PANDA: Policy-aware Location Privacy for Epidemic Surveillance

Yang Cao
Kyoto University, Japan
yang@i.kyoto-u.ac.jp

Shun Takagi *
Kyoto University, Japan
takagi.shun.45a@st.kyoto-u.ac.jp

Yonghui Xiao
Emory University, USA
yohuxiao@gmail.com

Li Xiong
Emory University, USA
lxiong@emory.edu

Masatoshi Yoshikawa
Kyoto University, Japan
yoshikawa@i.kyoto-u.ac.jp

ABSTRACT

In this demonstration, we present a privacy-preserving epidemic surveillance system. Recently, many countries that suffer from COVID-19 crises attempt to access citizen's location data to eliminate the outbreak. However, it raises privacy concerns and may open the doors to more invasive forms of surveillance in the name of public health. It also brings a challenge for privacy protection techniques: how can we leverage people's mobile data to help combat the pandemic without scarifying location privacy. We demonstrate that we can achieve this by implementing policy-based location privacy for epidemic surveillance. Our system has three primary functions for epidemic surveillance: people flow monitoring, epidemic analysis, and contact tracing. We provide an interactive tool allowing the attendees to explore and examine the usability of our system: (1) the utility of location monitor and disease transmission model estimation, (2) the procedure of contact tracing in our systems, and (3) the privacy-utility trade-offs w.r.t. different policy graphs. The attendees will find that we can have the high usability for epidemic surveillance while preserving location privacy.

PVLDB Reference Format:
Yang Cao, Shun Takagi, Yonghui Xiao, Li Xiong, Masatoshi Yoshikawa. PANDA: Policy-aware Location Privacy for Epidemic Surveillance. *PVLDB*, 13(12): 3001-3004, 2020.
DOI: https://doi.org/10.14778/3415478.3415529

1. INTRODUCTION

We are fighting with the pandemic of COVID-19 disease. To prevent the spread of such a highly contagious virus, the crucial information that we need for epidemic surveillance is people's location history. Recently, many countries that suffer from coronavirus crises attempt to access citizen's location data to eliminate the outbreak. The US pumped 500 million dollars of emergency funding into the CDC for building a surveillance and data collection system [1] and discussed with Facebook and Google for sharing people's location data to combat the coronavirus. In South Korea, the government created a public map of coronavirus patients using location data from telecom and credit card companies [15]. Italy's telecom companies are sharing location data with health authorities to check whether people are remaining at home [2]. China's giant tech companies provide a "health code" service to certify a user's health status based on her travel history, which are collected by the cellphone apps [9]. Although these special measures of personal data collection for public health emergency may be temporary and under stringent government regulation, it raises concerns over privacy, and people are worried that it may open the doors to surveillance activities in the name of public health. It also brings a challenge for location privacy protection techniques: how can we utilize people's mobile data to help combat the pandemic without sacrificing our location privacy.

Location privacy has been extensively studied in the literature [17]. However, the state-of-the-art location privacy models are not flexible enough to balance the individual privacy and public interest in an emergency as we are witnessing in the COVID-19 crisis. The early studies on location privacy were extending k-anonymity [19] and were flexible enough to be adapted to different scenarios such as personalized location anonymity [11]. But, the recent studies revealed that k-anonymity might not be rigorous enough since they suffer many realistic attacks [14, 16] when the adversary has background knowledge about the original dataset. The recent state-of-the-art location privacy models[3, 22, 21, 20, 23, 5, 6, 7] were extended from differential privacy (DP) [10] to private location release since DP is considered a rigorous privacy notion. Although these DP-based location privacy models are rigorously defined, they are not flexible and customizable for different scenarios with various requirements on privacy-utility trade-off. Taking an example of Geo-Indistinguishability[3], which is the first and influential DP-based location privacy metric, the strength of protection is solely controlled by a single parameter ϵ to achieve

*Yang and Shun contributed equally to this work.

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3415478.3415529

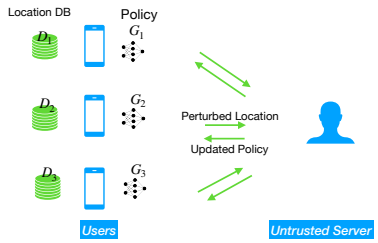

Figure 1: Private location sharing with Customizable Policy.

indistinguishability among all possible locations. It is hard to make a good privacy-utility trade-off using this single ϵ in a complicated setting.

We should have a flexible and rigorous location privacy model that enables customizable location privacy policy and can be used to define which locations are sensitive, which are not. The policy should be adjustable for different people, at different time, and with different purposes. For instance, under the emergency of COVID-19, a location privacy policy for contact tracing could be "*allowing to disclose a user's true locations of the past two weeks if she is a diagnosed coronavirus patient; otherwise, ensuring indistinguishability of the user's location*"; if the patient's location trace and the time period are confirmed, we can dynamically update the location privacy policy for each person to find all contacts of the confirmed patient. A policy for all other people could be "*allowing the app to access a user's true locations if she has been staying in the same location at the same period as an infected user; otherwise, ensuring indistinguishability of the user's location*". In this way, we can guarantee both full usability of contact tracing and reasonable privacy.

In this demonstration, we present PANDA, i.e., Policy-aware privAcy preserviNg epiDemic surveillAnce, which implements our recently proposed Policy Graph-based Location Privacy (PGLP) [4] and mechanisms for epidemic surveillance. Our system is featured by the customizable location privacy policy graph, which provides a new dimension to tune utility-privacy trade-off.

In our recent study [4], we proposed a formal representation of location privacy policy using a graph, which is inspired by a statistical privacy notion of Blowfish privacy [12]. In our setting of private location release, a privacy policy graph (such as the ones shown in Fig.2) includes all possible locations that need to be protected as its nodes, and the edges indicate indistinguishability between two possible locations. A user could arbitrarily customize the location policy graph according to her privacy and utility requirement and enjoy plausible deniability regarding her whereabouts. The definition of PGLP can be seen as a generalization of two influential DP-based location privacy models: Geo-Indistinguishability [3] and Location Set Privacy [22]. Under appropriate configuration of policy graphs, an algorithm satisfying PGLP w.r.t. the policy graphs could also satisfy Geo-Indistinguishability or Location Set Privacy. In [4], we also designed mechanisms for PGLP by adapting the Laplace mechanism and Planar Isotropic Mechanism (PIM) (i.e., an optimal mechanism for Location Set Privacy [22]) w.r.t. a given location policy graph.

However, it is not trivial to directly apply PGLP for a location-based application such as epidemic surveillance due to the following reasons. First, it is not clear how to design a proper policy graph with reasonable privacy and functional utility. Second, when there are multiple choices for location privacy policies, we lack a tool to explore and compare the utility gain w.r.t. different location privacy policies. Third, it is difficult for users to understand the privacy implications (i.e., the privacy risks) of a given location privacy policy.

1.1 Contributions

To address the above issues and motivated by the significant impact of the pandemic of COVID-19 in the world, we demonstrate a policy-based location privacy-preserving epidemic surveillance system. Our contributions are summarized below.

First, we design an epidemic surveillance system with three primary functions: *location monitoring, epidemic analysis*, and *contact tracing*. The scenario is shown in Fig.1, where users locally maintain location databases (e.g., all locations in the past two weeks) and share perturbed locations satisfying PGLP w.r.t. a specific policy graph with a semi-honest server. The policy graph essentially acts as an information filter to control what could be shared and what should not be shared.

Second, we demonstrate three policy graphs with distinct granularity that are appropriate for different functions in the epidemic surveillance. Specifically, we visualize the utility gain or loss between different policy graphs. It turns out that no policy is the best for all. The attendees of the conference can find that it is possible to have the full functionality of epidemic surveillance while preserving location privacy.

Third, we visualize the trade-off between privacy and utility. Although we can specify a policy graph that enables the full usability of the system, it is not clear what is the privacy implication given a policy graph. The policy graph itself could be semantically meaningful, but we lack a quantitative measurement. We provide empirical privacy metrics as the adversary's successful inference [18] with an interactive tool. The attendees can randomly generate a policy graph to explore its effect on the privacy-utility trade-off. The codes are available in github [1] [2].

2. BACKGROUND

2.1 Location Policy Graph

Inspired by Blowfish privacy[12], we use an undirected graph to define which location should be protected and which could not, i.e., location privacy policies. The nodes are secrets and the edges are the required indistinguishability, which indicate an attacker should not be able to distinguish the input secrets by observing the perturbed output. In our setting, we treat possible locations as nodes, and the indistinguishability between the locations as edges.

Definition 1. (Location Policy Graph) *A location policy graph is defined as an undirected graph* $\mathcal{G}=(\mathcal{S},\mathcal{E})$*, where* $\mathcal{S}$ *denotes all the locations (nodes) and* $\mathcal{E}$ *represents indistinguishability (edges) between these locations.*

Definition 2. (Distance in Policy Graph) *We define the distance between two nodes* $\boldsymbol{s}_i$ *and* $\boldsymbol{s}_j$ *in a policy graph as the length of the shortest path (i.e., number of hops) between them, denoted by* $d_{\mathcal{G}}(\boldsymbol{s}_i,\boldsymbol{s}_j)$.

In DP, the two possible database instances with or without a user's data are called *neighboring databases*. In our

[1]https://github.com/emory-aims/pglp
[2]https://github.com/tkgsn/covid-demo

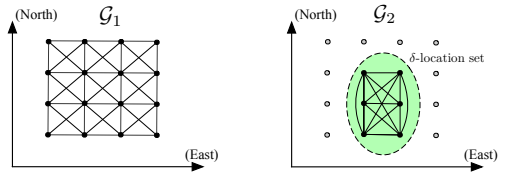

Figure 2: Two examples of location policy graphs.

location privacy setting, we define neighbors as two nodes with an edge in a policy graph.

DEFINITION 3. (Neighbors) *The neighbors of location $\boldsymbol{s}$, denoted by $\mathcal{N}(\boldsymbol{s})$, is the set of nodes having an edge with $\boldsymbol{s}$, i.e., $\mathcal{N}(\boldsymbol{s}) = \{\boldsymbol{s}' | d_{\mathcal{G}}(\boldsymbol{s}, \boldsymbol{s}') = 1, \boldsymbol{s}' \in \mathcal{S}\}$.*

In our system, we assume that the location policy graph is determined by the server for the purposed utility maximization. The user has the right to reject a privacy policy so that no location will be released. By making the policy graph public, the system has a high level of transparency.

2.2 Privacy Metrics

We now formalize Policy Graph-based Location Privacy (PGLP), which guarantees indistinguishability for every pair of neighbors (i.e., for each edge) in a location policy graph.

DEFINITION 4. ($\{\epsilon, \mathcal{G}\}$-location privacy) *A randomized algorithm $\mathcal{A}$ satisfies $\{\epsilon, \mathcal{G}\}$-location privacy iff for all $\boldsymbol{z} \subseteq Range(\mathcal{A})$ and for all pairs of neighbors $\boldsymbol{s}$ and $\boldsymbol{s}'$ in $\mathcal{G}$, we have $\frac{\Pr(\mathcal{A}(s)=z)}{\Pr(\mathcal{A}(s')=z)} \leq e^{\epsilon}$.*

In PGLP, privacy is rigorously guaranteed through ensuring indistinguishability between any two neighboring locations specified by a customizable location policy graph. The user enjoys plausible deniability about her whereabout.

LEMMA 1. *An algorithm $\mathcal{A}$ satisfies $\{\epsilon, \mathcal{G}\}$-location privacy, iff any two neighbors $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{G}$ are $\epsilon \cdot d_{\mathcal{G}}(\boldsymbol{s}_i, \boldsymbol{s}_j)$-indistinguishable.*

Lemma 1 indicates that, if there is a path between two nodes (locations) $\mathbf{s}_i, \mathbf{s}_j$ in the policy graph, the corresponding indistinguishability is required at a certain degree; if two nodes are not connected (i.e., $d_{\mathcal{G}}(\mathbf{s}_i, \mathbf{s}_j) = \infty$), the indistinguishability is not required by the policy. As an extreme case, if a node is not connected with any other nodes, it allows to release it without any perturbation.

2.2.1 Comparison with Other Location Privacy

We analyze the relation between PGLP and two influential DP-based location privacy models, i.e., Geo-Indistinguishability[3] and δ-Location Set Privacy [22]. We show that PGLP implies each of them under proper configurations of location policy graphs.

Geo-Indistinguishability [3] guarantees a level of indistinguishability between two locations $\mathbf{s}_i$ and $\mathbf{s}_j$ that is scaled with their Euclidean distance, i.e., $\epsilon \cdot d_E(\mathbf{s}_i, \mathbf{s}_j)$-indistinguishability, where $d_E(\cdot, \cdot)$ denotes Euclidean distance. Let $\mathcal{G}_1$ be a location policy graph that every location has edges with its closest eight locations on the map as shown in Fig.2 (left). We can derive the following theorem by the fact of $d_{\mathcal{G}}(\mathbf{s}_i, \mathbf{s}_j) \leq d_E(\mathbf{s}_i, \mathbf{s}_j)$ for any $\mathbf{s}_i, \mathbf{s}_j \in \mathcal{G}_1$ and Lemma 1. Note that the unit length used in Geo-Ind scales the level of indistinguishability. We assume that, for any neighbors $\mathbf{s}$ and $\mathbf{s}'$, the unit length used in Geo-Ind makes $d_E(\mathbf{s}, \mathbf{s}') \geq 1$.

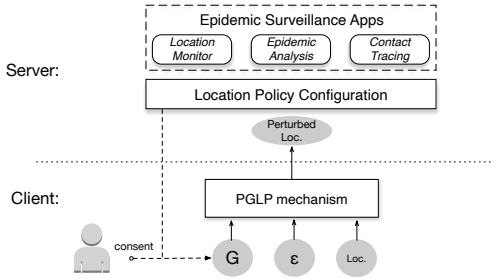

Figure 3: System Overview.

THEOREM 1. *An algorithm satisfying $\{\epsilon, \mathcal{G}_1\}$-location privacy also achieves ϵ-Geo-Indistinguishability.*

δ-Location Set Privacy [22] extends differential privacy on a subset of possible locations, which is assumed as adversarial knowledge. δ-Location Set Privacy ensures indistinguishability among any two locations in the δ-location set. Let $\mathcal{G}_2$ be a location policy that is a complete graph among locations in the δ-location set as shown in Fig.2 (right), an algorithm satisfying $\{\epsilon, \mathcal{G}_2\}$-location privacy also achieves δ-location set privacy for a certain δ.

3. SYSTEM OVERVIEW

3.1 Epidemic Surveillance

Our system consist of three main modules: PGLP mechanisms, Location Policy Configuration, and Epidemic Surveillance Apps as shown in Fig.3. PGLP mechanisms are proposed in [4] for achieving rigorous and customization location privacy. It takes inputs of ϵ, location policy graph G and the user's true location, and outputs a perturbed location to the server. The policy G is recommended by Location Policy Configuration and approved by the user. Location Policy Configuration defines different location policies according to the application of epidemic surveillance. Three primary functions (Apps) for epidemic surveillance are location monitoring, epidemic analysis and contact tracing. *Location monitoring* focuses on understanding people's movement between different cities or provinces in a coarse-grained level, which provides essential insights when combining with the incidence rate in each city along with the people's movement. It could also provide a "health code" service, i.e., allowing certification of the users health status, in a privacy-preserving way. A location policy for location monitoring can be "*ensuring indistinguishability inside each coarse-grained area and allowing the locations to be distinguishable in different coarse-grained areas*" such as $\mathcal{G}_a$ shown in Fig.4 since such a monitor is only focused on people moving between different cities. *Epidemic analysis* aims at building a predictive disease transmission model such as the SEIR model [13]. The fine-grained data would be beneficial for the estimation of parameters such as R0 (i.e., basic reproduction number). A location policy for epidemic analysis is similar to the previous one, but more fine-grained, such as $\mathcal{G}_b$ in Fig.4. *Contact tracing* attempts to find all contacts of a diagnosed case in order to stop the spread of disease by

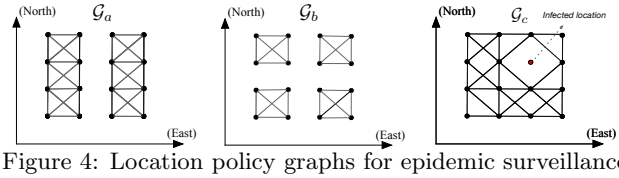

Figure 4: Location policy graphs for epidemic surveillance.

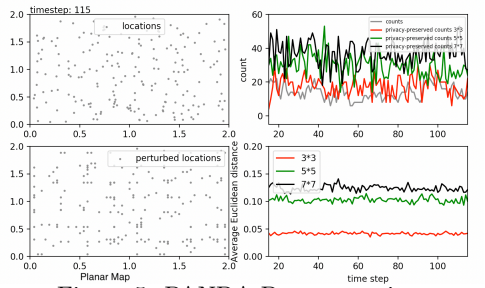

Figure 5: PANDA Demonstration.

finding and isolating patients. A policy for contact tracing can be "ensuring indistinguishability only if the user is not in an infected area, but allowing disclosure of true location if the user is in the vicinity or area of an infected location", which can be formally represented by a graph $\mathcal{G}_c$ in Fig.4. We introduce more details about contact tracing below.

3.2 Demonstration Scenario

We demonstrate the system using Geolife [24] and Gowalla [8] datasets. Interested readers can find a more detailed configuration in [4]. We provide an interactive tool that allows the attendees to explore and examine the usability of our system: (1) the utility of location monitor and coronavirus transmission model estimation, (2) the procedure of contact tracing in our systems, and (3) the privacy-utility trade-offs, as shown in Fig.5, w.r.t. different policy graphs. First, we evaluate the utility of location monitoring as the Euclidean distance between perturbed locations and real locations. We test the accuracy of transmission model estimation using the difference between (i.e., basic reproduction number) R0 estimated over accurate locations and the perturbed locations, respectively. Second, we demonstrate the procedure of contact tracing using our system and dynamic policy graphs (such as $\mathcal{G}_c$ in Fig.4). The goal is identifying the people who have the risk of infection (the decision rule of suspected infection could be advised by CDC or WHO; here we assume a simple rule of two persons have been in the same location at the same time at least twice). At each time point, each user sends the perturbed location w.r.t. her policy graph and stores the past two weeks of location history in a local database. When the server confirms a diagnosed patient's location history, the Policy Graph Configuration module will recommend an updated location privacy policy for the users who have the risk of infection during the past two weeks (according to our simple rule). Then, the corresponding user, upon accepting the policy, will re-send his past location using the updated privacy policy (the places where the diagnosed patient has been are allowed to be disclosed). In this way, the user can get alerted and tested in case of infection. Third, similar to the previous utility evaluation, we will also allow the attendees to evaluate the empirical privacy that is measured by adversary error [18]. One can choose predefined policy graphs, as shown in Fig.4, or randomly generate policy graphs to explore its effect on the privacy-utility trade-off.

4. ACKNOWLEDGEMENT

This work is supported partially by JSPS KAKENHI Grant No. 17H06099, 18H04093, 19K20269, U.S. National Science Foundation (NSF) under CNS-2027783 and CNS-1618932, and Microsoft Research Asia (CORE16).

5. REFERENCES

[1] CDC to set up a coronavirus 'surveillance and data collection system'. https://www.businessinsider.com/cdc-coronavirus-surveillance-and-data-collection-stimulus-package-2020-3, 2020. Business Insider.

[2] European mobile operators share data for coronavirus fight. https://www.reuters.com/article/us-health-coronavirus-europe-telecoms-idUSKBN2152C2, 2020. Reuters.

[3] M. E. Andrs, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *CCS*, pages 901–914, 2013.

[4] Y. Cao, Y. Xiao, S. Takagi, L. Xiong, M. Yoshikawa, Y. Shen, J. Liu, H. Jin, and X. Xu. Customizable and rigorous location privacy through policy graph. In *ESORICS*, 2020.

[5] Y. Cao, Y. Xiao, L. Xiong, and L. Bai. PriSTE: from location privacy to spatiotemporal event privacy. In *IEEE ICDE*, pages 1606–1609, 2019.

[6] Y. Cao, Y. Xiao, L. Xiong, L. Bai, and M. Yoshikawa. PriSTE: protecting spatiotemporal event privacy in continuous location-based services. *PVLDB*, 12(12):1866–1869, 2019.

[7] Y. Cao, Y. Xiao, L. Xiong, L. Bai, and M. Yoshikawa. Protecting spatiotemporal event privacy in continuous location-based services. *IEEE TKDE*, pages 1–1, 2019.

[8] E. Cho, S. A. Myers, and J. Leskovec. Friendship and Mobility: User Movement in Location-based Social Networks. In *ACM KDD*, 2011.

[9] H. Davidson. China's coronavirus health code apps raise concerns over privacy. https://www.theguardian.com/world/2020/apr/01/chinas-coronavirus-health-code-apps-raise-concerns-over-privacy, 2020. The Guardian.

[10] C. Dwork. Differential Privacy. In *ICALP*, pages 1–12, 2006.

[11] B. Gedik and L. Liu. Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms. *IEEE Transactions on Mobile Computing*, 7:1–18, 2008.

[12] X. He, A. Machanavajjhala, and B. Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. pages 1447–1458, 2014.

[13] M. Y. Li and J. S. Muldowney. Global stability for the SEIR model in epidemiology. *Mathematical Biosciences*, 125:155–164, 1995.

[14] N. Li, T. Li, and S. Venkatasubramanian. T-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *IEEE ICDE*, pages 106–115, 2007.

[15] K. Lyons. Governments are using cellphone location data to manage the coronavirus. https://www.theverge.com/2020/3/23/21190700/eu-mobile-carriers-customer-data-coronavirus-south-korea-taiwan-privacy, 2020. The Verge.

[16] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. In *ICDE*, pages 24–24, 2006.

[17] V. Primault, A. Boutet, S. B. Mokhtar, and L. Brunie. The Long Road to Computational Location Privacy: A Survey. *IEEE Communications Surveys Tutorials*, pages 2772 – 2793, 2018.

[18] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. Quantifying Location Privacy. In *IEEE SP*, pages 247–262, 2011.

[19] L. Sweeney. K-anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[20] S. Takagi, Y. Cao, Y. Asano, and M. Yoshikawa. Geo-Graph-Indistinguishability: Protecting Location Privacy for LBS over Road Networks. In *DBSec*, pages 143–163, 2019.

[21] G. Theodorakopoulos, R. Shokri, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec. Prolonging the Hide-and-Seek Game: Optimal Trajectory Privacy for Location-Based Services. In *WPES*, pages 73–82, 2014.

[22] Y. Xiao and L. Xiong. Protecting Locations with Differential Privacy Under Temporal Correlations. In *ACM CCS*, pages 1298–1309, 2015.

[23] Y. Xiao, L. Xiong, S. Zhang, and Y. Cao. LocLok: location cloaking with differential privacy via hidden markov model. *PVLDB*, 10(12):1901–1904, 2017.

[24] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma. GeoLife2.0: A Location-Based Social Networking Service. In *IEEE MDM*, pages 357–358, 2009.