Implementing & Comparing Noise Addition by Satisfying Geo-indistinguishability & Aggregating Data for k-Anonymity

Emin Berke Ay, Yiğit Ekin, Cenk Duran, Arda Eren, Harun Can Surav, Gamze Elif Çenesiz, Can Önal, Onuralp Avcı

İhsan Doğramacı Bilkent University

***Abstract* - Location-based services (LBS) rely on a device's precise location to offer users a range of services and information, including social networking, advertising, search, and navigation support. However, the gathering and use of location data poses privacy issues because it may reveal sensitive information about a person's daily activities and routines and may be accessible to unauthorized parties. Approaches like geo-indistinguishability (adding controlled noise to data) or satisfying k-anonymity in data sharing have been suggested as solutions to these problems. This study's objective is to put these methods into practice and evaluate their utility and respect for privacy. The evaluation of each approach's level of privacy protection and its effect on the usefulness of the location data will be included in the comparison. The comparison will be carried out via simulations or by putting the strategies into practice in the real world and gathering user data for analysis. The findings of this study will shed light on how well various strategies protect user privacy while keeping the value of location data for LBS.**

Introduction

Technology known as location-based services (LBS) makes use of a device's physical location to offer consumers a variety of services and information. These services may include, among others, location-based social networking, location-based search, and assistance with navigation.

Users' location data is gathered by LBS providers in order to provide these services using a variety of channels, including GPS, Wi-Fi, and cellular data. To pinpoint the user's position and deliver the necessary services or information, this data is used. However, user privacy issues are also brought up by the gathering and usage of location data.

Location information might be sensitive because it can reveal details about a person's daily habits and movements. Unauthorized parties may be able to access and use this data, which could result in privacy violations. Because of this, it is critical for LBS providers to ensure user privacy while also delivering on what customers want from them.

In the literature, a number of strategies have been put out to overcome these privacy issues. Adding controlled noise to the location data is one of the strategies. To make it more challenging to pinpoint the precise position of the user, this requires introducing a small bit of random noise to the data. This strategy, known as geo-indistinguishability, tries to maintain the location data's usefulness and accuracy (i.e., its capacity to offer location-based services) while still safeguarding the user's privacy.

Geo-indistinguishability can be done in a variety of ways, including by employing techniques like differential privacy to add noise to the data or by adding random noise to the location data within a specific radius. The objective is to make the data sufficiently noisy such that it is challenging to pinpoint the user's precise location while maintaining the data's value for location-based services.

Keeping location data private also involves satisfying k-anonymity requirements for data exchange. This entails making sure that no single user's location data in a dataset containing location data from several users can be identified. This is accomplished by combining the location information from at least k users, sharing the resulting aggregated information

rather than the discrete data points. This strategy aims to make it challenging to pinpoint any specific user's location data in the common dataset.

k-Anonymity can be applied in a variety of ways, for example by combining the location information of many users who are in close proximity to one another or by utilizing generalization to lessen the granularity of the data. While maintaining the data's value for location-based services, the objective is to make it challenging to identify any one user's data in the common dataset.

Implementing and contrasting these two strategies can help to determine how well they protect user privacy while retaining the usefulness of location data for location-based services. Comparing different approaches' levels of privacy protection and their effects on the usefulness of location data could be part of this analysis. This could be accomplished through simulations or by putting the strategies into practice in the real world and gathering user data for analysis.

For instance, an implementation could contrast the level of privacy protection offered by satisfying k-anonymity in data sharing with that provided by adding controlled noise to location data. The implementation could also evaluate how well a person's precise location can be pinpointed using the data from each method, as well as the effect on how useful the data is for offering location-based services.

Therefore, it is essential for LBS providers to maintain user privacy while still offering location-based services. Two efficient solutions we chose to implement in order to maintain user’s privacy while providing LBS are message perturbation engine and local k-search.

I. Problem Definition

In recent years, location-based services (LBS) have grown in popularity. These services give a variety of benefits to customers by utilizing location data. LBS have ingrained themselves into our daily lives and are used for everything from navigation to social networking to location-based advertising. However, there are possible threats to privacy that must be taken into account, just like with any technology that includes the collecting and use of personal data.

The possibility for exploitation or misuse of location data is one of the main issues with LBS. Without the user's knowledge or consent, location data may be gathered and transmitted, which may violate their privacy and put them at risk for crimes including identity theft and financial fraud. If location data is utilized to stalk or harass someone, there is also a chance of bodily injury.

It is quite critical for LBS providers to be open about how they gather and utilize users' data and to provide consumers control over their location information in order to allay these worries. This can entail giving consumers alternatives to opt out of data collection or restrict the kinds of data that are collected, as well as being transparent and thorough about how location data is gathered, utilized, and shared.

Users must also take precautions to preserve their privacy and be aware of the potential risks involved with sharing their location data. This can entail exercising caution when disclosing information and making sure they are aware of how their location data is being utilized. People can take preventative steps to safeguard themselves and their privacy by being aware of the potential risks associated with the exploitation of location data.

Thus, it is safe to say that LBS has a lot of potential benefits, but it's crucial to carefully weigh the dangers and take precautions to safeguard privacy. We can make sure that these services continue to be valuable to users while also preserving their privacy by balancing the advantages and hazards of LBS.

II. Proposed Solutions

A. Adding Noise by Satisfying Geo-Indistinguishability

Geo-indistinguishability is a privacy concept that aims to protect the privacy of individuals in a dataset by adding noise to their location data. This is done in

order to make it difficult for an attacker to determine the exact location of an individual, even if they have access to the location data.

To achieve geo-indistinguishability, the location data of individuals is modified by adding noise to it in a controlled way. The amount of noise added is typically chosen such that it is sufficient to make it difficult for an attacker to determine the exact location of an individual, but not so much that the location data becomes completely inaccurate or useless.

By adding noise to location data in this way, the privacy of individuals is preserved while still allowing for the analysis of general trends and patterns in the data.

B. Aggregating Data by Satisfying k-Anonymity

k-Anonymity is a privacy concept that aims to protect the privacy of individuals in a dataset by ensuring that no individual can be identified by their specific attributes. In the context of location privacy, k-anonymity can be used to protect the privacy of individuals by aggregating their location data such that no individual can be identified by their specific location.

To achieve k-anonymity, the location data of individuals can be grouped into "cells" or "bounded boxes", such that each bounded box contains at least k individuals. This means that for any given bounded box, there are at least k individuals with similar location data, making it difficult to identify a specific individual based on their location alone. By aggregating the location data in this way, the privacy of individuals is preserved while still allowing for the analysis of general trends and patterns in the data.

C. Expected Output for the Solutions

One of the main differences between k-anonymity and geo-indistinguishability is that k-anonymity is generally expected to provide better privacy protection, while geo-indistinguishability is expected to have better utility. This means that k-anonymity may be more effective at protecting the privacy of individuals, but it may also result in more loss of information and a lower level of accuracy in the data. On the other hand, geo-indistinguishability may provide a lower level of privacy protection, but it may also result in less loss of information and a higher level of accuracy in the data.

It is important to note that both k-anonymity and geo-indistinguishability have trade-offs between privacy and utility, and the best approach will depend on the specific requirements and constraints of the situation. In some cases, it may be necessary to use a combination of these techniques in order to balance privacy and utility in an appropriate way.

III. Research Methodology

A. Literature Search

The study "Location Privacy in Mobile Systems: A Personalized Anonymization Model" by Gedik and Ling Liu examines the problem of location privacy in mobile systems and suggests a personalized anonymization model as a remedy. The main concern of the essay is how to protect people's privacy while still allowing for the collecting and use of location data for a variety of reasons, including location-based services and advertising.

Authors contend that conventional methods for preserving location privacy, such as controlled noise augmentation or k-anonymity in data exchange, have limits and might not always be adequate. They suggest a tailored anonymization model as an alternate strategy that tries to strike a more flexible and successful balance between privacy and value.

The personalized anonymization model is based on the idea of personalized k-anonymity, which entails aggregating the location data of at least k users and then personalizing each user's controlled noise addition to the data. The user's privacy options and the sensitivity of the location data determine how much noise is added to the data. While still allowing for the collection and use of location data for various purposes, the objective is to protect the user's privacy.

The authors give a thorough explanation of the personalized anonymization model and all of its many parts, including the utility model, the noise

addition model, and the privacy preference model. In order to show how well the model protects location privacy while enabling the use of location data for location-based services, they also give a case study.

Gedik and Ling Liu's paper adds to the field of location privacy by presenting a promising approach for addressing the trade-off between privacy and utility in the context of mobile systems [1].

Additionally, the paper "Geo-indistinguishability: Differential Privacy for Location-Based Systems" by Andrés, Bordenabe, Chatzikokolakis, and Palamidessi, on the other hand, discusses the issue of preserving individuals' privacy in location-based systems and proposes a new approach called geo-indistinguishability.

Traditional approaches to safeguarding location privacy, such as satisfying k-anonymity in data exchange, the authors suggest, have limits and may not be sufficient in certain instances. They suggest geo-indistinguishability as a more flexible and effective solution to balancing the trade-off between privacy and utility.

Geo-indistinguishability is built on the concept of differential privacy, which includes adding controlled noise to data to make it impossible to pinpoint an individual's specific location while retaining the data's utility for various purposes. The quantity of noise added to the data is determined by the individual's privacy preferences and the sensitivity of the location data.

The authors describe the geo-indistinguishability strategy and its different components, such as the privacy parameter and the noise addition process, in detail. They also give a case study to demonstrate the approach's success in ensuring location privacy while permitting the usage of location data for location-based services.

The study by Andrés, Bordenabe, Chatzikokolakis, and Palamidessi makes a valuable contribution to the topic of location privacy and offers a viable strategy for resolving the conflict between privacy and utility in location-based systems [2].

IV. Result & Analysis

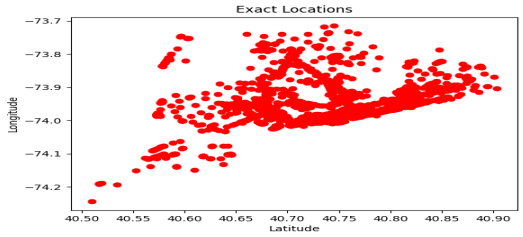

Figure 1. *Distribution of Wi-Fi Hotspot Locations Across New York.*

Figure 1 above shows the actual location of the Wi-Fi Hotspot Locations across New York. We have utilized our implementations on this dataset we have found [3].

A. Results on Geo-indistinguishability Algorithm Implementation

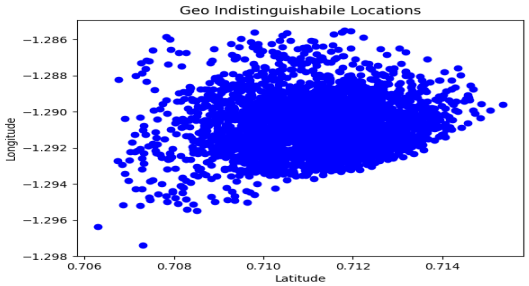

Figure 2. *Geo Indistinguishable Locations of Wi-Fi Hotspot Across New York for ε = 0.1.*

The epsilon value is the amount of noise or uncertainty added to location data to protect privacy. The epsilon value can be changed to achieve the desired level of privacy protection. A higher epsilon value indicates a greater amount of noise, which can provide better privacy protection but may reduce the data's utility for the intended purpose. A lower epsilon value indicates less noise, which can preserve data utility while providing less privacy protection.

The epsilon value chosen will be determined by the specific context and the trade-offs involved. In a situation where privacy is a high priority, for example, a higher epsilon value may be appropriate to provide stronger privacy protection. In contrast, if the utility of the data is important, a lower epsilon

value may be more appropriate in order to preserve the usefulness of the data; thus, we chose an appropriate epsilon value for our research purposes.

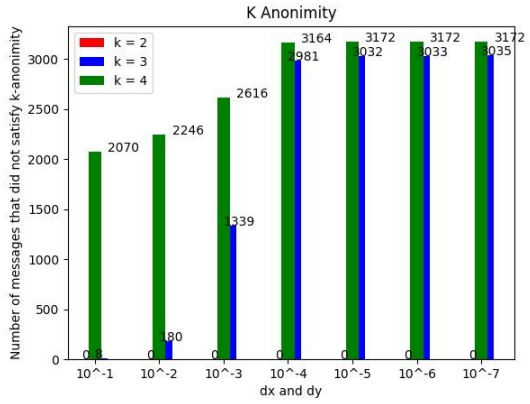

Figure 3. *Results for k-anonymity for all k-values combined.*

Figure 3 above shows that there is no location that is not bounded by a bounded box when we set k value as 2. This result shows that Wi-Fi hotspots are so close to each other that there are at least 2 Wi-Fi next to each other.

On the other hand, the number of locations that could not be bounded by aggregation of data are increasing as we increase k value. At this point, we can conclude that maintaining a higher level of privacy while preserving utility is not fully possible. In order to satisfy higher values of k values we need to increase the size of bounded boxes which results in more inaccurate results.

At this point, what is important to consider is the level of privacy of the data and how much we are willing to sacrifice accuracy of the data.

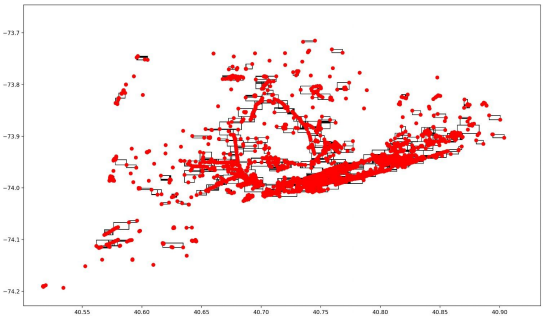

Figure 4. *Results of the implemented k-anonymity algorithm for k=3.*

Figure 4 shows that with k value equal to 3, we could achieve an adequate level of privacy as an adequate percentage of locations are bounded by a bounding box. However, deciding on the level of privacy would change based on context and in reality it would not be enough.

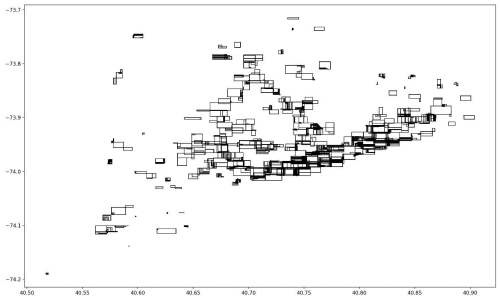

Figure 5. *Results of the implemented k-anonymity algorithm for k=3.*

Geo-indistinguishability offered more utility, or usefulness, for the intended goal, whereas k-Anonymity produced a higher level of privacy protection. However, the specific context and the associated trade-offs will determine the relative usefulness of these strategies. In order to strike the best possible balance between privacy protection and data utility, it could be necessary to take into account both strategies.

V. Conclusion & Future Work

Looking at the previous results from research and concluded results both from k-anonymity and geo-indistinguishability, the level of privacy and how to preserve it is understood. K-anonymity preserves it by ensuring that unless there are at least k other records in the collection with the same values for the sensitive attributes, no records can be linked to an individual. Geo-indistinguishability preserves it by adding noise so that it is more difficult for an attacker to detect the exact location.

In many different applications, including the publication of medical records, the sharing of

location data, and the release of census data, k-anonymity and geo-indistinguishability have both been widely employed. However, it has also been demonstrated that these methods have drawbacks and potential flaws, such as the potential for re-identification attacks or the loss of utility as a result of the additional noise.

For the future, the first thing to consider is to see the applicability of these methods in a wider perspective and other real possible cases.

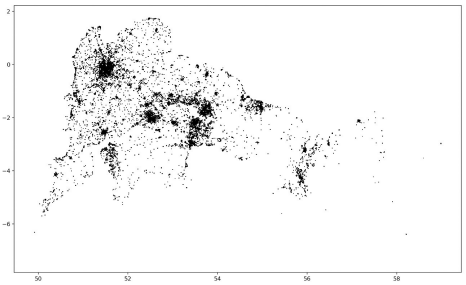

Figure 6. *Results of all of the pubs in the United Kingdom.*

Figure 6 above shows the result of k-anonymity applied to all the pubs in the United Kingdom. It can be seen that the volume of dots around the plot is reminiscent of the real map of the United Kingdom. Goal is to apply both methods in different areas with different datasets and discuss the results as seen in the figure.

References

[1] B. Gedik and L. Liu. Location privacy in mobile systems: A personalized anonymization model. In Proc. of ICDCS, pages 620–629. IEEE, 2005.

[2] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability," *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13*, 2013.

[3] C. of N. York, "NYC Wi-Fi hotspot locations," *Kaggle*, 01-Nov-2018. [Online]. Available: https://www.kaggle.com/datasets/new-york-city/nyc-wi-fi-hotspot-locations.