined as a vector of historical query probabilities in a 24-hour time period, during which the similarity between the semantic locations is calculated using the cosine similarity.

Zhang and Li [56] advocated for a dummy generation model that combines semantic diversity, location query probabilities, and physical dispersion for producing effective dummy locations. Semantic diversity helps maximize location privacy in scenarios where genuine

locations are situated in a semantically homogeneous area. However, this strategy fails to address the possibility of a temporal constraint attack in scenarios such as a residential true location. In this case, an adversary can exploit the semantic diversity in a temporal dimension [51] to eliminate dummies and to identify the residential true location.

Alyousef *et al.* [9] implemented a novel approach to creating dummy locations using deep learning. In this solution, the dummies are generated through a convolutional neural network, based on their similarity to the true location in terms of location query probability and the resulting maximum physical dispersion. The key limitation of using location query probabilities in a real-world implementation is the difficulty associated with obtaining query probability data from a trusted source for all locations within an area of interest, and the computational complexity associated with storing and processing the location query probability data. Jagarlapudi *et al.* [21] proposed a method of using a drone to assist with dummy generation. One of the major drawbacks of this approach lies in its reliance on a drone because it's impractical to have a drone linked to a user's device at all times.

The main goal of applying dummy locations for location privacy is to mask real locations from identification by an adversary, and this goal is achieved by producing dummy locations that are indistinguishable from the true locations [59]. The current dummy approaches employ a variety of factors such as semantic diversity, location query probabilities, physical dispersion [59], and spatial context [48] for evaluating locations. Except for semantic diversity, these approaches do not consider the primary purpose served by a particular location such as residential housing, commercial, or office buildings. Dummy approaches built on semantic diversity incorporate the location's primary purpose as a way to categorize locations to assist with the dummy generation process [10]. This approach, however, fails in integrating the unique temporal aspects of a semantic category (e.g., residential locations), which can be exploited by adversaries in temporal constraint attacks.

The major limitation of all the aforementioned approaches is that the spatial context is not factored in qualifying dummy locations. Dummy locations generated without accounting

for their similarity to real locations in a spatial context can be an easy target for location homogeneity and map matching attacks [36] by an adversary. Amid a location homogeneity attack, both true and dummy locations are located in the same parcel area such as a university campus or hospital area, allowing for an easy inference about the general whereabouts of the user by an adversary. In a map-matching attack, an adversary eliminates dummies that are situated in natural areas such as lakes, mountains, and green areas by overlaying the dummy and real locations on a map. A map-matching attack is also a possibility in cases where the dummy locations are located in geographically dissimilar areas compared to the true location. Here is an example: a legitimate location is a residential house and all the dummies are located on roads. Hence, the spatial context plays a vital role in determining the dummy locations that provide maximum location privacy.

Tadakaluru [48] proposed a parcel-based similarity scheme to create dummy locations that are similar in a spatial context to a true location. The study used real-world parcel data to assess and evaluate the spatial context of locations to forge dummy locations from the parcels that are spatially similar to the parcel of the input location [48]. In this study, the dummy locations are extracted from within the parcels that rank higher in similarity to the input parcel containing the true location of the user. Although Tadakaluru [48] successfully addressed the location-homogeneity attack, it's still prone to a map-matching attack. To the best of our knowledge, Tadakaluru [48] is the only study that employs spatial context for generating dummy locations to preserve location privacy. Therefore, the parcel-based dummy approach from Tadakaluru [48] is used as a benchmark to compare the effectiveness of enriched similarity search using building footprints and road proximity, spatial privacy zones, and location privacy quantification based on building footprint entropy.

Chapter 3

Context optimized and spatial aware dummy locations generation framework for location privacy

3.1 Motivation

The key limitations highlighted in Table 1.1 is further explained in the example shown below (Fig 3.1), where the assumed true location of the user is located within a large parcel area of a healthcare center. While three dummy locations are within green spaces located on the south-east side of the healthcare center, most of the dummy locations generated fall within the same parcel of the healthcare center, demonstrating the two key limitations stated in Table 1.1.

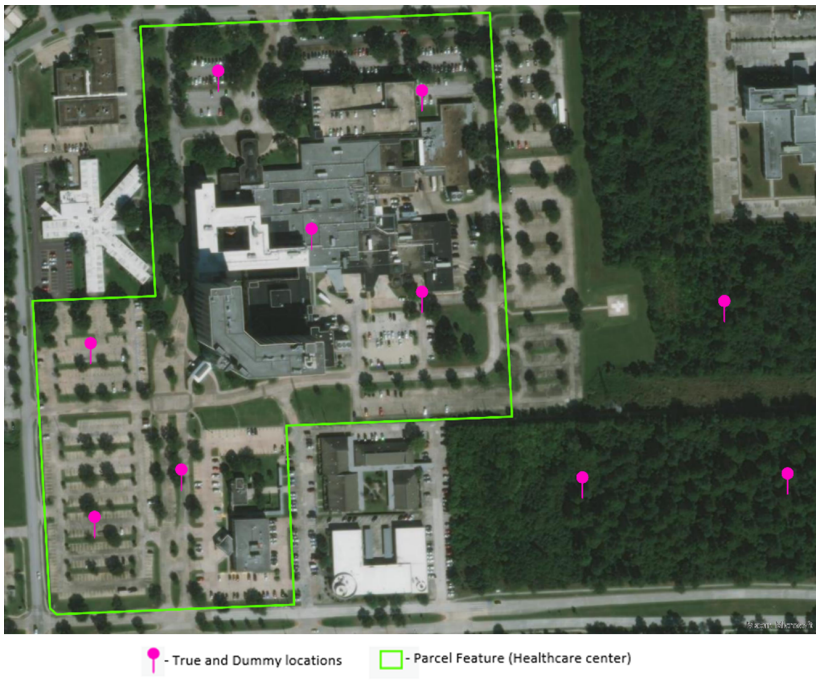

Figure 3.1: Example demonstrating the two key limitations in Table 1.1

The two main reasons that contribute to key limitations highlighted in Table 1.1 are as follows.

- **Spatial Context** The current techniques do not consider the true spatial context of the user's location. To optimize the degree of variation between the true and dummy locations, techniques should leverage geographical properties and spatial relationships using the spatial context of the current location. This would allow us to optimize the degree of variation between the true and dummy locations, so it would be difficult for an adversary to identify the true location of the user.

- **Resource Waste** Most of the current techniques do not consider the number of wasted resources, both computational and communication, that are used in performing operations on dummy locations and data transfer. A successful dummy generation technique should not only make the true location untraceable but also must achieve it by using the least number of dummy locations possible. Using less resources would also make it environmentally friendly.

Our proposed approach to address limitations with regards to spatial context and resource wastage is based on the conceptual notion of "where and what," i.e., where is the current location in terms of spatial context and what is the purpose or utility involved for sending the location information to the LBS provider in the first place. Taking the utility of the LBS service into account along with spatial context may help in reducing the number of dummies needed in preserving location privacy.The proposed dummy generation solution utilizes datasets such as parcels [2] and true color imagery [1] with spatial analyses to determine the true spatial context of the given location.

3.2 Proposed System Architecture

The main component in the proposed solution architecture is privacy geo-processor. The system workflow begins by providing the location information, i.e., latitude and longitude,

as input to the privacy geo-processor, which then geoprocesses the location information in conjunction with the other data sources such as the parcels. This step is implemented to identify and generate the dummy locations for the given true location. The main component is described in the following subsections ("Privacy Geo-processor" 3.2.1).

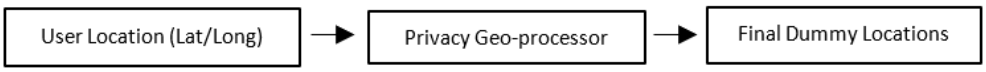

Figure 3.2: Proposed system architecture

3.2.1 Privacy Geo-processor

The privacy geo-processor is the main module in the proposed solution. The key function of the privacy geo-processor is to identify and generate dummy locations for the given user's true location. As shown in Fig. 3.3, the module consists of four major steps where the output of the module is the dummy locations that are generated for the given true location of the user.

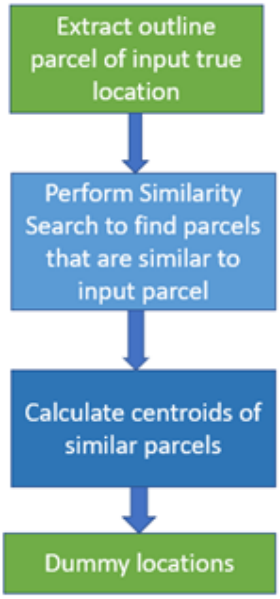

Figure 3.3: Privacy geo-processor workflow

The privacy geo-processor module takes two input items: (1) the true location coordinates of the user and (2) the parcel dataset for the county or city where the input true location is located. The feature type of the parcels in the parcel's dataset is a polygon. The first step in the privacy geo-processor involves extracting the parcel that contains the input location. A spatial join operation is performed between the input location and the parcel layer to obtain the outline parcel that contains the input location [38].

The second step involves identifying parcel polygons within the parcel dataset that are similar to the outline parcel identified in the previous step using a similarity search. The main purpose of the similarity search is to find parcel areas with spatial contexts that are similar to the input parcel. The user can add or remove the attributes of interest, such as length and area, to the similarity search, and the parcel polygons are compared for similarity with the input outline parcel using the attributes provided by the user. The sum of squared differences between standardized values of all attributes is calculated, whereby the lower the sum of squared differences, the higher the similarity between the parcels [38]. Dummy locations generated from parcel areas that are similar in a spatial context to the outline parcel containing the true location of the user are difficult to distinguish from the true location, making them less vulnerable to "map-matching attack" [36].

The final step in the privacy geo-processor is calculating dummy locations from the centroids of contextually similar parcels identified in the previous step. The centroid is calculated for each contextually similar parcel, and the final dummy location for each parcel is obtained by adding a standard offset distance in a random direction to the centroid of each parcel. The offset distance in the random direction is added to reduce the risk of identifying the true location by an adversary using spatial elimination. Spatial elimination is based on the variations in centroid locations of input parcel and contextually similar parcels, where the true location is the centroid of the input parcel, and the dummy locations are not parcel centroids or vice versa. The standard offset distance is equal to the Euclidean distance

between the centroid of the input outline parcel calculated in the first step and the true location of the user.

3.3 Preliminary Experimental Analysis and Results

3.3.1 Data Collection and Preprocessing

Out of 34 cities in Harris County in Texas, the city of Tomball was chosen to evaluate the proposed approach. This area was chosen because of its suitability to location privacy scenarios mentioned in Table 1 and the overall faster computation of a simulated study by focusing on a smaller study area. A new file geodatabase was built, and the downloaded shapefiles for Harris County parcels and cities [2] were imported as feature classes into the new file geodatabase. The 6,002 parcels within the city of Tomball were extracted from the complete parcels layer for Harris County (1.4+ million parcels) using the geoprocessing toolkit within ArcGIS Pro [38]. The extracted parcels for the city of Tomball (Fig. 3.4) were then stored in a new feature class in the file geodatabase.

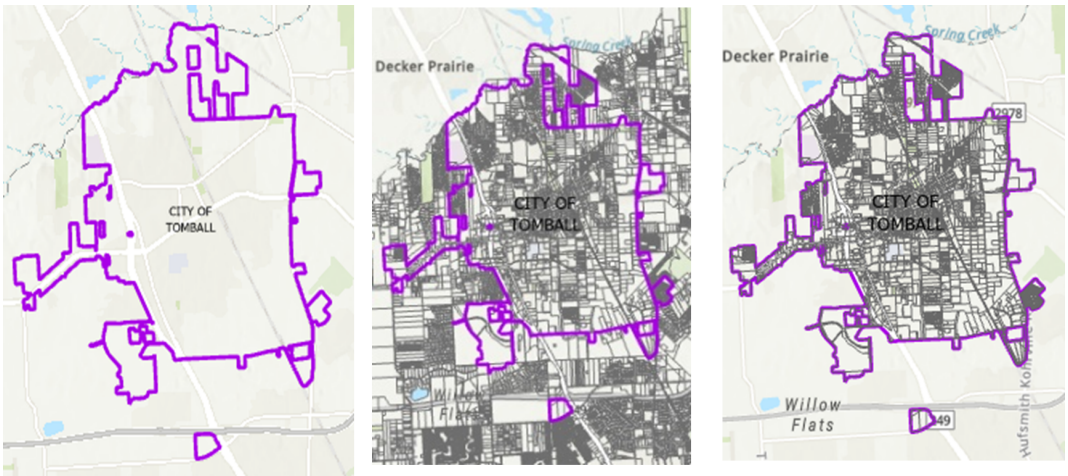

Figure 3.4: Parcel extraction for the city of Tomball in ArcGIS Pro software (ESRI 2021)

3.3.2 Privacy Geo-processor

The initial step in our proposed solution is to generate dummy locations that are similar in a spatial context to the true location of the user using the privacy geo-processor. The first step in the workflow of the privacy geo-processor is to extract the outline parcel containing the input location, i.e., user's true location. An example user's location located in the healthcare center was provided to the privacy geo-processor, and the outline parcel containing the input location was extracted, as shown in Fig.3.5

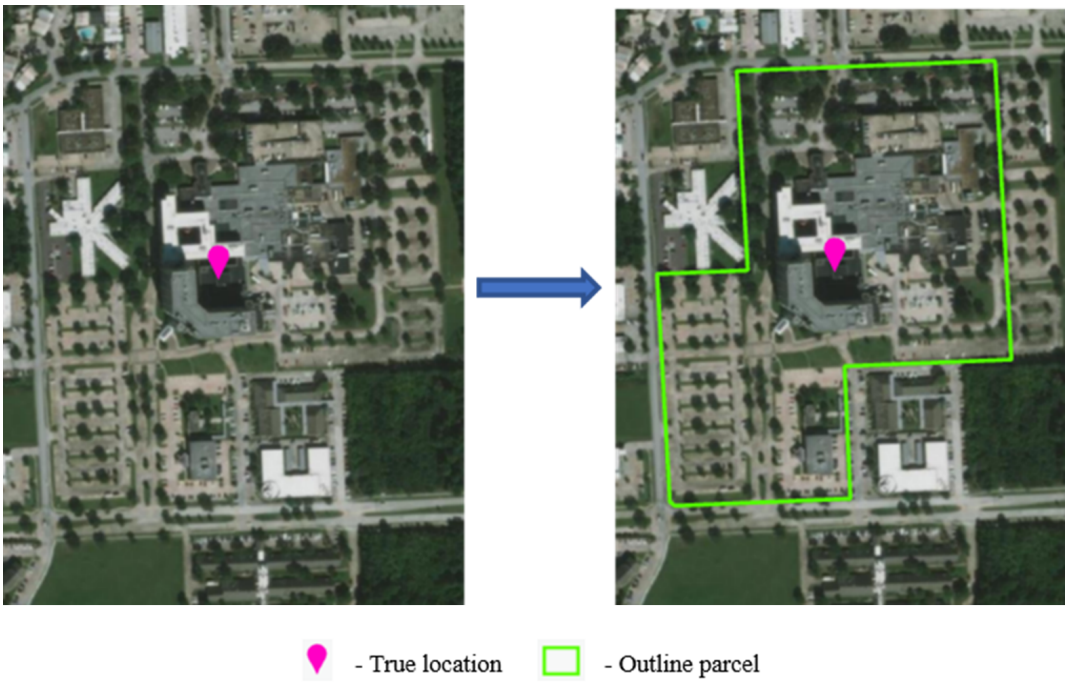

Figure 3.5: Extracting outline parcel of the user's true location

The outline parcel obtained above was used as input to the next step, where a similarity search is performed to find the parcels that are similar to the outline parcel. The input parcel provides the spatial context for the user's true location. In our current example, all the parcels within the city of Tomball were searched to find those parcels similar to the outline parcel. The similarity search was performed based on two attributes, area and length of the parcels, to compare and identify the parcels that were similar to the input parcel. The similarity search can also be extended to include other spatial and non-spatial

attributes in computing the similarity ranking between input and target parcels, such as building density derived from high-resolution airborne light detection and ranging (lidar) data [54]. Performing the similarity search based on the spatial context, i.e., input parcel of the current location, would allow us to optimize the degree of variation between the true and dummy locations, thus making it difficult to distinguish the true location from the dummy locations. The similarity rank is based on the sum of squared differences of all the standardized attribute values between two parcels, which in this case is the input outline parcel and the parcel that it is compared with. A lower value of the sum of squared differences represents a high degree of similarity between the parcels based on the attributes provided by the user. The similarity search ranks the output parcels based on similarity rank, where the most similar parcel is assigned a similarity rank equal to 1. The output parcels symbolized based on the similarity rank are shown in Figs. 3.6, 3.7 and 3.8. The parcel colored in red represents the outline parcel of the healthcare center containing the true location of the user. The parcels in different shades of blue are similar parcels identified by similarity search grouped into multiple ranges of their similarity rank. The similarity rank was calculated based on their similarity to the input outline parcel.

The next step in the privacy geo-processor is to calculate the dummy locations from the similar parcels, i.e., the output of the similarity search. For this, the centroid with an added standard offset distance in a random direction is calculated for each of these similar parcels, and these updated centroids are considered dummy locations for the given true location of the user ("Privacy Geo-processor" 3.2.1). This process was implemented on all the similar parcels obtained in the previous step for the true location (healthcare center), and the associated output centroids are shown in Fig.3.8.

3.4 Results and Discussion

When dummy locations are generated without consideration of the spatial context of the geographical areas they lie within, the dummy locations prove to be not as impactful in

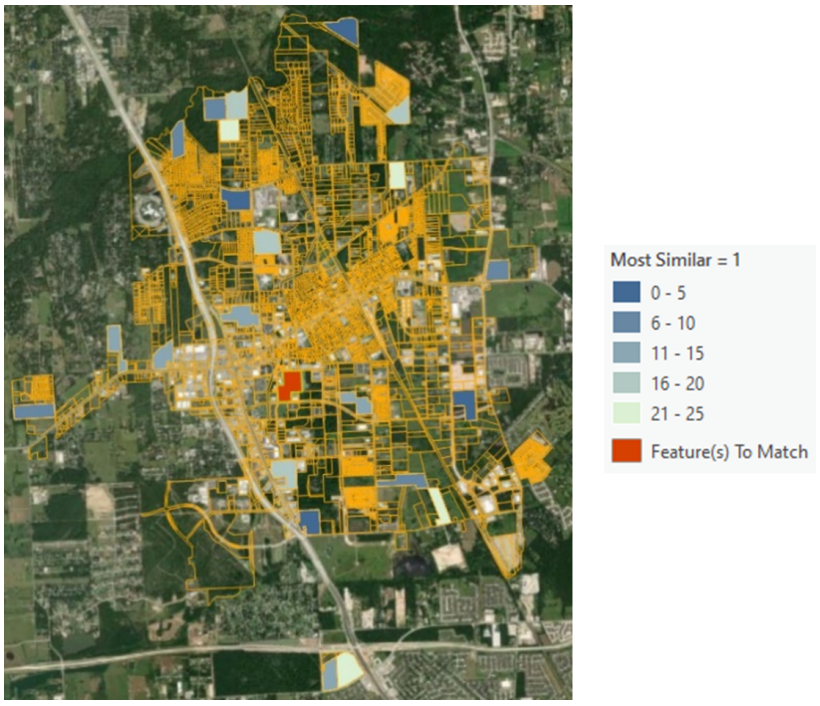

Figure 3.6: Privacy geo-processor output – showing similarity search results along with all parcels within the city

Figure 3.7: Privacy geo-processor output – showing similarity search results only

Figure 3.8: Dummy locations (centroids) generated by the privacy geo-processor: (a) centroids and their associated parcels and (b) centroids only

hiding the true location of the user [33] and fail to address the two key limitations presented in Table 1.1. Also, the location privacy approach based on the semantic similarity by Shi *et al.* [43] is vulnerable to a "location homogeneity attack" [36]. Based on our findings from the implementation of the proposed architecture, key limitations highlighted in Table 1 are addressed as follows.

Scenario 1: When the user's true location is in a large parcel area such as a hospital, university, or corporate work campus, the dummy locations generated can be within the same parcel or building compound. This makes the dummy locations vulnerable to a "location homogeneity attack" [36], where the true location can be easily identified because of the lack of diversity between true and dummy locations [36].

Our proposed solution uses the outline polygon, i.e., parcel of the given input true location as a spatial context and finds parcels that are similar over a larger spatial extent. By doing this, it eliminates the possibility of dummy location generation within the same geographic parcel extent of the user's true location, as seen in the healthcare center example.

Scenario 2: When the user's true location is a parcel or building structure in an area with few or no building structures in close proximity, the dummy locations generated are usually non-building structures and natural features such as rivers and green areas. This makes the dummy locations vulnerable to a "map-matching attack" [36], where dummies can be spatially eliminated to identify the true location.

The privacy geo-processor module in our proposed solution generates dummies based on the locations within the parcels, i.e., eliminating the possibility of dummy locations situated in natural features such as mountains, rivers, or non-building structures such as roads. In some cases, a dummy can be in a parcel without any building structures such as complete green area and rural area. Our proposed architecture provides a basic framework to embed new attributes such as building density derived from high-resolution airborne light detection and ranging (LIDAR) data [54] and integrate advanced filtering and classification methods. In general, the usage of parcels to generate dummy locations reduces the possibility of dummy location generation in non-building structures and features such as rivers and mountains.

Unlike approaches proposed by Niu *et al.* [35] and Anamala and Subramanian [31], our parcel-based approach does not require user query probability values to compute dummy locations, thus eliminating the complexities associated with both collection and usage of query probability values. The location privacy approaches such as those presented by Nisha *et al.* [33] and Wu *et al.* [52] involve the determination of location similarity, where a larger geographical area is divided into sub-areas based on arbitrary boundaries, without considering the true spatial context of the geographical areas. Our proposed solution uses widely available real-time parcel data for establishing the true boundaries of geographical areas, thus making the spatial context and location similarity calculated from these bounded areas more accurate. The location privacy approach by Nisha *et al.* [33] does not send the true location to the LBS server; instead, a proxy location along with the dummy location is sent to the LBS server. This results in an additional step for the client to filter and extract the results of the true location from the results received from the LBS server. This not only

increases computational overhead to the client but also can be impractical in cases where the client lacks the computational and storage capabilities. Our proposed solution is designed to overcome this limitation by sending the true location along with the dummy location, and it does not involve any offline filtering and extraction of results for the true location from the larger set of results on the client side.

3.5 Summary

The major contributions of this paper are summarized as follows.

- A novel dummy-based location privacy framework capable of computing location similarity based on the spatial context is proposed to generate context optimized and spatial aware dummy locations for a given true location of the user.

- The proposed location privacy framework represents the first time where a real-time parcel dataset is used to generate spatially aware dummy locations in an LBS snapshot query scenario. Usage of the real-time parcel data facilitates the accurate calculation of the spatial context for a given geographical parcel area with clearly defined boundaries, which in turn can be applied as a spatial context for any true or dummy location that lies within the parcel area. This has been a limitation in most location privacy frameworks, where the implementation is typically carried on manufactured spatial datasets with areas of interest without clearly defined boundaries. This, in turn, translates to a lack of true spatial context for the given area of interest. This limitation may render the location privacy frameworks built and tested on these manufactured datasets impractical because of their lack of applicability to real-time geographical datasets. Our proposed framework has successfully demonstrated the use of real-time geospatial data without modifications or arbitrary assumptions in the dummy location generation process for effective location privacy protection.

- The proposed framework successfully eliminates the possibility of a "location homogeneity attack" [36] where the true location and dummy locations are within the same larger parcel area, such as a hospital or university campus, which makes it easy for an adversary to identify the true location of the user. This is achieved by generating the dummy locations using the parcels outside the original scope of the input outline parcel containing the true location.

- The framework minimizes the vulnerability of the dummy locations to a "map-matching attack" [36], where the dummy locations are generated on natural features such as rivers and non-building features such as roads, which makes it easy for an adversary to identify the true location of the user by spatially eliminating the semantically different dummy locations. This is achieved by generating the dummy locations from within the parcels that are contextually similar to the outline parcel of the input true location.

- The location similarity is calculated and can be optimized based on the attributes provided by the user. The privacy geo-processor module in the proposed framework can be easily extended so that users can add and remove attributes from similarity search to optimize the location similarity and further improve the results.

Chapter 4

PLP+: An Enhanced Parcel-based Location Privacy Framework using Building Footprint Entropy for Spatially Similar Dummy Locations

This chapter presents the PLP+ framework and is organized as follows. Section 4.1 outlines the background followed by motivations in Section 4.2. Section 4.3 presents the PLP+ framework in depth. Section 4.4 walks through the quantification of location privacy using building footprint entropy. Section 4.5 showcases the experimental results followed by a discussion in Section 4.5.8.

4.1 Background

The current locations of users are often handled by LBS providers for effective customization of content and functionality provided to the users [28]. Although sharing of location information may help users with achieving a good service experience, location privacy concerns are raised because the users have limited or no control over how their location information is stored, used, and shared by the LBS providers [28]. This issue could also pose a serious threat to user privacy if locations are fallen into wrong hands, where an adversary can exploit location information by combining it with other publicly available background information and de-anonymize user identity [17].

LBS applications are broadly classified into two main groups, namely, (1) Snapshot LBSs and (2) Continuous LBSs. In snapshot LBS applications, the location information is sent to an LBS server only once to receive the intended service response from the applications. A good example of snapshot LBS applications is finding points of interest (POIs) such as restaurants, hotels, and gas stations near a given user location. In the case of continuous LBS, a user dispatches the location information to an LBS provider on a continuous basis to

retrieve the most updated service response such as driving directions and weather conditions, and it usually involves a trajectory of locations as the user continuously moves from one location to another [23].

A handful of approaches based on location obfuscation such as cloaking, differential privacy, dummy locations, and anonymization have been proposed to achieve location privacy in an LBS scenario [23]. The technique of delivering dummy locations along with genuine locations is a widely researched topic: an array of dummy and real locations are delivered to the LBS server. The main goal for sending dummies is that the LBS server or an adversary with background information would not be able to differentiate the true versus the dummy locations, thus keeping the authentic location of the user private [25, 23].

In location privacy approaches such as differential privacy, the obfuscation of the location information by adding random noise may result in the deterioration of the quality of services and may require a balancing act of achieving the best possible service experience with the loss of location privacy [23]. In the dummy locations technique, however, the LBS server processes both true and dummy locations, and returned results are filtered and extracted for genuine locations. Because of its use of unaltered true location without any noise, the dummy locations technique does not incur a loss in quality of service. By returning LBS query results that are specific to users' real locations, the dummy locations technique achieves the maximum quality of service with no reduction in location privacy, despite its wastage of resources associated with processing dummy locations [23].

In this study, we propose a novel and computationally efficient approach to generating dummy locations for true locations based on geographical context using real-world geospatial datasets. We also bring forth a novel strategy to quantify location privacy for a given location set comprised of true and associated dummy locations. This study demonstrates the significance of spatial context in successfully achieving location privacy through dummy locations and measuring the effectiveness of location privacy techniques.

4.2 Motivations

4.2.1 Motivation 1: Enhanced Parcel Similarity Search

The spatial context of a location is a crucial factor that is often overlooked when computing the similarity between the true and dummy locations. A dummy generation method that does not incorporate spatial context for identifying dummies that are spatially similar to true locations is likely to fail. A good example is when the generated dummy locations for a true location lie in non-building areas such as roads, lakes, and large green areas. A map overlay of these dummy locations over aerial imagery can give away users' true location to an adversary in a "map-matching attack" [36]. The dummy locations that mimic users' real locations in a spatial context are shown to achieve better results for location privacy. The only exception is when both true and dummy locations are situated within the same larger extent such as a hospital or university campus. This use case allows an adversary to deduce the user's general location area in a "location-homogeneity attack" [36] by exploiting the proximity between authentic and dummy locations despite their spatial similarity. By not considering both natural and artificial geographical features on the ground, most current dummy generation approaches are prone to originate dummy locations that are vulnerable to map-matching and location-homogeneity attacks [36].

4.2.2 Motivation 2: Spatial Privacy Zones

The search for dummy locations is performed within a certain geographical area surrounding true locations. This geographical area boundary - typically referred to as privacy area - can have a significant impact on both quality of the dummies generated and the suitability of the dummy approach for practical implementation. A large privacy area may impose an unnecessary computational burden rendering the approach inefficient and unsuitable for real-world implementation. A small privacy area, on the other hand, might omit perfect dummy candidates from a candidate pool and might generate low-quality dummy

locations that are vulnerable to identification by an adversary, offering poor location privacy to the true location. Hence, it is incredibly important to devise a privacy-area algorithm that produces an optimal search boundary containing the optimal dummy locations being found in a computationally efficient way. The privacy area algorithms in the current dummy location approaches are modeled based on the positional context of dummy locations relying mainly on distance and directional attributes. These methods do not factor in the spatial context of the locations, making the techniques prone to generating dummies that can be exploited in map-matching and location-homogeneity attacks [36] by adversaries.

COSA employed a parcel-similarity search for identifying spatially similar dummy locations. The target geographical area for parcel similarity search in COSA involved an entire city or county. This strategy can be highly problematic for cities with bigger spatial footprints as it involves searching through a large number of parcels. This drawback leaves COSA's current approach to search through parcels of an entire city not only a computationally intensive one but also challenging for real-world implementation.

4.2.3 Motivation 3: Location privacy quantification

While there are several solutions proposed for generating dummies to protect users' location privacy, little attention has been paid towards the quantification of location privacy. The capability to accurately measure total location privacy achieved in a location set of both true and dummy locations is crucial for evaluating the effectiveness of dummy location techniques. Importantly, gauging the quality of location privacy furnishes the comparison of a new design with any other existing approaches. The location privacy quantification also has a significant impact on the development of novel dummy approaches, especially in the areas of testing and optimization. As we have discussed previously, using spatial context is highly essential in generating dummy locations that offer maximum location privacy to the genuine location through high anonymity. To evaluate and develop such an approach, it is of paramount importance to have a location privacy quantification framework that can

effectively measure the total location privacy achieved by a given set of contextually spatially similar dummy locations for a given true location.

4.2.4 A Motivation Example

Figure 4.1 illustrates that dummy locations in Figure 4.1(b) are far better than the dummy locations in Figure 4.11(a) for the given sample input true location because the spatial similarity of dummy location parcels in Figure 4.1(b) is much higher than the dummy location parcels in Figure 4.1(a). In this example, we are able to visually compare and estimate the efficacy of the dummy locations because it is a small sample area with a limited number of parcels. Such an approach would not be possible or feasible when large datasets with millions of spatial features must be analyzed. Hence, it is crucial to be able to programmatically quantify location privacy, thereby automating the evaluation of the quality of dummy locations. By employing parcel entropy, we are positioned to gauge the collective impact that a given set of dummy locations has on the location privacy of an input location. Depending on the dummy locations and associated location privacy result, our PLP+ framework offers a strong potential to increase the SPZ area in an incremental fashion until a desired location privacy level is reached.

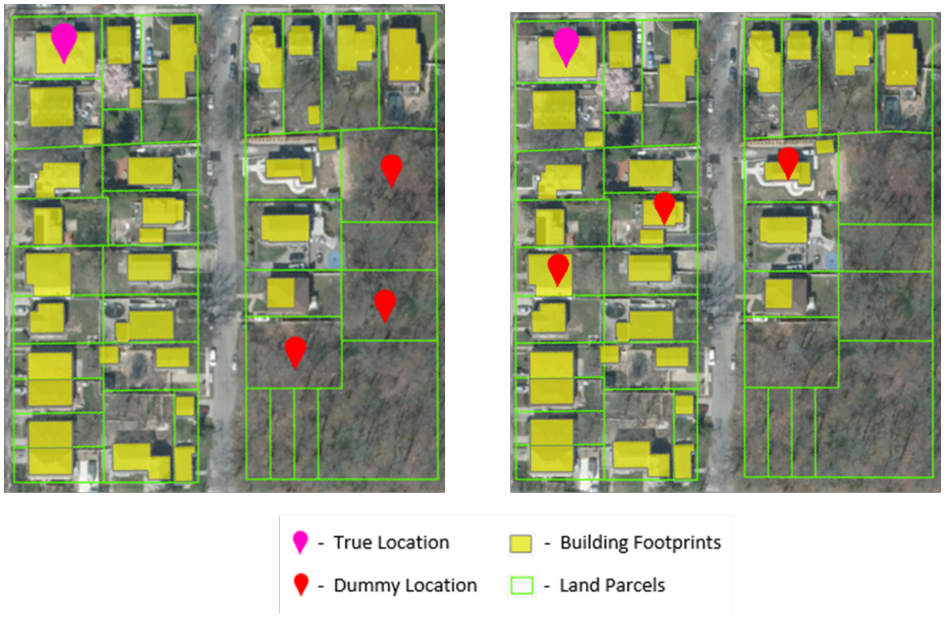

Figure 4.1: Comparing two sets of dummy locations generated for the same true location a. Dummy locations generated in parcels with green areas b. Dummy locations generated in spatially similar parcels

4.3 The Enhanced Parcel-based Location Privacy (PLP+) Framework

This section presents the PLP+ framework and is organized as follows. Sections 4.3.1 and 4.3.2 outline the foundational architecture and system design respectively. The subsequent sections4.3.3 to 4.3.6 present a step-wise walk-through of major steps within the PLP+ system presented in Section 4.3.2 .

4.3.1 PLP+ Foundational Architecture

Tadakaluru [48] proposed a parcel-based dummy generation framework for generating dummy locations anchored on parcels using a similarity search. The main module in the framework is the "privacy- geo-processor" [48] which is at the heart of the framework involving extracting an input parcel for a given user location, performing a similarity search for spatially similar parcels, and construction dummy locations from within the identified

parcels. At its core, our PLP+ framework depends on a privacy geo-processor module embracing major improvements to enhance the quality of dummy locations generated within the module and improve the overall location privacy achieved. The PLP+ framework represents enhanced functionality and capabilities of the privacy geo-processor module and any reference to the PLP+ framework for the rest of the study reported in this chapter is assumed to be applicable to the privacy geo-processor module of the parcel-based dummy generation framework proposed by Tadakaluru [48]. The details on the architecture of the parcel-based dummy generation framework can be found in Tadakaluru [48].

4.3.2 The PLP+ System Design

The main input and starting point in the functional workflow of the PLP+ framework is users' true location coordinates shown in Figure 4.2. Initially, the first step in the workflow is to extract the parcel profile outlining the input user location. The parcel profile for a given input location is comprised of both the outline parcel and the building footprint within that parcel. The user location is also provided to the SPZ module to create an optimal spatial privacy zone, which is used as a search boundary in the parcel similarity search. The user location accompanied by SPZ output are furnished as inputs to the similarity search module in the next step for searching spatially similar parcels within the SPZ. The last step involves generating dummy locations from similar parcels identified in the previous step.

4.3.3 Extract Parcel Profile

Recall that the first step in the PLP+ framework is to extract the parcel profile for the given input user location, which is of feature type point. In this part of the dissertation study, a parcel profile consists of two main components: (1) outline parcel containing the input location and (2) building footprint within the parcel outline containing the input location. In this module, the county or city-level parcels and building outline datasets of feature type polygon are provided as inputs along with the input location. The parcel profile is extracted

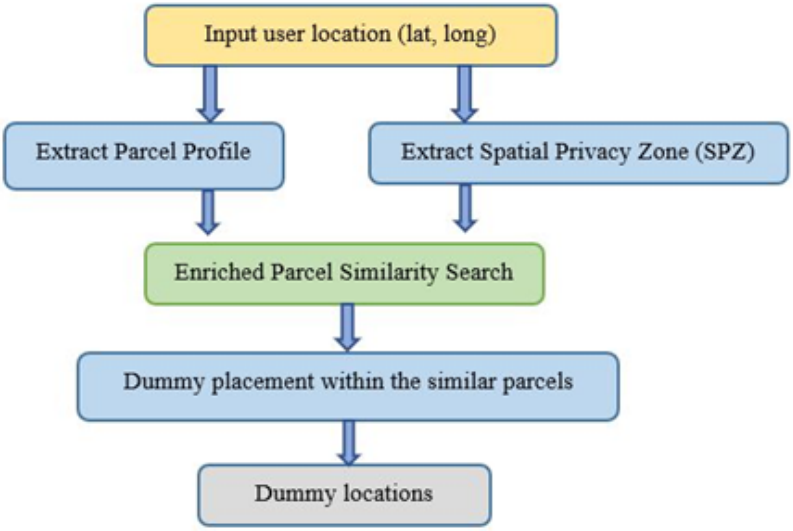

Figure 4.2: The PLP+ System Design. Adapted from Tadakaluru [48]

for the given input location by performing a spatial join between the input location point feature and parcels and building outlines datasets. The spatial join extracts the parcel and building outline polygons that contain the input location from the input datasets [38]. The outline parcel extracted in this step is applied as an input in both similarity search and dummy placement modules. The building footprint outline is leveraged in the dummy placement module for calculating an optimal position to place a dummy within a similar parcel.

4.3.4 Extract Spatial Privacy Zone (SPZ)

I proposed the parcel-based similarity search [48], which implements a full search on a complete city parcel dataset. This approach may not be computationally efficient in scenarios where a substantial number of final output similar parcels lie within a subset bounded geographic area within the city. If a similarity search can begin the search procedure within this specific geographic area first, the procedure could avoid searching through all the parcels that are outside of this geographic area, thus saving a lot of computational resources. More often than not, the approach to searching through an entire parcel dataset tends to be slow

– being unfeasible in a scenario where the parcel dataset is for a large city with millions of parcels. This scenario presents a pressing need for the identification of an optimal geographic area encompassing the input location that may have all the spatially similar parcels to serve as a primary search region for the parcel-based similarity search.

We address this need in the PLP+ framework by proposing a novel approach to leverage the clustering of building footprint information associated with each parcel within the city for calculating a set of optimal SPZs for the overall geographical region represented by the input parcel dataset. The city or county-level generation of SPZs adopting building footprint centroids is a one-time task, which is invoked as needed depending on the data update frequency of building footprints. For each new dummy generation request, the SPZ associated with the input true location is extracted and used in performing the similarity search. As shown in Figure 4.3, the process to generate SPZs involve two key steps: 1. generate density-based clusters using a DBSCAN algorithm [13] on the building outline centroid points for the complete input parcel dataset and 2. calculate bounding geometry for each generated cluster.

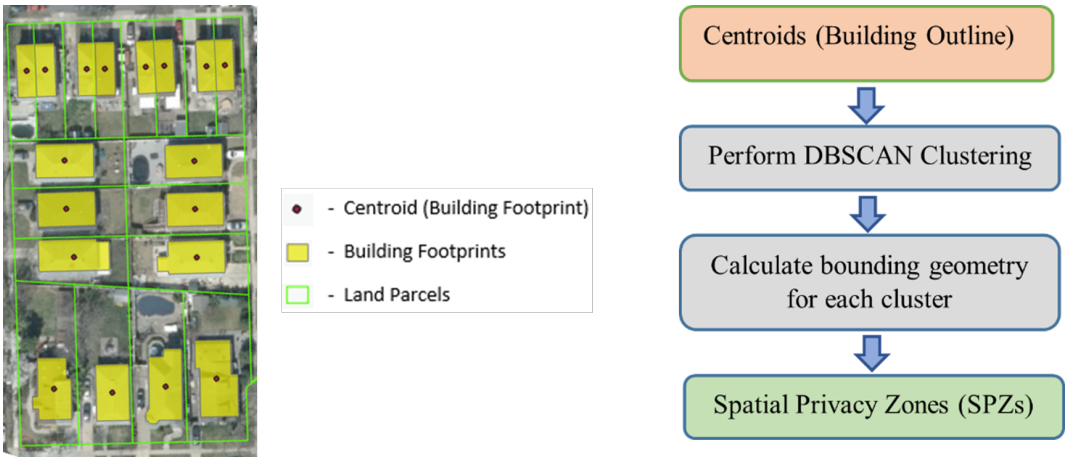

Figure 4.3: a. An example illustrating building footprint centroid along with building footprint and parcel overlaid on imagery base map [1] b. The process architecture for generating spatial privacy zone (SPZ)

The DBSCAN algorithm [13] is specifically chosen thanks to its ability to identify clusters within high-density regions and to label the points in low-density regions as noise.

Within clusters, the likelihood of locations being similarly based on their building type is high because of spatial autocorrelation [46]. Hence, the classification of high and low-density regions naturally aligns, with the key premise that the more similar a dummy is to the user's true location, the harder it is to distinguish between the dummy and true location, thus providing better location privacy [59]. The locations classified as noise by the DBSCAN can serve as another conceptual cluster on its own and can be leveraged as an independent SPZ for user locations that are located within the lower-density regions of the city. This alignment between the DBSCAN algorithm and the dummy locations approach makes the search for spatially similar dummy locations within a cluster outline polygon generated using DBSCAN an optimal choice.

The search radius eps also commonly referred to as ϵ and the minimum number of neighboring points minPts are the two main input parameters for the DBSCAN algorithm. The eps value is used for identifying the core points and having a significant impact on the clusters generated by DBSCAN. A radius that is too small separates valid clusters into even smaller clusters, and a radius that is too large can combine valid clusters into a few large clusters. In both cases, the forged clusters may not be genuine representatives of the scope of the given spatial region which can negatively impact the usage of clusters. Hence, it is essential to choose the best possible ϵ to generate an optimal number of clusters. A heuristic based on the distance to k^{th} nearest neighbor is proposed in Ester *et al.* [13] where a k-dist graph is constructed using distances for each point to its k^{th} nearest neighbor. In this approach, a k-dist graph is plotted using the sorted distances to k^{th} nearest neighbor for each point, and based on the estimated percentage of noise, a k-dist value within the elbow region is chosen as an optimal eps value. In this study, we propose a novel approach to determine the optimal eps parameter value for DBSCAN based on the original heuristic method proposed in [13] . The algorithm for determining the DBSCAN eps value is as follows

-

Algorithm 1: Estimating Optimal Search Radius (eps) for a Given Input Location Dataset.

Require: minPts, Locations dataset S=$\{l_1, l_2, l_3...,l_n\}$ with n locations {Input Datasets: Building footprint centroids }

Ensure: Optimal search radius (*eps*)

1: k = minPts -1 // minPts includes the core point itself.
2: Initialize array D to store k^{th} nearest distances
3: **for** each location l_i in dummy locations set S **do**
4: Calculate distance d_i to k^{th} nearest neighbor
5: If $d_i \neq 0$ then append di to D
6: **end for**
7: Identify and remove outliers from D
8: *eps* = Mean of all the k^{th} nearest distances within the D
9: **return** *eps*

The key difference between the heuristic algorithm proposed in the original DBSCAN paper by Ester *et al.* [13] and our approach is that we advocate for a measure of central tendency to detect and eliminate outliers from kth nearest distances calculated for all the points, as opposed to relying on an arbitrary estimation of noise percentage by the user. Another discrepancy is that we gauge the mean of all the k^{th} nearest distances minus outliers to be used as eps instead of choosing a single value from a "threshold point" within the "valley" on the k-dist graph. Finally, our eps estimator algorithm does not require sorting of locations based on their kth nearest distances as originally proposed in Ester *et al.* [13] .

SPZs are calculated for each high-density cluster by generating a convex-hull boundary encompassing all the points in a given cluster. Figure 4.4 depicts that an SPZ provides a polygonal scope [8] of a spatial region based on the input scoping criteria, which in this case is the building footprint information. The minimum enclosing convex-hull boundaries for each cluster are extracted and stored as polygon features in a geodatabase. When a user

submits a dummy generation request for a true input location, a spatial join in performed to extract the relevant SPZ containing the input location. The extracted SPZ is employed in the next step as a geographical search boundary for the enriched parcel similarity search.

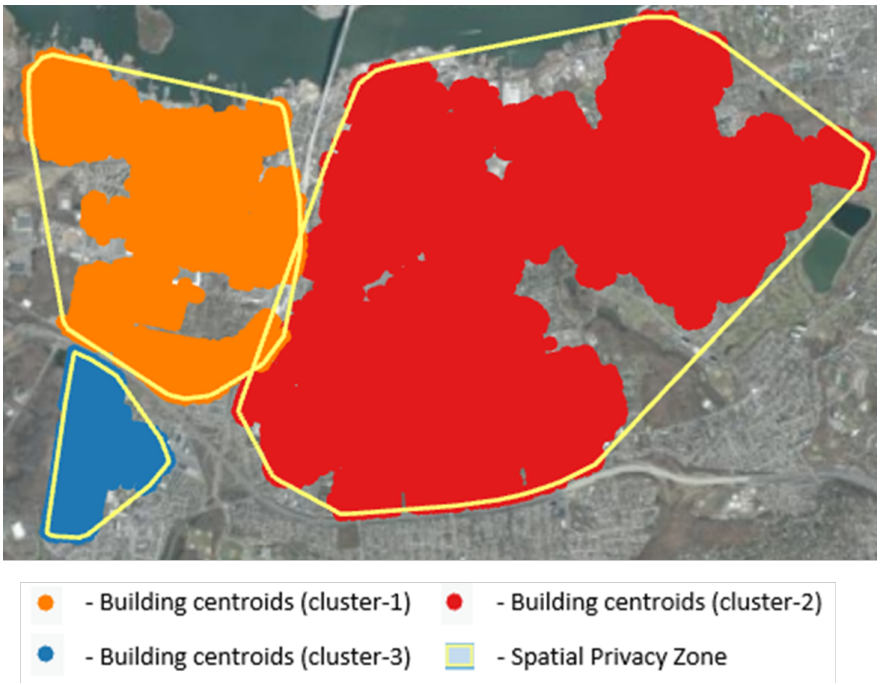

Figure 4.4: Sample illustration of three clusters generated using DBSCAN and their respective spatial privacy zones

4.3.5 Enriched Parcel Similarity Search

The parcel-based similarity search proposed in my early design (see also [48]) finds parcels that are similar to an input outline parcel based the the parcel's area and length. This approach is susceptible to returning parcels that are similar in area and length but differ significantly in terms of other geospatial attributes such as building area within the parcel and the location remoteness [12]. In other words, using dummy locations located in non-building or remote areas is not an effective and viable solution when the input legitimate location is in an urban setting. These remote dummy locations can be eliminated spatially using a map, clearing the way to isolate and identify the true location. Our proposed PLP+

framework mitigates this obstacle by leveraging two additional datasets for the similarity search, namely, (1) building footprints and (2) Street centerlines.

The building footprints dataset is incorporated in the PLP+ framework through the building volume information for each individual parcel. The attribute building volume– being computed by multiplying the building footprint area by the height of the building [37] – is applied in the similarity search to identify dummy locations from spatially similar parcels. The direct relationship between location privacy and population density is assessed in Koufogiannis and Pappas [26], where an areas with larger population density provide higher location privacy levels. The building volume, which is considered a good representative of the population distribution within a given area [37], can be associated with location privacy through their common relationship to the population density within a given region [12]. This inference of association between the building volume and the location privacy further supports our usage of building volume as an attribute of interest in searching for similar parcels. The street centerlines dataset is leveraged to assess the geographical remoteness of a location and is used as one of the factors for comparing the similarities between parcels outlining true and dummy locations. The Euclidean proximity distance to the nearest road feature is computed for each individual parcel and used as one of the attributes in the similarity search for finding spatially similar parcels.

The similarity search in the PLP+ framework relies on a Euclidean distance-based approach to measure the similarity between an input outline parcel and the other candidate parcels within the SPZ. For each pair of input and candidate parcel, the squared difference between the standardized values for each attribute is calculated and added to compute a total sum of squared differences (SSD) for all the attributes within the similarity search. The SSD values obtained for each input and candidate parcel pair are used as a sorting index, such that the candidate parcels are arranged in the ascending order of their computed SSD value with the input parcel. The candidate parcels are then ranked in the order of their sorted SSD values where the lowest SSD is assigned the highest rank, indicating that the

candidate parcel is most similar to the input parcel. The SSD measure is calculated for a given pair of parcels A and B each with n attributes using the following equation

$$SSD(A, B) = \sum_{i=1}^{n} (\mathrm{A}_i - B_i)^2. \tag{4.1}$$

4.3.6 Dummy Placement within the Similar Parcels

The last step in the PLP+ framework is the placement of dummy locations within similar parcels. The proposed approach in Tadakaluru [48] calculates the dummy location at a standard offset distance in a random direction from the parcel's centroid. The standard offset distance in Tadakaluru [48] is built on the distance between the centroid of the input outline parcel and the user's real location. This approach is inapplicable for a scenario where the true location is within a parcel's building footprint and the dummies generated are placed in a non-building area within the parcel, even though both genuine and dummy locations are placed using the same offset distance from the parcel centroid.

In the PLP+ framework, a dummy placement within similar parcels is performed based on the spatial context of a given true location within an input outline parcel. The spatial context of the true location is determined based on the associated building footprint within the input outline parcel. If the real location lies within the building footprint area of the input parcel, the dummies are also placed randomly within the building footprint areas within the parcels. If the legitimate location lies outside of the building footprint area within a parcel in a non-building area such as a parking lot, the dummies also ought to be located randomly in non-building areas within the parcels as illustrated in Figure 4.5. This dummy placement approach guarantees an otpimized location privacy by positioning the dummies in positions that are contextually similar to users' genuine locations where all locations are either in building areas or non-building areas.

Figure 4.5: Example illustrating placement of dummy locations a. when the true location lies in the building area of the parcel b. when the true location lies in the non-building area of the parcel

4.4 Quantifying Location Privacy Using Building Footprint Entropy

Location privacy quantification is a key step in determining the effectiveness of location privacy approaches. In the PLP+ framework, location privacy is achieved by sending spatially similar dummy locations along with true locations of a user. These dummy locations are chosen from geographical areas that are similar in a spatial context to the user's real location such that the land parcel polygon enclosing a location is used in determining the spatial context. In an urban setting, the building footprint within a land parcel area is a crucial factor that has a major influence on the spatial context of the parcel. In this part of the dissertation study, we propose a novel technique that uses Building Footprint Entropy – FPE – to quantify the location privacy attained by a set of dummy locations for a given user's true location. When being delivered with a legitimate location, dummy locations provide location privacy by introducing uncertainty in predicting the true location from the complete location set. Thus, we assess the effectiveness of a given set of dummy locations by measuring the uncertainty introduced by these dummy locations in predicting the true location. The uncertainty associated with a given set of locations S=s_1, s_2, s_3...,s_n can be measured using entropy [34, 32, 42]. More formally, the uncertainty is computed as follows:

$$H(S) = -\sum_{i=1}^{n} \mathrm{p}(s_i) \log_2 p(s_i), \tag{4.2}$$

where $p(s_i)$ is the probability of selecting location s_i from a given set of locations S and $\sum_{i=1}^{n} \mathrm{p}(s_i) = 1$.

The probability of selecting location s_i from parcel $parcel_i$ is computed using the building footprint area of $parcel_i$ as shown in the expression below.

$$p(s_i) = \frac{\text{Building footprint area}(parcel_i)}{\sum_{j=1}^{n} \text{Building footprint area}(parcel_j).} \tag{4.3}$$

Building Footprint Entropy (FPE) captures the uncertainty associated with the given location set S, which encompasses both real and dummy locations. A large entropy value represents a high level of uncertainty in determining the location's identity – whether it is a true or a dummy location. So, when two sets of dummy locations are compared against each other, the location set with higher entropy value is considered to provide better location privacy. The entropy value is maximum when all the locations, both dummy and true in the location set, have equal probabilities of identification [34]. It can be inferred that a higher entropy value for a given set of dummy and true locations represents greater similarity between the locations translating to better location privacy [59].

Algorithm 2: Calculating Building Footprint Entropy (FPE) for a Given Set of Dummy Locations.

Require: Dummy locations set S=$\{s_1, s_2, s_3...,s_n\}$ with n locations
{Input Datasets: Parcels, Building footprints}

Ensure: Building Footprint Entropy (FPE)

1: TotalBldgFtprArea = 0
2: **for** each location s_i in dummy locations set S **do**
3: find $parcel_i$ outlining the location s_i
4: find building footprint area $BldgFtprArea_i$ of $parcel_i$
5: TotalBldgFtprArea += BldgFtprArea
6: **end for**
7: **for** each location s_i in dummy locations set S **do**
8: $p(s_i)$ = BldgFtprArea$_i$/$TotalBldgFtprArea$
9: FPE += -$p(s_i)$ * log($p(s_i)$)
10: **end for**
11: **return** FPE

The FPE value is calculated for an input location set using the parcel and building footprints dataset as shown in Algorithm 2. The total building footprint area $TotalBldgFtprArea$ is computed for a given input location set using the building footprint and parcel associated with each location. The probability of electing a location p(s_i) for each location s_i is computed and used in calculating the total FPE of the input location set. In this study, we advocate for FPE to gauge the effectiveness of the PLP+ framework in preserving users' location privacy when PLP+ is compared to the parcel-based approach proposed by Tadakaluru [48]. To this end, we calculate the FPE for dummy location sets generated by the competitors such as the PLP+ framework and the parcel-based approach [48].

4.5 Experimental Results

In this section, we present the experimental results which are organized as follows. Section 4.5.1 provides details on the datasets and pre-processing steps. Sections 4.5.2 and 4.5.3 demonstrates the implementation of the initial steps of the PLP+ architecture. Section 4.5.4 demonstrates the enhanced parcel similarity search followed by the evaluation of the effectiveness of SPZ in Section 4.5.5. Finally, Section 4.5.7 shows the evaluation of the location privacy quantification approach proposed in this chapter.

4.5.1 Data and Preprocessing

The three datasets used in the evaluation study are building footprints, street centerlines, and land parcels, for Richmond County (Staten Island) in the state of New York [4, 6, 5]. The parcels in the parcel dataset are polygon features representing physical land ownership boundaries based on city tax and land ownership records. The building footprints dataset contains polygon features representing the building outline along the perimeter as seen from the top [4]. The street centerlines are comprised of line features representing roadways of type street in a road-bed format [6]. These datasets were extracted from New York's statewide datasets using the geoprocessing toolkit available in ArcGIS Pro [38] as depicted in Figure 4.6. The number of features in parcels, building footprints, and street centerlines datasets extracted for Richmond County (Staten Island) were 123,849 , 141,946, and 14,722 respectively [3].

4.5.2 Extract Parcel Profile

The first step in the PLP+ framework workflow involves extracting the parcel profile for a given input location. The parcel profile consists of both the parcel outlining the input location and the building footprint associated with the outline parcel. The main purpose of this step is to extract the spatial context associated with the user's genuine location from the parcel profile containing the parcel outline and building footprint. The parcel profile is

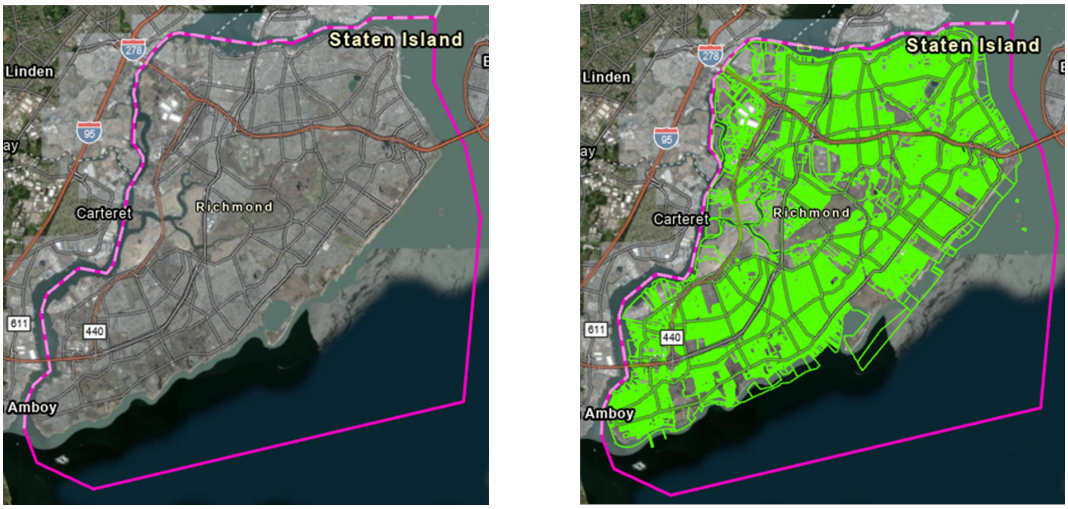

Figure 4.6: a. Boundary for Richmond County (Staten Island), New York b. Parcels within the county boundary.

extracted by implementing a spatial join between the input location, parcels, and building footprint datasets. Figure 4.7 illustrates the input location, and extracted profile parcel profile view at both larger and smaller scales.

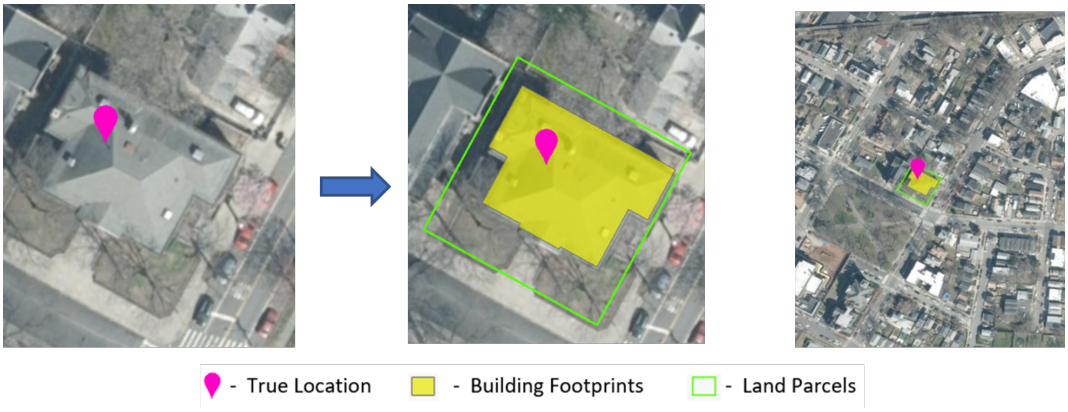

Figure 4.7: a. Input user location b. Extracted parcel profile for the given input location c. Input location and associated parcel profile at a smaller scale.

4.5.3 Extract Spatial Privacy Zone (SPZ)

The similarity search in the PLP+ framework begins the search for similar parcels from within the SPZ of the input location. As explained in Section 4.3.4, the DBSCAN algorithm is deployed to generate an optimal number of clusters, which are then used in calculating SPZs for Richmond County (Staten Island) using building footprint [17] centroids as the

input dataset. Given the large volume of the input dataset, a larger minPts parameter value is suggested to work better by Schubert *et al.* [41]. The minPts parameter was set as 1% of the total number of building footprint centroids (minPts = 1,419). For the given input building footprint centroid dataset, we estimate the optimal eps of 1,814 feet using our proposed eps estimator stipulated in Algorithm 2. For this experiment, we leverage the box and whisker plot [20] of distances to k^{th} nearest neighbor (K=1,418) for all points (Figure 4.8a) to detect outliers and calculate the optimal eps of 1,814 feet. The optimal eps of 1,814 feet was in the lower end of the "valley" or "elbow" region as shown in k-dist graph in Figure 4.8b, further verifying the validity of the generated optimal eps using Algorithm 2. We construct clusters and enclosing SPZs using the DBSCAN algorithm governed by the input parameters minPts=1,419 and eps = 1,814 feet.

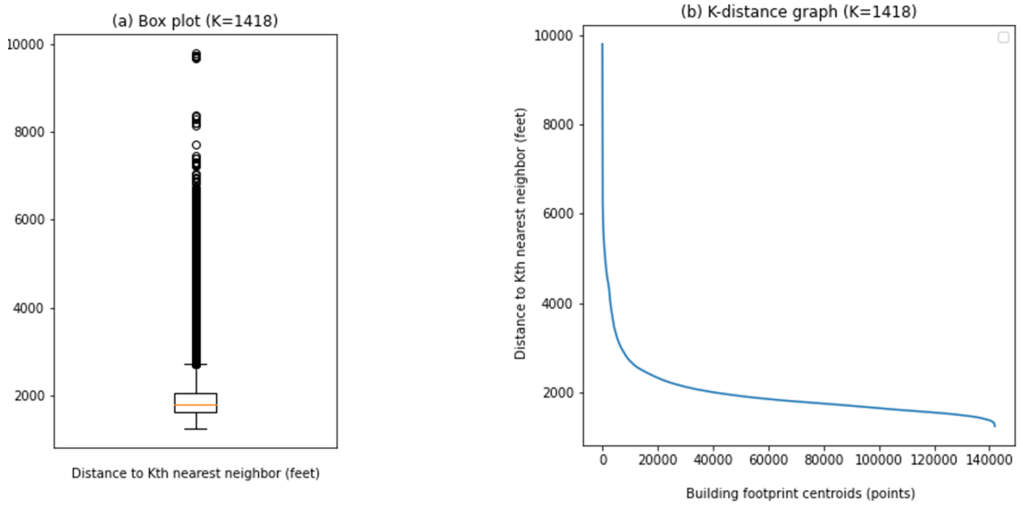

Figure 4.8: a. Box plot showing the distance to k^{th} nearest neighbor (K=1418) for all the input centroid locations b. K-distance graph showing sorted distances for all the input centroid locations to k^{th} nearest neighbor (K=1418)

The SPZs, which are bounding convex hull geometries generated for each DBSCAN cluster as shown in Figure 4.9a, are stored in the geodatabase as polygon features [38]. The SPZ enclosing the input user's true location is extracted using a spatial join between the input location and layer containing SPZs as shown in Figure 4.9b. Next, the extracted SPZ

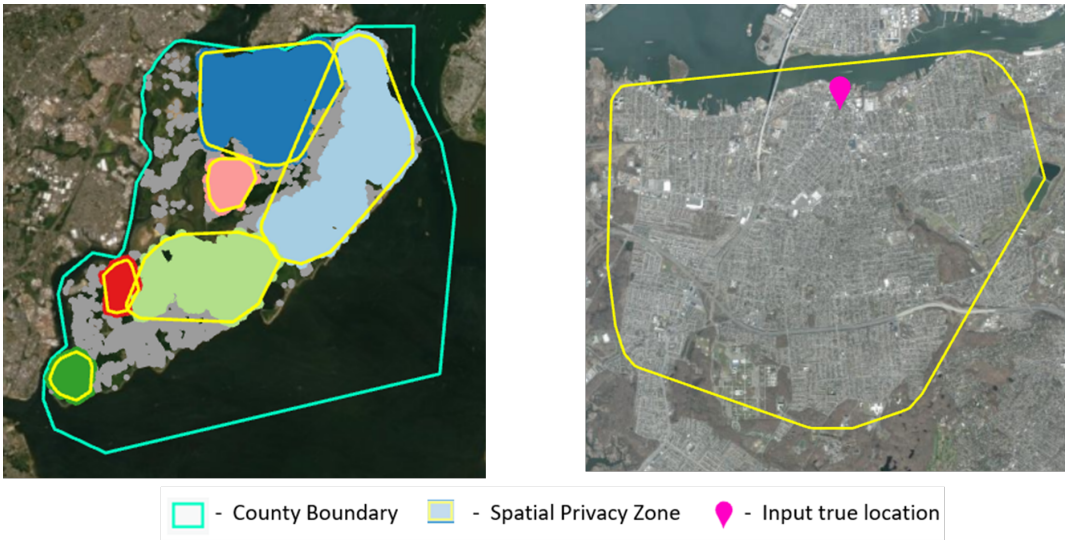

Figure 4.9: a. Clusters and enclosing SPZs using DBSCAN algorithm for parameters minPts=1419 and eps = 1814 feet b. SPZ extracted for input user location.

polygon is applied as a search query boundary to limit the candidate parcels used for the similarity search in the next step.

4.5.4 Enriched Parcel Similarity Search

In this step, the candidate parcels within the SPZ boundary identified in Section 4.5.3 are searched to identify parcels that are similar to the input outline parcel (see also Section 4.5.1). The parcel similarity search is enriched with two new additional attributes: 1. building footprint area within the parcel and 2. proximity to the closest road along with the area and length of the parcel that was used previously in the baseline approach [48]. The layers containing the input parcel, candidate parcels within the SPZ, and the number of output similar parcels to generate (N), are provided as input to the similarity search. The search process computes the SSD for the four attribute values between each input and candidate parcel combination for all the candidate parcels within the SPZ and identifies the top N similar parcels within the SPZ by sorting the parcels from lowest to highest SSD.

Figure 4.10 unveils the results from the similarity search with 15 similar parcels (N =15) found within an SPZ containing 31,412 candidate parcels. The 15 similar parcels are ranked based on their SSD ordered from the lowest to the highest measures (Table 4.1). The

SPZ-based similarity search in the PLP+ framework is computationally efficient because it only employs parcels within the SPZ boundary instead of searching through all the parcels within the county. Specifically, for the given input outline parcel, the similarity searches through only 31,412 candidate parcels within the SPZ (Figure 4.10) as opposed to searching through the entire 123,848 candidate parcels in Richmond County (Staten Island). As discussed in Section 4.3.4, these SPZs are created from DBSCAN clustering using our novel eps estimator orchestrated by Algorithm 2. To consider our approach of using SPZ for a query boundary as efficient, it must be able to identify N highly similar parcels from within the SPZ boundary without any loss in similarity compared to N similar parcels that can be identified from the entire city parcel dataset. In other words, for a given input parcel, if we can prove that N similar output parcels from a similarity search based on SPZ and N similar output parcels from a similarity search based on the entire city parcel dataset are not statistically significantly different, we conclude the former approach is more efficient by utilizing a subset of an entire parcel dataset.

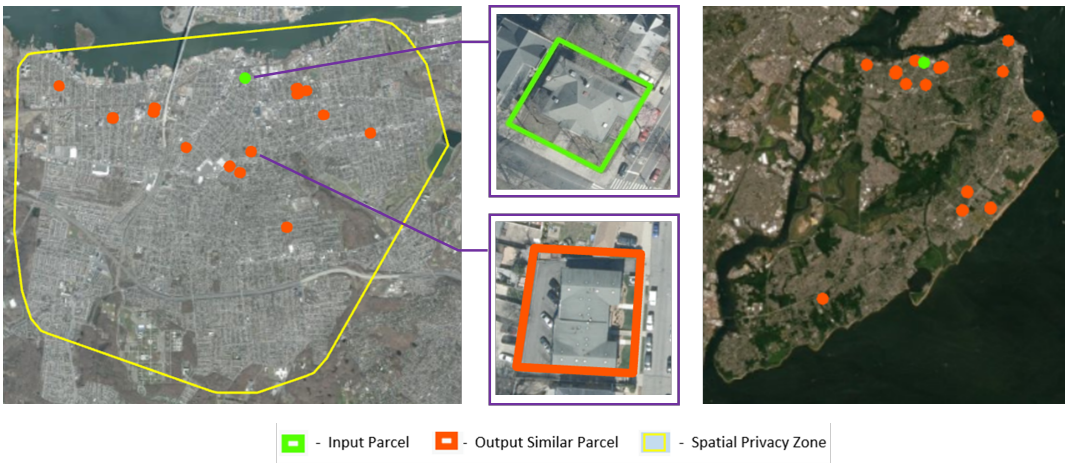

Figure 4.10: Input parcel and output parcels from similarity search for 15 similar parcels (N=15) within a. SPZ with 31,412 candidate parcels b. entire Richmond County (Staten Island) with 123,848 candidate parcels.

Output Similar Parcel ID	Sum Of Squared Differences (SSD)	Similarity Rank
001	0.001618	1
002	0.001717	2
003	0.004959	3
004	0.006934	4
005	0.00817	5
006	0.009133	6
007	0.013153	7
008	0.01427	8
009	0.014392	9
010	0.016487	10
011	0.016789	11
012	0.019067	12
013	0.019201	13
014	0.024572	14
015	0.024794	15

$$\overline{SSD} = 0.013017 \tag{4.4}$$

$$\sum SSD = 0.195256 \tag{4.5}$$

Table 4.1: Showing the SSD values and similarity rank for output parcels from similarity search with in a SPZ for N=15

4.5.5 Evaluating SPZ

Table 4.1 shows that the PLP+ similarity search computes the sum of squared differences (SSD) of the attributes between the input parcel and each of the candidate parcels to determine and rank the parcel similarity. The mean of SSDs (MSSD) measures the quality of similarity achieved for a set of N similar parcels and in turn, the measures can be used as a metric to compare multiple approaches to similarity search. To evaluate the effectiveness of using SPZ and the underlying DBSCAN eps estimator algorithm in the context of similarity search, we evaluate and compare the MSSD between two sets of output similar parcels, where one is generated using SPZ and another is not. The MSSD for a set of N similar parcels identified by similarity search can be computed as follows.

$$\overline{SSD} = \frac{\sum_{i=1}^{N} SSD_i}{N}. \tag{4.6}$$

For this experiment, we randomly select 10 input locations (k=10) in Richmond County (Staten Island) with five locations in SPZ-1 and another five in SPZ-2 (Figures 4.11a and 4.11b respectively). When it comes to each of 10 input locations, we extract input outline parcel as explained in section 4.5.2, performed similarity search using geoprocessing toolkit available in ArcGIS Pro [5, 38] and identified 50 similar parcels (N=50) from the candidate parcels within their respective SPZs. Given each of the same 10 locations, we also conducted another similarity search and identified 50 similar parcels (N=50) using candidate parcels from entire parcel dataset of Richmond County (Staten Island).

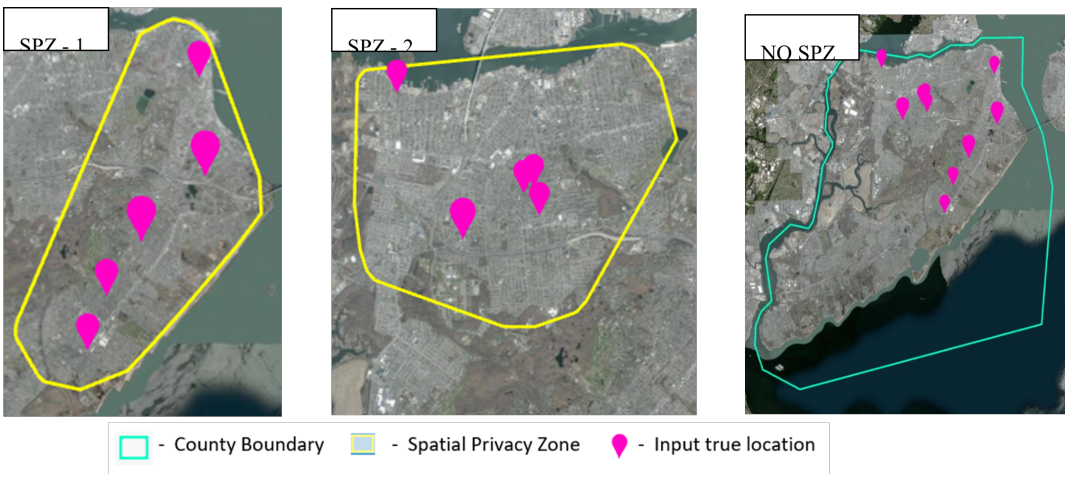

Figure 4.11: a. Showing 5 out of 10 input locations for similarity search using SPZ-1 with 43,563 candidate parcels b. Showing 5 out of 10 input location in SPZ-2 with 31,412 candidate parcels c. 10 input location in the context of entire Richmond County (Staten Island) with 123,848 candidate parcels without any SPZ.

We calculate, for each input location, $MSSD_{SPZ}$ – the MSSD of 50 similar parcels identified within SPZ, and the $MSSD_{NO-SPZ}$ – the MSSD of 50 similar parcels identified from the entire parcel dataset of Richmond County. The $\Delta MSSD$ for each input location is gauged by taking the difference between Mean SSD_{SPZ} and Mean SSD_{NO-SPZ}. Table 4.2 shows the 10 input locations and their calculated $MSSD_{SPZ}$, $MSSD_{NO-SPZ}$, and $\Delta MSSD$ values, where the input locations are grouped by the SPZ they belong to. For each SPZ group of observations, the paired sets of observations for $MSSD_{SPZ}$ and $MSSD_{NO-SPZ}$ are statistically evaluated for significant discrepancies using Paired Samples t-Test and to determine whether there is a loss in similarity by using a smaller candidate parcel set for similarity search instead of larger complete parcel dataset. We apply the R-studio [39] software to perform the statistical analysis presented in this part of the study.

SPZ_1 Observations				SPZ_2 Observations			
Input Location	**$MSSD_{SPZ}$**	**$MSSD_{NO\text{-}SPZ}$**	**$\Delta MSSD_1$**	**Input Location**	**$MSSD_{SPZ}$**	**$MSSD_{NO\text{-}SPZ}$**	**$\Delta MSSD_2$**
1	0.03320342	0.00911344	0.02409	6	0.00340062	0.00041158	0.00298904
2	0.12559	0.04378404	0.081806	7	0.000189	0.00002918	0.00015982
3	0.00053154	0.00013132	0.0004	8	0.00197416	0.00039854	0.00157562
4	0.00123502	0.00027496	0.00096	9	0.000738	0.00014356	0.00059444
5	0.0337978	0.0109364	0.022861	10	0.00056198	0.00009192	0.00047006

Table 4.2: Showing the input locations within each SPZ and the calculated Mean SSD_{SPZ}, Mean SSD_{NO-SPZ} and $\Delta MSSD$ values calculated based on SSD values of 50 similar parcels (N=50) for each location input location.

A paired t-test is adopted to compare two sets of paired observations and check if there is a statistically significant difference between the means of the two sets of observations. In this experiment, we are performing paired t-test to compare the two sets of observations 1. $MSSD_{SPZ}$ 2. $MSSD_{NO-SPZ}$ for a group of locations within the same SPZ and check if there is any statistical evidence to show that are significantly different. Proving there is no statistically significant difference between $MSSD_{SPZ}$ and $MSSD_{NO-SPZ}$, we assume that the similar parcels identified in an SPZ are as good as the similar parcels pinpointed by searching the entire dataset.

One of the core requirements for performing a paired t-test on sets of paired observations for $MSSD_{SPZ}$ and $MSSD_{NO-SPZ}$ is that the $\Delta MSSD$ must be normally distributed. The Shapiro-Wilk's normality test is conducted for values in $\Delta MSSD_1$ and $\Delta MSSD_2$, which resulted in p-values of 0.09372 and 0.2568, respectively. Since the p-value > 0.05, we fail to reject the null hypothesis and concluded that both $\Delta MSSD_1$ and $\Delta MSSD_2$ are normally distributed. A separate two-tailed paired t-test is performed on each group of observations belonging to SPZ1 and SPZ2, with results and null hypothesis tabulated in Table 4.3.

Hypotheses : H_0 (null hypothesis) : $\mu_1 - \mu_2 = 0$ H_A (alternative hypothesis) : $\mu_1 - \mu_2 \neq 0$ μ_1 - Population mean of $MSSD_{SPZ}$ μ_2 - Population mean of $MSSD_{NO\text{-}SPZ}$	
SPZ_1 Observations	**SPZ_2 Observations**
test statistic (t) = 1.7525 degrees of freedom (df) = 4 p-value = 0.1546 Result: Since p-value > 0.05, we fail to reject the null hypothesis that the population mean difference between $MSSD_{SPZ}$ and $MSSD_{NO\text{-}SPZ}$. Hence, we assume that there is no statistically significant difference between the observations for $MSSD_{SPZ}$ and $MSSD_{NO\text{-}SPZ}$	test statistic (t) = 2.2461 degrees of freedom (df) = 4 p-value = 0.08803 Result: Since p-value > 0.05, we fail to reject the null hypothesis that the population mean difference between $MSSD_{SPZ}$ and $MSSD_{NO\text{-}SPZ}$. Hence, we assume that there is no statistically significant difference between the observations for $MSSD_{SPZ}$ and $MSSD_{NO\text{-}SPZ}$

Table 4.3: The hypotheses for paired t-test and separate paired t-test results for both SPZ1 observations and SPZ2 observations.

4.5.6 Generating Dummy Locations from Similar Parcels

The last step in the PLP+ framework is to produce dummy locations from within the similar parcels identified using a similarity search in the previous step. A dummy location is placed within each similar parcel using our proposed approach for dummy placement articulated in Section 4.3.6. Figure 4.12 sketches 15 dummy locations (N=15) generated for the input location indicated in Section 4.5.2, and the figure highlights the dummy location

placement in the inset for a couple of dummy locations based on approach using building footprints discussed in Section 4.3.6. These dummy locations are constructed using enriched similarity search implemented in the previous section using two additional attributes: 1. building footprint area within the parcel, and 2. proximity to the closest road, with the ultimate goal of improving total location privacy achieved. To measure and evaluate total location privacy achieved, we implement an entropy-based quantification approach for location privacy as explained in Section 4.5.7.

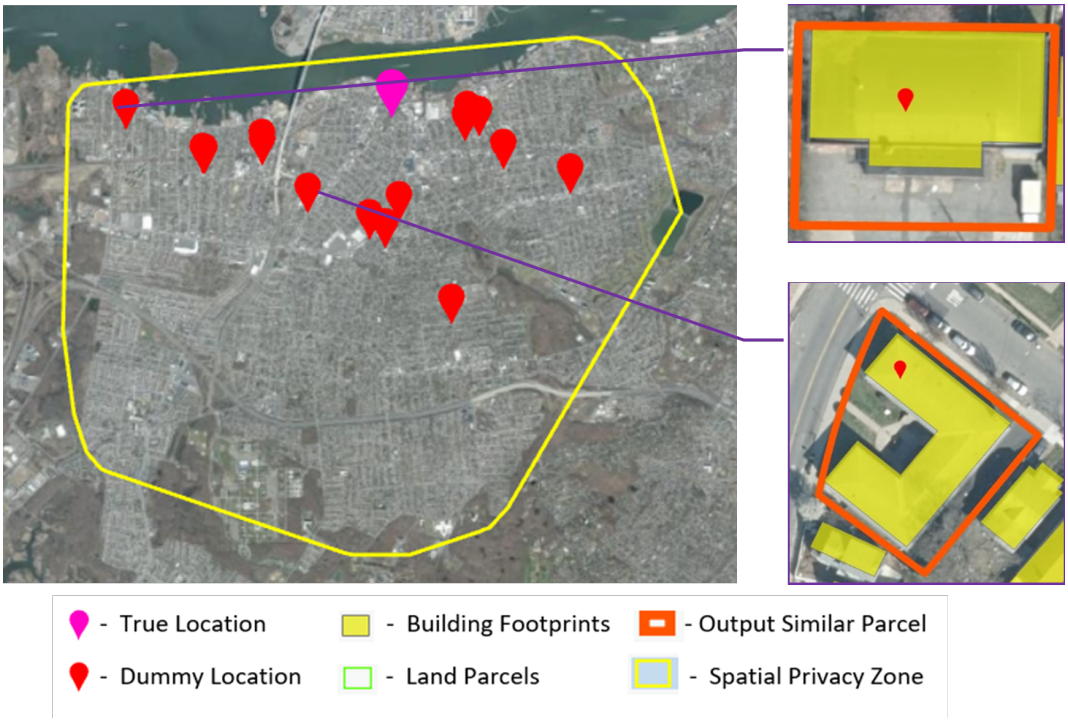

Figure 4.12: Dummy locations generated for input location using PLP+ enriched similarity search for 15 similar parcels (N=15) within a SPZ with 31,412 candidate parcels.

4.5.7 Location Privacy Quantification

The cardinal purpose of the proposed PLP+ framework (Sections 4.3.2) is to create dummy locations for a given input true location. The end goal of these dummy locations is to preserve the location privacy of a user by being sent along with the input real location to the LBS server for processing. The maximum location privacy is achieved when the dummy locations are truly indistinguishable from the input true location. As explained in

Section 4.4, the indistinguishability of the dummy locations from the true location in an urban setting can be measured using building footprint entropy (FPE), which is the degree of uncertainty involved in picking a location from a set of true and dummy locations. A larger FPE value denotes a higher degree of uncertainty associated with choosing a location from the given set of input and dummy locations, thus indicating greater location privacy.

We implement our proposed Algorithm 2 on 10 input locations randomly chosen in Richmond County (Staten Island). Three different FPE values for each input location are derived from three sets of 50 dummy locations (N=50) generated using three different similarity search criteria (Figure 4.13). The three similarity search criteria used include: 1. Search entire dataset for 50 dummy locations using two parcel attributes, area, and length (SPZ = NO, ATTR=2, N=50) 2. Search SPZ for 50 dummy locations using two parcel attributes, area, and length (SPZ = YES, ATTR=2, N=50) and 3. Search SPZ for 50 dummy locations using four parcel attributes- area, length, building footprint area and proximity to closest road (SPZ = YES, ATTR=4, N=50). The FPE value for each input location is computed based on 50 dummy locations generated for that input location using one of the three similarity search criteria. The three sets FPE values for 10 sample input locations are plotted using multi-line chart as illustrated in Figure 4.13, with input location number on the X-axis and FPE values on the Y-axis.

In Section 4.5.5, we evaluate our approach of generating an SPZ from DBSCAN and eps estimator algorithm (Algorithm 2) using MSSD of an input location. There is no statistically significant difference between $MSSD_{SPZ}$ and $MSSD_{NO-SPZ}$ and; therefore, we demonstrate that there is no loss in quality of similarity in output parcels generated by similarity search using SPZ when compared to the output parcels generated using the complete dataset. The same effect can be observed in two sets of FPE values; 1. SPZ = YES, ATTR=2, N=50 2. SPZ = NO, ATTR=2, N=50 (Figure 4.13). The only distinction between these two sets of FPE values is the usage of a SPZ versus entire parcel dataset in the underlying parcel similarity search. Figure 4.12 unravels that there is minor difference between the FPE values

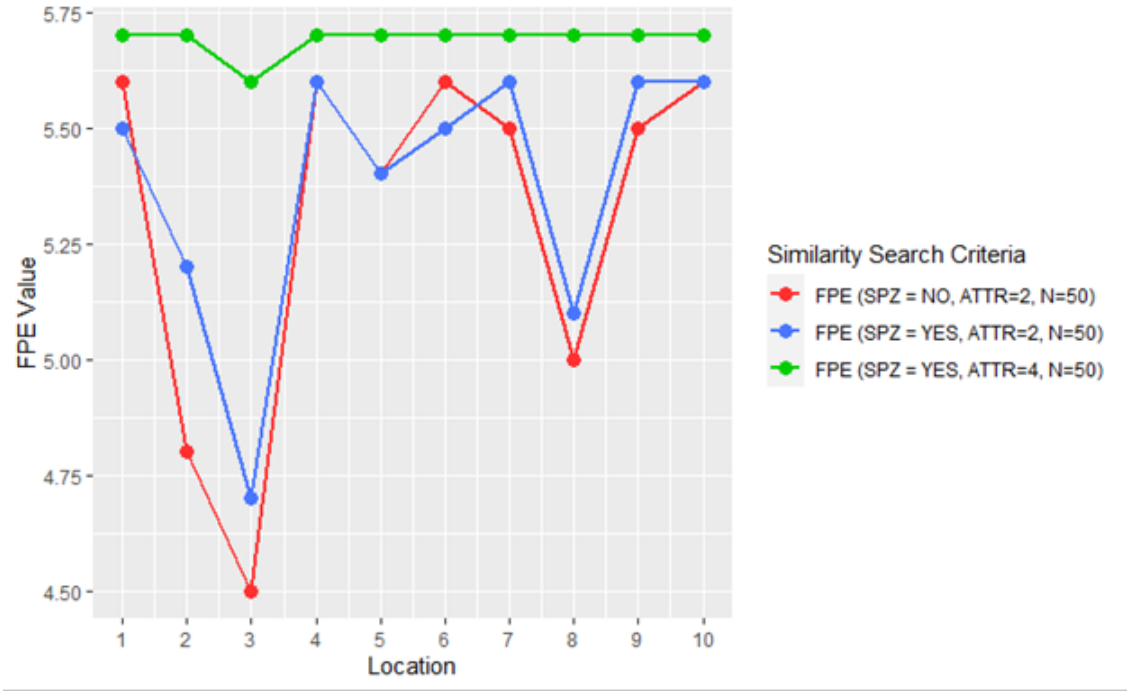

Figure 4.13: Line chart [12] showing FPE values computed for ten sample input locations using three different similarity search criteria. The legend shows each search criteria indicating the use of SPZ (SPZ = YES / NO), number of attributes used in similarity search (ATTR), number of dummy locations generated (N)

generated for sets 1 and 2 indicating that there is no resultant loss in total location privacy achieved by using SPZ instead of the entire dataset for the underlying similarity search. This is in line with our previous finding in Section 4.5.5, in that an equivalent number of similar parcels can be found within the proposed SPZ. This finding also further validates our novel approach of generating SPZ using DBSCAN and the proposed eps estimator algorithm (Algorithm 1).

The enriched similarity search using building footprints and proximity to the closest road is one of the key features of our proposed PLP+ framework. To evaluate the impact of these two attributes, we compute a third set of FPE values where the underlying similarity search to generate dummy locations employs four parcel attributes – area, length, building footprint area, and proximity to the closest road. The FPE values are plotted in green on the line chart shown in Figure 4.13, under the label SPZ = YES, ATTR=4, N=50.

	Building footprint in all of the -		
Similarity Search Criteria	**Top 10 similar parcels**	**Top 25 similar parcels**	**Top 50 similar parcels**
SPZ = NO, ATTR=2, N=50	NO	NO	NO
SPZ = YES, ATTR=2, N=50	NO	NO	NO
SPZ = YES, ATTR=4, N=50	YES	YES	YES

Table 4.4: Showing whether the output parcels from the similarity search contain parcels without building footprint in the Top 10, Top 25, and Top 50 categories under three similarity search criteria.

The efficacy of enriched similarity search is further validated by testing the three sets of 50 dummy parcels under three similarity search criteria for all the 10 sample input locations as summarized in Table 4.3. The ten dummy parcel sets under each similarity search criteria are checked for any parcels without a building footprint in the Top 10 similar parcels. A search criterion is considered to pass the check and marked as YES only if all of the top 10 similar parcels in all the ten dummy parcel sets consist of a building footprint.

As shown in Table 4.3, the similarity search based on building footprints area and road proximity listed under search criteria SPZ = YES, ATTR=4, N=50 did not return any non-building or green area parcels in all of top 50 similar parcels. This indicates that the dummy PLP+ framework is capable of generating up to 50 dummy locations that can withstand a map-matching attack by an adversary.

4.5.8 Discussions

Quantifying Dummy-based Location Privacy

Quantifying location privacy is an important aspect of comparing and determining the efficacy of location privacy algorithms. Theoretical quantification of location privacy is a well-known limitation when it comes to approaches involving dummy locations [23]. Niu *et al.* [34]

proposed entropy-based location privacy derived from the historical query probabilities. To the best of our knowledge, there are no other studies addressing the quantification of location privacy in a spatially similar dummy locations scenario. In this part of the dissertation research, we explore a novel approach to quantify location privacy using land parcel entropy. We use this approach to quantify and compare location privacy results achieved before and after incorporating SPZ, building footprints and street centerline data into parcel similarity search for identifying spatially similar dummy locations.

The Effectiveness of Enriched Similarity Search

As discussed in Section 4.2, being prone to "map-matching attack" and "location-homogeneity attack" is a major drawback in a majority of the existing dummy-based location privacy approaches. By using the outline parcel encompassing the location for spatial context, Tadakaluru [48] successfully addressed the issue of a "location-homogeneity attack"[2 11]. Tadakaluru [48] used a similarity search for finding parcels that are like the input parcel within a city boundary. The two main attributes adopted in the similarity search were the parcels' length and area. In this study, the usage of land parcels stopped dummy locations from being generated in non-building areas such as roads and natural features like rivers, when a true location is located within a parcel area. In some cases, despite being ranked high in similarity with the input original parcel comprised of building area, the dummy locations are generated in parcels with barely any building area and instead consisted of non-building type areas such as green areas.

When dummy locations are forged in areas that are isolated and different in their spatial context from a true location, the dummies are more susceptible to map matching attack by an adversary [30]. We address this in the study through the PLP+ by exploring the use of building footprints [4] and street centerlines [6] along with the land parcels [5] to improve the similarity search results and to find parcels that are greatly similar in a spatial context to the input parcel. As shown in Figure 4.13. , the set of FPE values generated using

building footprints and road proximity (SPZ = YES, ATTR=4, N=50) are larger than those generated based on only two attributes, shape, and area (SPZ = YES, ATTR=2, N=50, and SPZ = NO, ATTR=2, N=50). The FPE values based on building footprints area and road proximity using SPZ consistently outperformed the FPE values from the other two sets. This demonstrates the effectiveness of the attributes - building footprint area and proximity to closest road in increasing the FPE value, resulting in better location privacy for a given input location.

The Effectiveness of Spatial Privacy Zones (SPZ)

Since processing complex and large datasets such as building footprints may incur a heavy computational cost, it is prudent to set an outer boundary and limit the geographical area used for searching the dummy locations while achieving maximum location privacy. In this study, we bring forth the novel concept of spatial privacy zone - SPZ), which is a custom query area of interest for a given true-location input. To obtain a spatial privacy zone for a given input true location, we first run a clustering algorithm on building footprints using optimal search radius – being subsequently used in performing a spatial boundary operation around the user's true location.

As shown in Table 4.2, there is no statistically significant difference in the quality of similarity between the output parcels generated by similarity search using SPZ versus the entire dataset. This has been proven for two different sets of input locations, with one set of locations in SPZ1 and the other in SPZ2. Being able to successfully identify equally similar parcels from a smaller candidate parcel set instead of an entire candidate parcel dataset is not only computationally efficient but also increases the viability of implementing a dummy locations approach using parcel-based similarity search without compromising performance on the client side for the end user. Thus, we have shown that there is no need to search the entire dataset of 123,848 candidate parcels for similar parcels since similar parcels with

no significant difference in the quality of similarity can be identified from a smaller sample dataset, such as 43,563 candidate parcels in SPZ-1, and 31,412 candidate parcels in SPZ-2.

4.6 Empirical Evaluation

The widespread usage of mobile and smart Internet of Things (IoT) devices in our daily lives have led to the universal adoption of Location Based Services (LBS) as a way to customize service offerings anchored on users' geographic locations [47]. In a typical LBS scenario, users share their current location with the LBS service provider in exchange for geographically personalized services without much control over what happens to their location information after service delivery [28].

Dummy generation techniques for location privacy in location-based services or LBS are well-studied. In general, dummy locations techniques operate under the premise that sending dummy locations along with a true location will help conceal the true location from identification by LBS servers or adversaries [59]. Hence, for dummy locations to successfully preserve the location privacy of the true location, it is of paramount importance for the dummies to be indistinguishable from the legitimate one. Without the high degree of similarity between true and dummies, an adversary can exploit the dissimilarities for eliminating dummies and identifying the true location. The dummy identification process plays a vital role in ensuring that both dummy and true locations are indistinguishable from one another, thereby maximizing location privacy.

Lu *et al.* [29] proposed two algorithms, CirDummy and GridDummy, where the dummy locations are generated within a circular or rectangular privacy area computed based on the locations within the query and the privacy area requirements of the user. It addresses the limitations in the earlier studies by generating dummies in a systematic manner within the computed privacy area. But within the privacy area, the dummy locations are chosen without any consideration of the spatial context of the dummies to that of the true location. Niu *et al.* [35] proposed V-circle and V-grid algorithms that not only generate dummy

locations within a privacy area but also blur them to locations that are similar in query probability of the true location.

4.6.1 Comparison with existing approaches

The minimum bounding area (MBAR) has been demonstrated as a good indicator of the degree of location privacy achieved by a location set comprising both true and dummy locations [29, 35, 11, 22]. The MBAR of a location set is calculated as the area of the convex hull encompassing the locations within the location set containing both true and dummy locations. A higher MBAR value of a location set indicates better location privacy achieved by the dummies within the location set. To showcase the efficacy of the parcel-based location privacy approach, the maximum MBAR achieved by a location set by PLP+ was compared with MBAR results from the existing approaches such as CirDummy, GridDummy, [29] V-circle, V-grid [35].

The privacy area comparison results from Niu *et al.* [35] are directly adopted and used in our comparative analysis due to a lack of relevant data. Niu *et al.* [35] used the Borlange dataset [16] associated with the City of Borlänge in Sweden to compute the results for comparison between MBAR referred to as the privacy area in Niu *et al.* [35] versus the location set size (k). To facilitate a fair comparison, the sample input locations for PLP+ are chosen from a SPZ with the size of an area similar to the City of Borlänge. The MBAR results for PLP+ are calculated from the maximum convex-hull area bounding the location set at various sizes, ranging from 3 to 10 for multiple sample input true locations.

As shown in 4.14 the location set size (k) is plotted against the MBAR in square miles (mi^2) for each of the five dummy-based algorithms including PLP+. The MBAR values for PLP+ at various location set sizes ranging from 3 to 30 are comparable to the MBAR values of the other algorithms plotted in 4.14. At the outset, PLP+ appears to perform better than the circle base algorithms V-Circle and CirDummy but not as well as the GridDummy and V-grid algorithms. The PLP+ is designed to consider spatial context and skip non-building

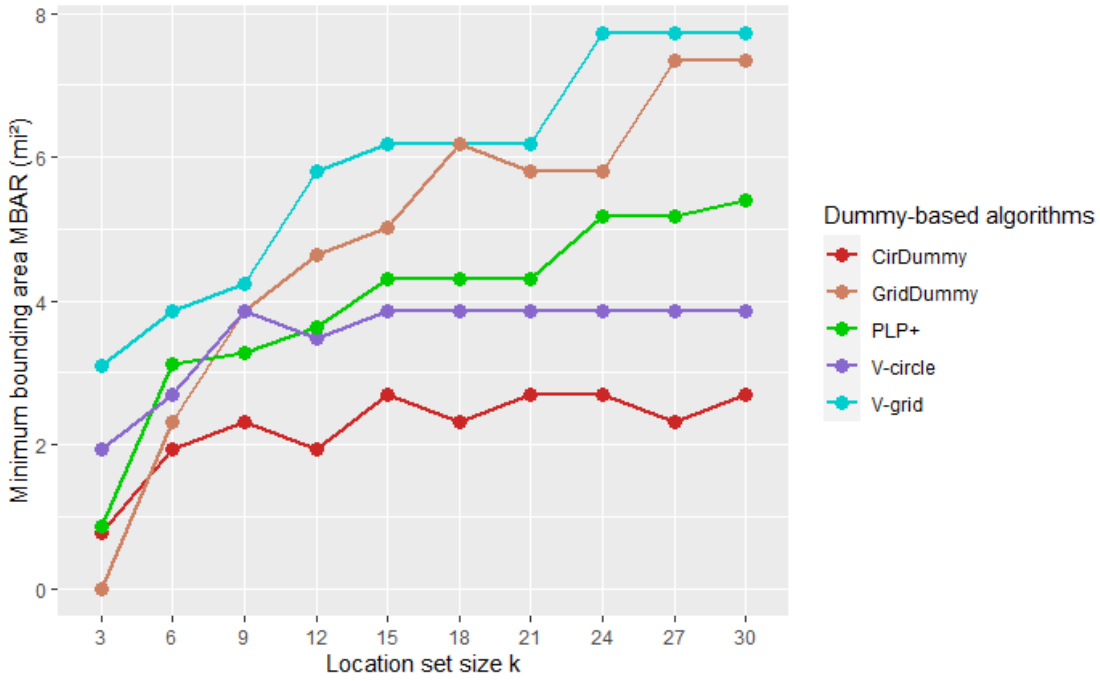

Figure 4.14: Line chart comparing location set size (k) versus minimum bounding area (MBAR) for dummy-based location privacy algorithms.

areas such as roads, lakes, green areas, etc. to mitigate the possibility of map-matching attacks on the dummy locations. This can reduce the total amount of candidate area available for searching the dummy locations within an SPZ and can lead to lower increments in MBAR as the location set size increases, as shown in the above plot. To the best of our knowledge, the PLP+ is the only algorithm that considers spatial context, allowing it to produce dummy locations that are capable of withstanding map-map matching attacks. The dummy locations generated by the other algorithms are susceptible to map-matching attacks because of the lack of consideration of the spatial context of the true location in the dummy generation process. In conclusion, the PLP+ achieved a comparable MBAR at diverse location set sizes, while successfully addressing map-matching attacks.

4.6.2 Location Privacy Threat Analysis

Sharing location information has become a standard practice for accessing location-based services or LBS. It is well understood that in order to receive geographically customized

results, users must share their current locations with LBS service providers. The lack of control over what happens with users' location information after being delivered to an LBS server coupled with concerns for potential misuse or mishandling of the location information demands the need to protect the location privacy of the users. This scenario highlights a pressing need for users to be able to receive geographically customized results without compromising location privacy. Dummy location approaches become one such technique that allows users to enjoy geographically customized services of the highest quality without giving away users' true locations to LBS servers.

The validity of the presumption that LBS servers or adversaries will not be able to distinguish between true and dummy locations depends significantly on how similar the dummy locations are to their real counterparts. The greater the similarity between the true and dummy locations, the harder it is for an adversary to decipher a true location from dummy locations, which in turn leads to optimized location privacy [59]. A raft of studies has used a variety of information, such as query probability of a location [35, 15], geographical distribution [35, 34], and location semantics [60, 10] for evaluating the similarities between the true and dummy locations.

The two major limitations of the state-of-the-art dummy-generation techniques: the existing solutions are subject to map-matching attacks and location homogeneity attacks as explained in Section 1.2. Relying on the similarity in spatial context, the COSA framework successfully safeguards against location homogeneity attacks [48]. Despite this success, the COSA failed to adequately address concerns about map-matching attacks. This issue is more evident in scenarios when a real location is located in an input parcel with substantial building area. In this scenario, COSA is susceptible to generating the dummies in parcels that are barely developed or mostly green areas rendering the dummies an easy target for elimination in a map-matching attack by an adversary.

In a location homogeneity or a map-matching attack, an adversary discerns the true location through the process of identifying and eliminating the dummies and thus pose a

threat to the location privacy of the user. In a location homogeneity attack, the dummies fail to protect the true location from identification since they are placed within the same parcel as the true location. The parcel-based dummy approach originally proposed in COSA and enhanced in PLP+ addressed the location homogeneity attack by placing the dummy locations outside of the true location's parcel. For this, a total of 500 dummy locations generated by PLP+ for ten sample input locations with 50 dummy locations each were evaluated. Out of 500 dummy locations, the total number of dummy locations sharing the same parcel as their true location was zero. In addition, none of the 500 dummy locations were in a parcel with zero building footprint, indicating their effectiveness against a map-matching attack. This demonstrates that PLP+ is capable of producing dummies that are effective against both location homogeneity and map-matching attacks.

4.7 Summary

Our major contributions made throughout this chapter are as follows:

- We propose a novel dummy generation approach to generate dummy locations based on the geographical feature similarity to true locations, and we implement the Enhanced Parcel-based Location Privacy (PLP+) framework. Unlike many existing approaches that use synthetic datasets, the PLP+ framework is built and evaluated using real-world geospatial datasets such as parcels, building footprints, and street centerlines for a county in the state of New York.

- We implement enriched parcel similarity search powered by building footprints and street centerline data to identify spatially similar parcels and dummy locations.

- We develop a novel privacy area generator anchored on building footprints and DBSCAN clustering to generate spatial privacy zones for the entire city dataset. We propose and implement a novel DBSCAN parameter estimator algorithm to calculate

an optimal search radius (eps) value for any given input location dataset. We successfully and seamlessly integrate the application of spatial privacy zone, SPZ, within PLP+ framework where the relevant SPZ for a given input real location is retrieved and used as a query boundary within the enhanced similarity search.

- We statistically evaluate the results of enhanced similarity search for multiple sample input location datasets using SPZ versus no SPZ. We demonstrate that the similarity search within the SPZ boundaries is adept at fetching equally similar parcels compared to searching the entire county parcel dataset, thus proving the effectiveness of our novel DBCAN-based SPZ generation approach and eps estimator algorithm.

- We delve into the implementation of a novel approach to quantify location privacy using building footprint entropy to measure uncertainty associated with a given location set that includes both true and dummy locations. Using our building footprint entropy metric, we accurately measure and compare the total location privacy offered by dummy locations forged using enriched similarity search in the PLP+ framework versus dummies from the other parcel-based dummy generation methods.

- Our results confirm that enriched parcel similarity search using building footprint and road proximity within an SPZ yields better dummy locations with higher building footprint entropy than the existing approaches in an urban setting. More importantly, none of the Top 50 dummy locations are placed in vacant parcels non-building, or green areas. This finding indicates that the PLP+ framework is capable of generating dummy locations that are resistant to map-matching attacks. Since PLP+ is a parcel-based framework, our technique already eliminates the possibility of location-homogeneity attacks articulated in a prior study[2 11]. Therefore, our PLP+ framework meets our proposed design goal and objectives, stipulated in Section 1.3.2, by simultaneously addressing map-matching and location-homogeneity attacks.

Chapter 5

A Voronoi-based Semantically Balanced Dummy Generation Framework for Location Privacy

In this chapter, we present the VSBDG framework which is organized as follows. 5.1 presents the background. 5.2 describes the temporal constraint attack.Section 5.3 presents the proposed methodology. In Section 5.4, we provide a detailed explanation of the VSBDG algorithm. Section 5.5 articulates the experimental implementation of VSBDG on one sample location followed by detailed experiments to verify the VSBDG algorithm's effectiveness. Section 5.6 evaluates the results from Section 5.5.3.

5.1 Background

More often than not, mobile apps and websites that offer LBS services require the user to provide their current locations. The potential for misuse of location information by LBS providers coupled with the concerns over a possible breach of LBS servers resulting in exposing user information to an adversary makes a strong case for location privacy protection [23]. In many cases, the quality of LBS services is directly related to the accuracy of user locations. In other words, lowering the accuracy of the true locations lowers the quality of results returned by an LBS query. The main goal of any location privacy algorithm is to maximize quality of service while protecting user locations by sharing as little as possible or not sharing the exact locations of users. The current approaches for location privacy can be classified into four main categories – cloaking, dummy location, obfuscation, and cryptographic [23].

Cloaking and obfuscation do not send the true location of the user to the LBS server, thus resulting in lower-quality of services, specifically LBS services that require the exact location of the user [53]. The cryptography-based approaches are computationally intensive

rendering them impractical [28]. The dummy location approaches involve sending a user's genuine location along with the dummy locations making these preferable for achieving high-quality LBS. The two main categories of LBS services are (1) snapshot and (2) continuous services [23]. In a snapshot LBS, the user's location is submitted only once to the LBS server for query results, whereas in a continuous LBS, user locations are continuously reported to an LBS server to receive up-to-date query results. An example of snapshot LBS would be searching for the nearest point of interest (POI) such as restaurants or hotels. An example of continuous LBS services would be using an application for driving directions where the user's current location is continuously sent to the LBS server to track the user's movement and provide up-to-date driving directions [23]. This study's research mainly focuses on generating dummy locations for location privacy in the context of a snapshot LBS.

5.2 Temporal Constraint Attack

In a temporal constraint attack, an adversary uses location information from historical dummy-based LBS requests delivered by a specific user and exploits the differences in temporal constraints between semantic categories such as residential versus POI in a historical timeline to separate the dummy locations from a true user location. By eliminating the dummies using a temporal constraint attack, an adversary is likely to identify the real location of the user breaching the location privacy of the user. The primary purpose of a residential location is housing, and it is where people live [40]. This fundamental assumption that people live in their houses and not in their workplace or a restaurant forms the basis for our argument that a residential location has different temporal constraints from point of interest (POI) based locations. When dummy locations are generated without consideration of the distinction between residential and non-residential locations, an adversary can exploit the semantic difference to eliminate dummy locations based on temporal constraints and identify the true location of a user.

	Location Semantic Type				
Time	**l_1**	**l_2**	**l_3**	**l_4**	**l_5**
2:20 AM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station
4:01 AM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station
6:10 AM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station
9:15 AM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station
11:50 AM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station
3:00 PM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station
8.00 PM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station
11:20 PM	Restaurant	$\text{Residential}_{\text{True}}$	Supermarket	Shopping mall	Gas station

Table 5.1: Showing the semantic information associated with locations in sample LBS requests.

For a given location, an adversary can use approaches such as reverse geocoding [27] and other background information to find the specific address and the semantic category of the location such as a residential, or supermarket. Using this technique, an adversary can retrieve semantic categories for all the locations in a dummy-based historical LBS request. The adversary can then compare and evaluate the semantic categories of all these locations in a historical timeline. By doing this evaluation within the context of temporal constraints associated with semantic categories such as residential versus non-residential purposes, an adversary can eliminate the dummy locations and increase the probability of true location identification. Depending on the amount of historical location requests possessed and the extent of semantic diversity implemented in generating dummy locations, an adversary tends to be able to successfully eliminate all dummy locations and to identify users' legitimate locations.

In the following example, we demonstrate how an adversary can exploit historical dummy-based LBS requests by viewing them in a temporal context of a single 24 hour period beginning at midnight (12 AM) to 11.59 PM on the same day. Table 5.1 displays the semantic information associated with locations in sample LBS requests containing true and dummy locations sent to the LBS server by a single user at various times over one day. Each row corresponds to a request made at a certain time of the day and shows semantic categories for five locations $\{l_1 \dots l_{k=5}\}$ where k is the total number of locations dispatched to the LBS server in each request, with k-1 dummy locations and one true location ($Residential_{True}$).

In this concrete example, we assume that an adversary possesses background information that all the LBS requests listed in the table belong to the same user and request type, and occurred on the same day within 24 hours between midnight (12 AM) to 11.59 PM a specific time zone. We also assume that the dummy locations submitted for a given true location don't change over time because the same dummy algorithm is used to generate the dummy locations for every new request. Moreover, it is assumed that a POI location is a place of business and is not used for residential purposes. The main goal of an adversary in a temporal constraint attack is to eliminate the k-1 locations and identify the k^{th} location that is also a true location. In the above example, it is easy to deduce that a location with a semantic type shopping mall is a dummy location since its unusual to be at a shopping mall at 2.20 AM in the morning and also to be at a shopping mall throughout the entire day. Using the same logic, we demonstrate that the two semantic types, -restaurant and supermarket, can be further eliminated. Although being at a gas station at 2.20 AM is possible, it can be pruned as a dummy as it's unusual to be at the same gas station throughout the day. By successfully eradicating the four locations, the fifth location of the residential semantic type that remains is identified as the legitimate location. The adversary can also erase all four business locations since it is not usual for a user to be at a place of business for an entire day, and can identify the residential location as the true location.

The example shows that despite the maximum semantic diversity between true and dummy locations, the adversary can still exploit the semantic information from the historical requests data and identify the true location by wiping out the k-1 dummy locations. The current dummy approaches fail to integrate the intrinsic semantic particularity, which in this case is the notion of home associated with a residential location [40], and instead treat the residential location as a general semantic type without any special attention. This could result in potential dummy locations that are susceptible to temporal constraint attack as shown in the example above.

5.3 Proposed Methodology

Our proposed methodology is a Point of Interest (POI)-based approach to producing semantically balanced dummy locations. A physical location associated with a POI is employed for delineating the influence of the POI within a geographical space through Voronoi polygons. Within each Voronoi POI influence area, the parcel-based dummy generation framework devised by Tadakaluru [48] is then deployed to create dummy locations that are spatially similar to legitimate ones. These topics are further articulated in detail in Sections 5.3.1- 5.3.4.

5.3.1 Relationship between Geographic Location, Address, and Land Parcel

A geographic location is an exact physical place on the Earth's surface usually represented by a unique latitude and longitude pair. In a majority of the state-of-the-art location privacy studies, the term location is generally used to refer to a geographic location. An adversary can leverage the street address associated with the geographic location to obtain sensitive information about a user [24] and; hence, this trick plays an important role in building privacy-preserving mechanisms. An address is generally associated with a parcel of land that has designated property ownership boundaries. When someone refers to a specific geographic location (i.e., longitude, latitude) in an urban setting, it is normally linked to an address that is representative of a land parcel. In other words, any physical location located within a land parcel is associated with a unique address assigned to the land parcel [55]. Despite the importance of address and land parcels, most dummy-location generation approaches in location privacy assume that a single physical location is enough to represent an address and the underlying land parcel. As a result, these solutions use physical locations as a sole spatial component within their algorithms. In this study, we make use of land parcels rather than single locations to elect areas that are similar in a spatial context to the area containing real locations. Applying the land parcels, we can delineate the geographical

boundary of a POI point location to support the distinction between residential and non-residential locations in the VSBDG algorithm. Without this delineation offered by the land parcel, it is impossible to guarantee that a chosen location is of a certain semantic type – an important requirement for building a semantically balanced dummy location set.

5.3.2 Modeling POI Influence Using Voronoi Polygons

A Voronoi diagram anchored on POI locations is employed to divide a geographical area into polygons, with each enclosing a single POI location such that any location within a given polygon is closer to the related POI than any other POI locations [14]. The locations lying on the edge of a Voronoi polygon are equidistant to POIs associated with the two Voronoi polygons sharing an edge. The main purpose for adopting Voronoi polygons in the proposed approach is to avoid dynamic runtime analysis of POIs in such a way that an efficient dummy generation process becomes viable. This idea is made possible by the intrinsic property of a Voronoi polygon, which contains only a single POI. This novel design facilitate the selection of deterministic proportions of POIs and residential locations within a geographic area influenced by given POI, a Voronoi polygon. The Voronoi polygons are generated as part of data pre-processing and reused for each new dummy generation request. The predictability of POIs within a group of Voronoi polygons eliminated a pressing need for proximity queries looking for the POIs within a geographical area influenced by a POI during runtime. Given the Voronoi polygons for a POI dataset, a semantically balanced location set L containing both true and dummy locations can be formally expressed as:

$$\mathrm{L} = \{l_1, l_2, \ldots, l_m, l_{m+1}, l_{m+2}, ..., l_n\} \text{ , where}$$

$$n : \text{total number of locations,}$$

$$m : \text{number of non-residential locations (POIs),}$$

$$n\text{-}m : \text{number of residential locations.}$$

- m is defined as the number of Voronoi polygons used in generating a dummy location set because each Voronoi polygon contains a one POI location

- With at least two locations selected from each Voronoi polygon, there is *2m+1* minimum number of dummy locations and one legitimate location.

5.3.3 Cosine Similarity between Voronoi Polygons

In this study, we advocate for cosine similarity to find spaces that are similar to the Voronoi polygon embracing users' genuine locations. The cosine similarity between two vectors A and B is measured using the cosine angle between the vectors, which can be calculated in the Euclidean space [50] using the following formula:

$$\text{Cosine Similarity (A , B)} = \frac{\sum_{i=1}^{n} \mathrm{A}_i B_i}{\sqrt{\sum_{i=1}^{n} \mathrm{A}_i^2}\sqrt{\sum_{i=1}^{n} \mathrm{B}_i^2}} \tag{5.1}$$

The cosine similarity gauges similarity based on the direction of vectors using cosine angle instead of the magnitude of vectors. This measure aligns with the goal to identify Voronoi polygons that are similar to the input Voronoi polygon in POI influence in a spatial context rather than the magnitudes of feature attributes. For this reason, the attribute vectors of both target and candidate Voronoi polygons are compared and ranked using cosine similarity ordered with highly similar polygons at the top and least similar at the bottom. Dummy locations are picked from the Voronoi polygons that are most similar in cosine relationship to the input Voronoi polygon containing the true location.

5.3.4 Parcel-based Similarity Search

The parcel-based location privacy framework proposed by Tadakaluru [48] uses similarity search to elect dummy locations that are similar in a spatial context to real locations. In this study, the parcel-based similarity search is deployed to forge residential dummy locations within each candidate Voronoi polygon that is cosine similar to the Voronoi polygon containing the input location. The parcel-based similarity search is driven by the Euclidean distance between the attribute values of target and candidate parcels. For each target and

candidate parcel, the total sum of squared differences (SSD) between standardized attribute values is calculated. The candidate parcels are then ranked based on their SSD values with the target parcel, where the candidate parcel with the lowest SSD value ranked higher is considered to be the most similar one to the input parcel. The SSD between two parcels P and Q with n attributes can be calculated as follows.

$$SSD(P, Q) = \sum_{i=1}^{n} (\mathrm{P}_i - Q_i)^2 \tag{5.2}$$

5.4 Voronoi-based Semantically Balanced Dummy Generation (VSBDG)

The main objective of the proposed approach is to generate semantically balanced dummy locations that can effectively withstand a temporal constraint attack, thereby maximizing location privacy for users. The semantic balance between residential and non-residential dummy locations is accomplished by dividing a geographical area, such as a city or county bounding a true location, into separate regions based on POI influence. The Voronoi polygons are generated using the POI dataset, where each Voronoi polygon is associated with a single POI location (see also Section 5.3.2). The key rationale behind the deployment of Voronoi is to facilitate a guaranteed and predictable selection of one POI within each Voronoi polygon, control the ratio of residential versus non-residential (POI) dummy locations, and preserve the semantic balance of a given location set.

Algorithm 1 originates a semantically balanced location set for a given legitimate location using land parcels, POI-based Voronoi polygons, and POIs dataset(s) for a geographical region like a city or county. The first steps in rows 1 and 2 identify the relevant land parcel p_{true} and Voronoi polygon v_{true} outlining the true location l_t using spatial join. Then, the next major step involves using cosine similarity search to identify top m Voronoi polygons that are similar to v_{true}. A next similarity search based on the Euclidean distance between attributes is performed to pinpoint the land parcel ($parcels\_similar_i$) that is most similar to p_{true} in each of the m Voronoi polygons. Finally, the associated POI and the centroid of

$parcels\_similar_i$ for each of m Voronoi polygons are appended to the dummy location set, returning *2m+1* dummy locations in total.

Algorithm 3: VSBDG - To identify semantically balanced dummy locations for a given residential true location

Require: True location l_t (Longitude, Latitude), Location set size k

{Input Datasets: Land Parcels P, POI-based Voronoi polygons dataset V}

Ensure: Location set D of size k

1: Determine land parcel p_{true} outlining l_t using spatial join between l_t and P

2: Determine Voronoi polygon v_{true} outlining l_t using spatial join between l_t , V

3: m = (k-2)/2 {m is total number of similar Voronoi polygons to be identified}

4: $V_{similar}$ = Cosine Similarity Search(target= v_{true}, candidate set = V, output length = m)

{Perform cosine similarity search to identify top m Voronoi polygons from V that are similar to v_{true}}

5: **for** each Voronoi polygon v_i in $V_{similar}$ **do**

6: Set $candidate\_parcels\_set_i$ = parcels within Voronoi polygon v_i

{Perform Euclidean Similarity Search to identify top 1 land parcel from $candidate\_parcels\_set_i$ that is most similar to p_{true}}

7: $prclSim_i$ = Euclidean Similarity Search(target = p_{true} , candidate set = $candidate\_parcels\_set_i$, output length=1)

8: Calculate $dummy_{residential}$ using parcel centroid of $prclSim_i$

9: $dummy_{poi}$ = POI location associated with v_i

10: Add $dummy_{residential}$, $dummy_{poi}$ to D

11: **end for**

12: $dummy_{vt}$ = POI location associated with v_{true}

13: Add $dummy_{vt}$ to D

14: Add l_t to D

15: **return** D

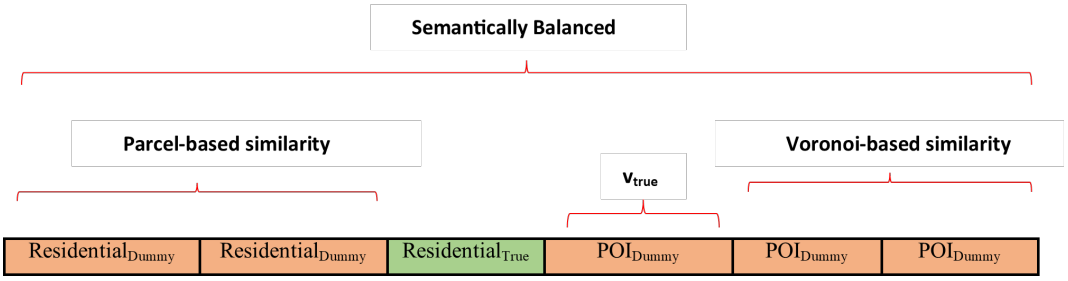

Figure 5.1: A three-tiered location privacy protection for a location set that contains one legitimate location and five dummy ones (k=6).

The semantic composition of the location set from Algorithm 1, containing both true and dummy locations, consists of both residential and non-residential locations regardless of the semantic classification of the real location. This intention ensures that an adversary is unable to single out the real location by exploiting the background information associated with a temporal constraint, such as general operating hours for a certain type of POI [57]. The example in Table 5.2 illustrates the possible residential versus POI semantic composition of a location set that contains one true location and six (k =5) dummy locations over a period of one day, like the example shown in Table 5.1.

Our algorithm provides location privacy for the true location at three different tiers as shown in Figure 5.1. First, by consistently generating the same number of residential and POI dummy locations each time, it decreases the chance to single out the real location using a temporal constraint attack by an adversary. Second, the POI dummies are identified based on similarity in POI influence related to the true location through the use of Voronoi polygons. Third, the residential-type dummies are identified based on their parcel similarity to the parcel outlining the true location using a parcel-based similarity search. Whether the genuine location is a residential location or a POI, it is arduous for an adversary to differentiate a particular location as a true location: for any given location set, there are at least one-half of the locations that are similar to the real location, while the other half of the locations are similar to each other. The following example, revisiting the scenario

		Location Semantic Type				
Time	l_1	l_2	l_3	l_4	l_5	l_6
2:20 AM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI
4:01 AM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI
6:10 AM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI
9:15 AM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI
11:50 AM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI
3:00 PM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI
8.00 PM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI
11:20PM	Residential	Residential	$\text{Residential}_{\text{True}}$	POI	POI	POI

Table 5.2: Showing the semantic information associated with locations in sample LBS requests.

presented in section 5.2, evaluates how VSBDG addresses the temporal constraint attack in that scenario.

Table 5.2 illustrates the semantic information associated with an example location set containing both true and dummy locations generated and submitted to an LBS server at various times during a single 24 hour period beginning at midnight (12 AM) to 11.59 PM on the same day. Even if an adversary employs a temporal constraint attack similar to the scenario described in Section 5.2 and identifies all the three POIs as dummies, there are still two more residential locations to identify and eliminate for the true location. The two dummy residential locations, being similar in a spatial context to the real one, are generated using parcel-based similarity search that is proven to be effective against location homogeneity attacks and less prone to map-matching attacks [48]. By leveraging Voronoi and parcel-based similarity search, the VSBDG algorithm generates a semantically balanced dummy location set that effectively withstands temporal constraint attack while preserving the indistinguishability of real locations.

5.5 Experimental Analysis and Results

5.5.1 Data Collection and Preprocessing

In the empirical study, we test land parcels [5] and POI datasets [7] for spatial analysis and generation of dummy locations. These datasets are gleaned for the Richmond County

(Staten Island) in the state of New York, USA. We run the geoprocessing toolkit in ArcGIS Pro [38] to extract county-level parcel and POI data from statewide New York datasets and for the rest of the analysis in this section. The parcels dataset is comprised of parcel features that are stored as polygon features, and the POI datasets contains POIs that are point features as shown in Figure 5.2(a). The total number of land parcels and POI datasets of the Richmond County (Staten Island) are 123,849 and 1,288, respectively [3]. As stated in Section 5.3.2, the Voronoi polygons are generated for 1,288 POI locations during preprocessing and referred to within the proposed algorithm as candidate set V.

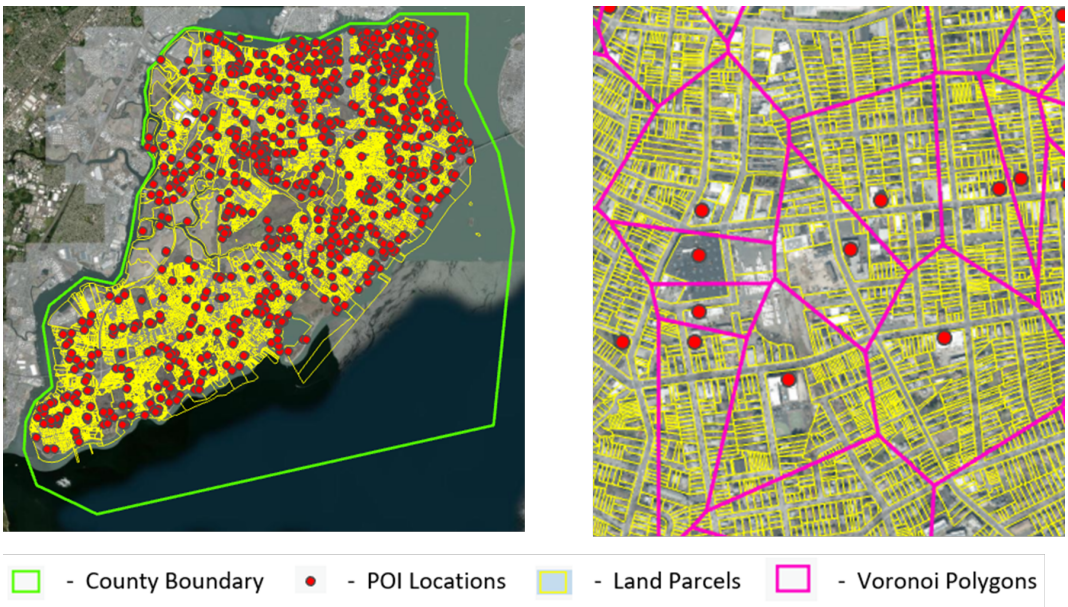

Figure 5.2: (a) Land parcels and POI locations within the Richmond County (Staten Island) overlaid on imagery basemap [1] (b) Voronoi polygons and their associated POIs within a section of Richmond County.

5.5.2 Electing Dummy Locations using VSBDG

This section demonstrates a step-by-step implementation of the VSBDG algorithm for a sample input true location l_t and location size (k=2). The first steps in the algorithm are to determine parcel p_{true} and Voronoi polygon v_{true} outlining a true location as shown in Figure 5.3. The next step is to perform cosine similarity search and find top m Voronoi polygons similar to v_{true}. The m is calculated as half of k-1 value, where k is the number of

dummy locations to be created. This step is to achieve semantic balance to ensure that for every POI dummy location, there is a residential dummy location chosen.

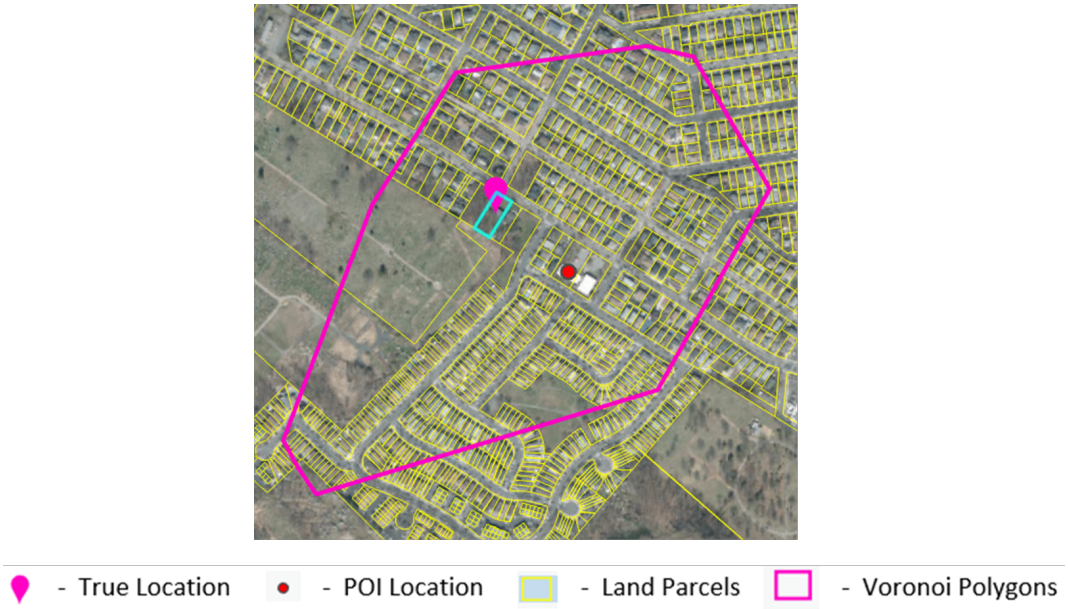

Figure 5.3: Input parcel p_{true} outlining the true location (highlighted in blue) and the Voronoi polygon v_{true} containing the input true location l_t as indicated in steps 1 and 2 of VSBDG (Algorithm 1)

The next step in the algorithm is to determine the top m Voronoi polygons similar to v_{true} from the candidate set V generated during the preprocessing phase. As delineated in Section 5.3.3, the search for similar Voronoi polygons is driven by cosine similarity and uses the attributes area, length, and the number of land parcels within the Voronoi polygon to perform cosine similarity search. The cosine similarity search is executed for two dummy locations (k=2) and the output for the top one similar Voronoi polygon v_1 (m=1) is shown in Figure 5.4. The next steps involve identifying the residential parcel that is being the most similar to the input parcel of the real location p_{true} within each of the m similar Voronoi polygons. Figure 5.4. shows a residential parcel $prclSim_1$ that is most identical to p_{true} from all the candidate parcels within v_1 identified by the Euclidean similarity search for similar parcels. A residential dummy location $dummy_{residential}$ is calculated using the centroid of $prclSim_1$ and the POI location linked with v_1 is chosen as the non-residential POI dummy $dummy_{poi}$. As explained in the algorithm, Steps 5 to 10 involving parcel similarity search, is

repeated for each of the m similar Voronoi polygons identified using cosine similarity search in the previous step. The POI location $dummy_{vt}$ associated with v_{true} is included as one of the dummies. Thus, two dummies constructed from each of m Voronoi polygon and one additional dummy $dummy_{vt}$ results in total *2m+1* dummy locations for a given genuine location p_{true}. With *2m+1* dummy locations evaluating to three dummy locations, there are total of four locations (k=4) including the true location in the location set as shown in Figure 5.5. Figure 5.5a depicts all the four locations in the context of Voronoi polygons, and Figure 5.5b shows only the locations.

Figure 5.4: Voronoi parcel similar to v_{true} from cosine similarity search, residential parcel $prclSim_1$ similar to p_{true} from Euclidean similarity search and the two dummies ($dummy_{residential}$ and $dummy_{poi}$) identified.

5.5.3 Results

Physical dispersion of dummy locations in a location set is employed to evaluate the effectiveness achieved by the location privacy algorithms [10, 60]. The physical dispersion is measured as the minimum distance between any two locations in a location set containing both true and dummy locations [59]. A location set with a higher physical dispersion indicates that locations are scattered much farther, implying better location privacy. The core

Figure 5.5: For a location set with four locations (k=4), (a) shows the true location and three dummy locations with their respective Voronoi polygons, and (b) shows the true and dummy locations only.

idea for building a semantically balanced location set is to ensure that for every residential location, there exists a POI location and vice-versa to reduce the probability of being identified by adversaries. This goal is achieved through Voronoi polygons by choosing a set of POI and residential dummy locations from within multiple geographical areas that are spatially similar to the Voronoi area of real locations.

For a semantically balanced location set to be effective, dummy locations should not only have a higher physical dispersion within each category but should also have a similar physical dispersion as the other semantic category. gauge the effectiveness of the semantically balanced location set generated by VSBDG (Algorithm 1), we bring forth an algorithm on three input true locations elected using random sampling with different location set sizes. The *VSBDG (True location lt , Location set size k)* algorithm is implemented for three input true locations with location sizes (k) ranging from 4 to 22. Given each sample input true location, ten location sets containing both true and dummy locations are generated with sizes ranging from k=4 to k=22. For each location set, the minimum dispersion distance (MDD) is separately calculated for locations in residential and semantic categories. Table

5.3. tabulates the physical dispersion of residential locations for 10 location sets generated for the input legitimate location in three columns T-RES-1, T-RES-2, T-RES-3. Table 5.4. shows the physical dispersion of POI locations for 10 location sets generated for input true location in three columns T-POI-1, T-POI-2, T-POI-3. The minimum dispersion distance for each location set is plotted in RStudio[3.34] against location set size (k) separately for residential and POI locations as shown in Figure 5.6 and Figure 5.7.

	Minimum dispersion distance - Residential (meters)		
Location set size (k)	**True Location-1 (T-RES-1)**	**True Location-2 (T-RES-2)**	**True Location-3 (T-RES-3)**
4	9291.308514	13685.24076	26978.01745
6	9291.308514	10923.37298	11920.92343
8	9291.308514	2073.75585	11834.15819
10	9291.308514	2073.75585	6944.580188
12	4076.586856	2073.75585	6944.580188
14	4076.586856	2073.75585	3461.144276
16	3194.775997	2073.75585	3461.144276
18	3101.138163	2073.75585	2069.387356
20	3101.138163	1618.185695	2069.387356
22	3101.138163	1618.185695	2069.387356

Table 5.3: Showing physical dispersion of residential locations in a location set of different sizes (k) of the three input locations.

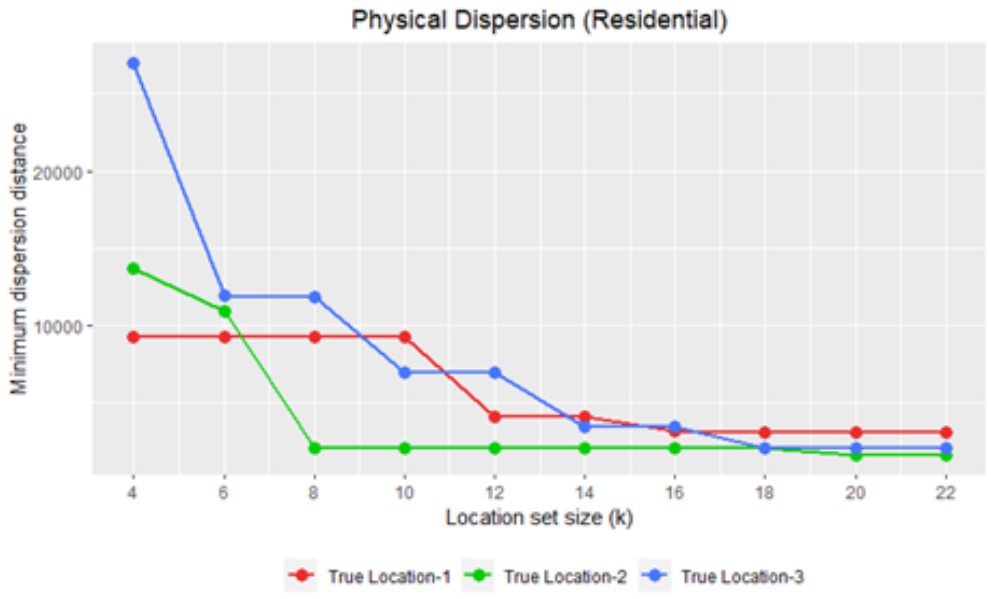

Figure 5.6: Plots show physical dispersion of residential locations in a location set with the size of location set (k) on X-axis and minimum dispersion distance (meters) on the Y-axis.

Location set size (k)	Minimum Dispersion Distance – POI (meters)		
	True Location-1 (T-POI-1)	True Location-2 (T-POI-2)	True Location-3 (T-POI-3)
4	10368.38844	14322.61609	25800.26266
6	10368.38844	8369.51998	12034.79693
8	10368.38844	3641.198031	11580.28315
10	10368.38844	3641.198031	7445.09195
12	3427.856197	3641.198031	7248.846177
14	3427.856197	3641.198031	3037.160013
16	3427.856197	3641.198031	3037.160013
18	2799.307792	3310.462408	2191.07635
20	2799.307792	3310.462408	2191.07635
22	2799.307792	3310.462408	2191.07635

Table 5.4: Physical dispersion of POI locations in a location set of different sizes (k) for the three input locations.

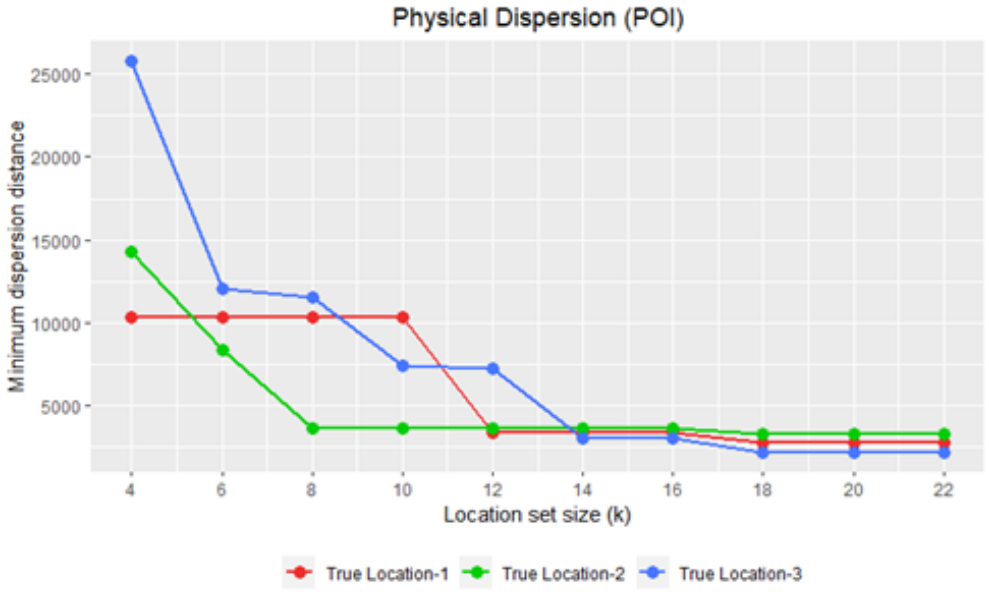

Figure 5.7: Plots show physical dispersion of POI locations within a location set with the size of location set (k) on X-axis and minimum dispersion distance (meters) on the Y-axis.

Input True Location	Vectors measured	Physical Dispersion Cosine Similarity (PDCS)
1	T-RES-1 and T-POI-1	0.9966669
2	T-RES-2 and T-POI-2	0.9641056
3	T-RES-3 and T-POI-3	0.9993439

Table 5.5: Cosine similarity is measured between residential and POI semantic categories for the three input true locations.

The effectiveness of a semantically balanced location set is evaluated by measuring physical dispersion similarity between residential and POI MDD values at various location set sizes as shown in Table 5.3 and Table 5.4. For this, the cosine similarity is calculated between MDD values of residential and POI categories of each of the three input true locations separately.

5.6 Discussions

5.6.1 Evaluating VSBDG

The MDD of residential locations in the 10 location sets with sizes ranging from 4 to 22 are almost similar and follows a nearly identical trend as the MDD of POI locations in the same 10 location sets. This trend holds true for all the three sample locations plotted using red, green, and blue colors in Figure 5.6 and Figure 5.7. The same is empirically proven by measuring the cosine similarity between the MDD values of residential locations versus POI locations in 10 location sets for all three input locations as shown in Table 5.5. The PDCS measures of input locations 1, 2, and 3 is 0.997, 0.994, and 0.99 respectively with an average PDCS value of 0.988, indicating a high cosine similarity between MDD of residential locations versus POI locations in the 10 location sets with sizes ranging from 4 to 22.

The high average PDCS value of 0.988 between the MDD values of residential and POI locations indicate the effectiveness of the VSBDG algorithm in generating a semantically balanced location set. This high cosine similarity between the two semantic categories also demonstrates a strong semantic balance between the semantic categories that is consistent

even at higher values of k. This result also unveils that VSBDG is capable of making a scalable semantic balance even at high values of k by equally creating efficient dummy locations, despite an increase in the size of the location set (k). This finding further unravels that our proposed algorithm is adroit at originating dummy locations that are consistent with their resistance to temporal constraint attacks despite the increase in the size of the location set k.

Without a high physical dispersion, the dummy locations are prone to location distribution attacks [51] where, in this case, an adversary can target locations from a specific semantic category. The adversary employs techniques such as clustering for eliminating dummies, either to identify true locations or the neighborhood area of the true location. The latter would pose a much higher risk in a case where the real location is residential since the adversary infers a lot of background information by knowing the neighborhood area associated with clusters where a residential user resides [45]. The average MDD of residential locations for all the 10 location sets for each input location shown in Table 5.3 is 5861.894 meters. The average MDD of POI locations for all the 10 location sets for each input location listed in Table 5.4 is 6258.046 meters. These high MDD values within each semantic category demonstrate that locations within each category are scattered farther apart indicating optimized location privacy.

5.6.2 Comparison with the Existing Dummy Approaches

Recall that to the best of our knowledge, this study is the first of its kind that introduces temporal constraint attacks, which is tackled by the VSBDG algorithm. Because of the novelty of VSBDG, it is not feasible to conduct a direct comparison of our results with relatable results produced by the other state-of-the-art dummy generation approaches. Nevertheless, we perform a comparison between VSBDG and the other state-of-the-art solutions by summarizing their handling of known location privacy vulnerabilities and features. Table 5.6 compares the handlings of three location privacy attacks by VSBDG and the other existing

Vulnerability	VSBDG	COSA[3.6]	k-LPP [3.15]	VLBS [3.18]	DLSS[3.14]	V-Cir/ V-grid [3.12]	DLIP [3.19]
Location homogeneity attack	✓	✓	✓	$✓_p$	X	X	X
Map matching attack	$✓_p$	$✓_p$	X	X	X	X	X
Temporal Constraint attack	✓	X	X	X	X	X	X
✓ - Addresses $✓_p$ — Partially addresses X – Fail to address							

Table 5.6: A comparison of the proposed (VSBDG) and the existing dummy approaches based on how various vulnerabilities are addressed.

state-of-the-art dummy algorithms in general. Our proposed VSBDG approach successfully handles all the three vulnerabilities listed in Table 5.6.

The VSBDG algorithm utilizes parcel-based similarity search from COSA[3.6] to seek dummy locations from parcels that are spatially similar to the parcel of an input location. VSBDG and COSA[3.6] are built on real-world geospatial datasets and leverage spatial context for dummy identification. This idea not only helps to generate dummy locations that are resistant to location homogeneity attacks but also makes them less prone to map matching attacks [48]. k-LPP [60] VLBS [44] and DLIP [56] addresses location homogeneity attacks through dummy generation based on semantic diversity. The other three approaches [35, 33] neither use spatial context nor semantic diversity making them prone to location homogeneity attacks. The approaches [60, 44, 33, 35, 25] do not consider the spatial context in dummy generation making them prone to map matching attacks. To our knowledge, VSBDG is the only framework that is capable of addressing temporal constraint attacks by using that semantically balanced location set. Table 5.7. provides a comparison of general features between VSBDG and the other existing dummy approaches. Overall, the VSBDG algorithm not only addresses all the three location privacy attacks (Table 5.6) but also offers the key benefits (Table 5.7) when compared to the existing dummy approaches.

The potential applications of VSBDG include location privacy scenarios where a semantic category is inherently different from the other semantic categories, and one example is

Key benefits	VSBDG	COSA[3.6]	k-LPP [3.15]	VLBS[3.18]	DLSS[3.14]	V-Cir/ V-grid [3.12]	Random[3.10]
Physical dispersion semantic similarity for larger k values.	✓	X	X	X	X	X	X
Do not use location query probability	✓	✓	✓	✓	X	X	✓
Use spatial context in dummy identification process	✓	✓	X	X	X	X	X
Do not submit proxy instead of true location to LBS server	✓	✓	✓	✓	X	✓	✓
Built on real-world geospatial dataset(s)	✓	✓	X	X	X	X	X
✓ - YES X – NO							

Table 5.7: A comparison of benefits addressed by the proposed (VSBDG) and the existing dummy approaches.

residential locations. In the case of residential locations, the VSBDG algorithm is leveraged to protect the real locations of a user whose locations belong to a residence. The smart devices located in a residential location are also potential candidates for location privacy protection offered by VSBDG.

5.7 Summary

The major contributions of this chapter are as follows –

- We introduce a new type of location privacy attack called 'temporal constraint attack' where an adversary can exploit the location semantics from a temporal dimension for eliminating dummies and identifying the true location. In doing so, we provide evidence on how a true location of residential semantic type can be compromised in a temporal constraint attack.

- A novel Voronoi-based semantically balanced dummy generation (VSBDG) approach is proposed to generate dummy locations that are capable of withstanding a temporal constraint attack by an adversary. In general, VSBDG algorithm can achieve location privacy protection regardless of the semantic type of the true location, whether it is residential or non-residential. This is due to the semantically balanced nature of the location set generated by VSBDG.

- One of the major drawbacks of existing dummy location studies is that they do not consider the spatial context of the location, which is not possible unless the technique is built upon real-world geospatial datasets. At best, the current approaches are tested on simple real-world location datasets that contain a collection of point locations. The VSBDG algorithm is built and tested on real-world geospatial datasets such as land parcels and point of interest (POI) locations. The VSBDG algorithm leverages spatial relationships and operations to identify spatially similar dummy locations for a given true location.

- The Voronoi polygons is applied to model and delineate POI influence. We establish an approach that uses a cosine similarity search for finding geographical areas within the city with similar POI influence and performs a parcel-based similarity search to identify the residential dummy location within each similar Voronoi polygon. This allowed us to identify spatially similar residential and POI dummy locations and build semantically balanced location sets that are resistant not only to temporal constraint attacks but also to location homogeneity attacks, location distribution attacks, and map matching attacks.

Chapter 6

Conclusion and Future Research

6.1 Context optimized and spatial aware dummy locations generation framework for location privacy

In chapter 3, we proposed a novel approach to generate dummy locations to preserve location privacy in location-based services. The location involved in location privacy typically extends beyond location coordinates, as the user is concerned with protecting their general whereabouts that can be exploited by an adversary to obtain other sensitive information about the user. The core assumption for the successful implementation of dummy locations for preserving location privacy in location-based services is that dummy locations are similar to the true location. When the dummy locations are indistinguishable from the true location, it makes it difficult for the LBS servers or an adversary to identify the true location of the user. In the scenario where the user location is a health care center located in a large parcel area, it is easy to identify the general whereabouts of the user if dummy locations also lie within or closer to the input large parcel area. An adversary can exploit this information along with other available data such as the type of healthcare center and frequency of visits. In the absence of location privacy protection, the adversary can possibly deduce sensitive information, such as the user's health condition.

Our proposed solution successfully eliminated the possibility of the above location privacy limitation scenario by extending the scope of the dummy location generation beyond the large parcel area. Our approach is spatially aware because of its core dependency on the spatial context of the user's true location and usage of this spatial context in the identification of the dummy location. By implementing the workflow involving an actual county parcel

dataset, we demonstrated the applicability of the solution in real-world scenarios involving location privacy in LBS services.

When dummy locations are sent to LBS servers along with the true location, the LBS servers must compute and communicate the results back for both dummy and true locations. Only the results generated for true location are used by the client. The computation and communication resources spent on producing results for dummy locations are considered waste since the data generated for dummy locations is discarded by the client. Hence, it is important to keep the number of dummy locations to a minimum to reduce the wastage of resources.

Most of the current studies assume that dummies are generated on the client side such as mobile devices. The proposed solution can be implemented in a service-oriented architecture (SOA) setting where a dummy generation service is built on the cloud and could be leveraged by mobile on demand.

6.2 PLP+: An Enhanced Parcel-based Location Privacy Framework using Building Footprint Entropy for Spatially Similar Dummy Locations

Generating dummy locations that are vulnerable to location-homogeneity and map-matching attacks is the most common limitation in a host of existing dummy locations approaches catering to location privacy. The main reason behind such a limitation in the existing solutions is the lack of considering the true spatial context of a location as a criterion for creating dummy locations. The current parcel-based location privacy schemes generate dummy locations that are similar in spatial context, thereby successfully averting the possibility of location-homogeneity attacks. Unfortunately, these techniques do not fully address vulnerability to map-matching attack, since the solutions produce dummy locations in parcels with green areas and natural features.

In chapter 4, we proposed an PLP+ framework, a novel parcel-based dummy locations approach to forge dummy locations that protect against both location homogeneity and

map-matching attacks by an adversary. We implemented enriched similarity search to find spatially similar parcels based on building footprints and road proximity information. The enriched similarity search uses a SPZ - a smaller area within the city - as a query boundary to search for similar parcels instead of entire city. We developed a novel eps estimator algorithm to calculate the search radius (eps) parameter value to be used in DBSCAN clustering for generating SPZs.

We proposed a novel location privacy quantification approach that incorporates building footprint entropy to gauge the total location privacy achieved by a given location set that includes both true and dummy locations. The building footprint entropy metric - FPE - measures the level of uncertainty introduced by dummy locations within a location set while enabling us to evaluate the effectiveness of a map-matching attack by an adversary. Through the FPE metric, we also evaluated and compared the efficacy of an array of dummy approaches applied to create dummies. We implemented the PLP+ framework using real-world geospatial datasets for land parcels, building footprints and street-centerline data for a county in the state of New York. The functionality of enriched similarity search and SPZ were seamlessly integrated into the PLP+ framework, where the only input to the framework is a user's true location and dummy locations as an output.

We used statistical analysis to demonstrate that our enriched similarity search within an SPZ can identify parcels that are no less in similarity quality compared to searching the entire parcel database. We implemented our location privacy quantification method and calculated building footprint metric on a sample input location set to compare the efficacy of dummy locations under the three different similarity search criteria. The dummy locations generated by our enhanced similarity search with building footprints and road proximity within an SPZ consistently resulted in higher FPE measures compared against the other approaches: none of our output dummies are within a green or non-building area. The empirical study demonstrates that our PLP+ framework is adept at producing dummies that keep map-matching attacks at bay.

6.3 A Voronoi-based Semantically Balanced Dummy Generation Framework for Location Privacy

Locations are unique and may differ in characteristics from locations of the other semantic types. One such a case is residential locations that are unique and different in temporal constraints from locations of the POI semantic types. A dummy generation approach ought to address these intrinsic differences to construct robust dummies that are expected to effectively conceal true locations and secure the privacy of the locations.

In chapter 5, we identified and explored a new type of attack called temporal constraint attack, in which an adversary exploits differences in temporal constraints between locations of different semantic types to eliminate dummy locations and single out real locations. We demonstrated how residential locations are susceptible to temporal constraint attacks when an adversary possesses historical request data on dummies submitted for a residential location. The existing techniques, including the ones that are built based on semantic diversity, are prone to temporal constraint attacks because the difference in temporal constraints of a semantic category such as residential location is not taken into account. The key takeaways from this study are summarized below.

- We proposed a novel VSBDG algorithm, which is conducive to generating dummies that can keep temporal constraint attacks at bay.
- The VSBDG algorithm is capable of handling both location homogeneity attacks and map matching attacks because 1. POI influence in a spatial area is modeled using Voronoi polygons and leverages cosine similarity search to find areas within a city that has similar POI influence. 2. Parcel-based similarity search [48] is adopted to construct dummy locations within each Voronoi polygon from parcels that are spatially similar to a legitimate location's parcel.

- Our findings show a high average MDD of 5861.894, 6258.046 meters for residential and POI locations respectively, entailing that the locations are distributed further apart indicating optimized location privacy.

- The results unfold an average PDCS of 0.988 between MDD values of residential and POI locations in location sets with sizes ranging from 4 to 22, thereby demonstrating a strong and scalable semantic balance within an output location set of the VSBDG algorithm suggesting good location privacy protection against a temporal constraint attack.

The temporal constraint attack model discussed in this study, accompanied by the proposed VSBDG algorithm, is specific to snapshot LBS scenarios involving a single real location. On the other hand, a continuous LBS request involves a trajectory with a series of locations [23]; hence, the temporal constraint attack in a continuous LBS scenario should be explored in a future study. Since this investigation is the first study that addresses the concerns of temporal constraint attacks, this work will pave the way for further research into the applicability of time constraint attacks under new scenarios and potential ground-breaking solutions addressing the time constraint attacks.

Bibliography

[1] Esri inc. world imagery. "imagery" [basemap]. https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9, [Accessed: 01-May-2022].

[2] Harris county appraisal district (hcad). tax parcels 2021. city boundary. shapefile. https://hcad.org/pdata/pdata-gis-downloads.html, [Accessed: 01-May-2022].

[3] Gis.ny.gov. 2022. nys civil boundaries. shapefile. [online]. https://gis.ny.gov/gisdata/inventories/details.cfm?DSID=927, [Accessed 1 September 2022].

[4] Nyc opendata 2022. building footprints. shapefile. [online]. https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh, [Accessed 1 September 2022].

[5] Nyc opendata 2022. department of finance digital tax map. shapefile. https://data.cityofnewyork.us/Housing-Development/Department-of-Finance-Digital-Tax-Map/smk3-tmxj, [Accessed 1 September 2022].

[6] Nyc opendata 2022. nyc street centerline (cscl). shapefile. [online]. https://data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b, [Accessed 1 September 2022].

[7] Nyc opendata 2022. points of interest. shapefile. [online]. https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj, [Accessed 16 November 2022].

[8] F. Akdag, C. Eick, P. Amalaman, and A. Tadakaluru. A framework for discriminative polygonal place scoping. 11 2013.

[9] A. S. Alyousef, K. Srinivasan, M. S. Alrahhal, M. Alshammari, and M. Al-Akhras. Preserving location privacy in the iot against advanced attacks using deep learning. *International Journal of Advanced Computer Science and Applications*, 13(1), 2022.

[10] S. Chen and H. Shen. Semantic-aware dummy selection for location privacy preservation. In *2016 IEEE Trustcom/BigDataSE/ISPA*, pages 752–759, 2016.

[11] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar. Preserving user location privacy in mobile data management infrastructures. In G. Danezis and P. Golle, editors, *Privacy Enhancing Technologies*, pages 393–412, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[12] L. Dunne, E. Bamford, and D. Taylor. Quantifying remoteness-a gis approach. In *The 11 th Annual Colloquium of the Spatial Information Research Centre*, pages 13–15, 1999.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[14] D. G. Evans and S. M. Jones. Detecting voronoi (area-of-influence) polygons. *Mathematical geology*, 19:523–537, 1987.

[15] F. Fei, S. Li, H. Dai, C. Hu, W. Dou, and Q. Ni. A k-anonymity based schema for location privacy preservation. *IEEE Transactions on Sustainable Computing*, 4(2):156–167, 2019.

[16] E. Frejinger. Route choice analysis: data, models, algorithms and applications. 01 2008.

[17] S. Gambs, M.-O. Killijian, and M. Núñez del Prado Cortez. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614, 2014. Special Issue on Theory and Applications in Parallel and Distributed Computing Systems.

[18] 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*, MobiSys '03, page 31–42, New York, NY, USA, 2003. Association for Computing Machinery.

[19] T. Hara, A. Suzuki, M. Iwata, Y. Arase, and X. Xie. Dummy-based user location anonymization under real-world constraints. *IEEE Access*, 4:673–687, 2016.

[20] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[21] H. N. S. S. Jagarlapudi, S. Lim, J. Chae, G. S. Choi, and C. Pu. Drone helps privacy: Sky caching assisted k-anonymity in spatial querying. *IEEE Systems Journal*, 16(4):6360–6370, 2022.

[22] C. S. Jensen, H. Lu, and M. L. Yiu. Location privacy techniques in client-server architectures. In *Privacy in Location-Based Applications*, 2009.

[23] H. Jiang, J. Li, P. Zhao, F. Zeng, Z. Xiao, and A. Iyengar. Location privacy-preserving mechanisms in location-based services: A comprehensive survey. *ACM Comput. Surv.*, 54(1), jan 2021.

[24] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias. Preventing location-based identity inference in anonymous spatial queries. *IEEE transactions on knowledge and data engineering*, 19(12):1719–1733, 2007.

[25] H. Kido, Y. Yanagisawa, and T. Satoh. Protection of location privacy using dummies for location-based services. In *21st International Conference on Data Engineering Workshops (ICDEW'05)*, pages 1248–1248, 2005.

[26] F. Koufogiannis and G. J. Pappas. Location-dependent privacy. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7586–7591. IEEE, 2016.

[27] O. Kounadi, T. J. Lampoltshammer, M. Leitner, and T. Heistracher. Accuracy and privacy aspects in free online reverse geocoding services. *Cartography and Geographic Information Science*, 40(2):140–153, 2013.

[28] B. Liu, W. Zhou, T. Zhu, L. Gao, and Y. Xiang. Location privacy and its applications: A systematic study. *IEEE Access*, 6:17606–17624, 2018.

[29] H. Lu, C. S. Jensen, and M. L. Yiu. Pad: Privacy-area aware, dummy-based location privacy in mobile services. In *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '08, page 16–23, New York, NY, USA, 2008. Association for Computing Machinery.

[30] R. D. N. C. M. Machado$_2$020. $Privlbs : Preservingprivacyinlocation - basedservices. JournalofInformationandDataManagement, 10(2) : 81–96, Feb.2020.$

[31] A. B. Manju and S. Sumathy. Dispersed dummy selection approach for location-based services to preempt user-profiling. *Concurrency and Computation: Practice and Experience*, 33(20):e6361, 2021.

[32] L. Ni, F. Tian, Q. Ni, Y. Yan, and J. Zhang. An anonymous entropy-based location privacy protection scheme in mobile social networks. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):1–19, 2019.

[33] N. Nisha, I. Natgunanathan, and Y. Xiang. An enhanced location scattering based privacy protection scheme. *IEEE Access*, 10:21250–21263, 2022.

[34] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li. Achieving k-anonymity in privacy-aware location-based services. In *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pages 754–762, 2014.

[35] B. Niu, Z. Zhang, X. Li, and H. Li. Privacy-area aware dummy generation algorithms for location-based services. In *2014 IEEE International Conference on Communications (ICC)*, pages 957–962, 2014.

[36] D. Parmar and U. P. Rao. Dummy generation-based privacy preservation for location-based services. In *Proceedings of the 21st International Conference on Distributed Computing and Networking*, ICDCN 2020, New York, NY, USA, 2020. Association for Computing Machinery.

[37] F. Qiu, H. Sridharan, and Y. Chun. Spatial autoregressive model for population estimation at the census block level using lidar-derived building volume information. *Cartography and Geographic Information Science*, 37(3):239–257, 2010.

[38] C. E. I. Redlands. Esri inc (2021) arcgis pro (version 2.8.2). software. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.

[39] B. M. RStudio, PBC. Rstudio team (2021). rstudio: Integrated development environment for r. http://www.rstudio.com/.

[40] P. M. Schirmer, M. A. Van Eggermond, and K. W. Axhausen. The role of location in residential location choice models: a review of literature. *Journal of Transport and Land Use*, 7(2):3–21, 2014.

[41] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[42] V. Sharma and C.-C. Shen. Evaluation of an entropy-based k-anonymity model for location based services. In *2015 International Conference on Computing, Networking and Communications (ICNC)*, pages 374–378. IEEE, 2015.

[43] X. Shi, J. Zhang, and Y. Gong. A dummy location generation algorithm based on the semantic quantification of location. In *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 172–176, 2021.

[44] X. Shi, J. Zhang, and Y. Gong. A dummy location generation algorithm based on the semantic quantification of location. In *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 172–176, 2021.

[45] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. Quantifying location privacy. In *2011 IEEE symposium on security and privacy*, pages 247–262. IEEE, 2011.

[46] S. Steiniger, T. Lange, D. Burghardt, and R. Weibel. An approach for the classification of urban building structures based on discriminant analysis techniques. *T. GIS*, 12:31–59, 02 2008.

[47] G. Sun, V. Chang, M. Ramachandran, Z. Sun, G. Li, H. Yu, and D. Liao. Efficient location privacy algorithm for internet of things (iot) services and applications. *Journal of Network and Computer Applications*, 89:3–13, 2017.

[48] A. Tadakaluru. Context optimized and spatial aware dummy locations generation framework for location privacy. *Journal of Geovisualization and Spatial Analysis*, 6, 2022.

[49] A. Tadakaluru and X. Qin. A voronoi-based semantically balanced dummy generation framework for location privacy. *Analytics*, 2(1):246–264, 2023.

[50] S. Van Dongen and A. J. Enright. Metric distances derived from cosine similarity and pearson and spearman correlations. *arXiv preprint arXiv:1208.3145*, 2012.

[51] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel. A classification of location privacy attacks and approaches. *Personal and ubiquitous computing*, 18:163–175, 2014.

[52] Z. Wu, G. Li, S. Shen, X. Lian, E. Chen, and G. Xu. Constructing dummy query sequences to protect location privacy and query privacy in location-based services. *World Wide Web*, 24, 01 2021.

[53] X. Xu, H. Chen, and L. Xie. A location privacy preservation method based on dummy locations in internet of vehicles. *Applied Sciences*, 11(10):4594, 2021.

[54] B. Yu, H. Liu, J. Wu, Y. Hu, and L. Zhang. Automated derivation of urban building density information using airborne lidar data and object-based method. *Landscape and Urban Planning*, 98(3):210–219, 2010. Climate Change and Spatial Planning.

[55] P. A. Zandbergen. A comparison of address point, parcel and street geocoding techniques. *Computers, environment and urban systems*, 32(3):214–232, 2008.

[56] A. Zhang and X. Li. Research on privacy protection of dummy location interference for location-based service location. *International Journal of Distributed Sensor Networks*, 18(9):15501329221125111, 2022.

[57] C. Zhang, H. Liang, K. Wang, and J. Sun. Personalized trip recommendation with poi availability and uncertain traveling time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 911–920, 2015.

[58] P. Zhang, C. Hu, D. Chen, H. Li, and Q. Li. Shiftroute: Achieving location privacy for map services on smartphones. *IEEE Transactions on Vehicular Technology*, 67(5):4527–4538, 2018.

[59] S. Zhang, M. Li, W. Liang, V. K. A. Sandor, and X. Li. A survey of dummy-based location privacy protection techniques for location-based services. *Sensors*, 22(16), 2022.

[60] Y.-B. Zhang, Q. yu Zhang, Z.-Y. Li, Y. Yan, and M. yi Zhang. A k-anonymous location privacy protection method of dummy based on geographical semantics. *Int. J. Netw. Secur.*, 21:937–946, 2019.