Private-HERMES: A Benchmark Framework for Privacy-Preserving Mobility Data Querying and Mining Methods

Nikos Pelekis
Dept. of Statistics & Ins. Science
University of Piraeus
Piraeus, Greece
npelekis@unipi.gr

Aris Gkoulalas-Divanis
Information Analytics
IBM Research–Zurich
Rüschlikon, Switzerland
agd@zurich.ibm.com

Marios Vodas
Dept. of Informatics
University of Piraeus
Piraeus, Greece
mvodas@gmail.com

Anargyros Plemenos
Dept. of Informatics
University of Piraeus
Piraeus, Greece
plem21@gmail.com

Despina Kopanaki
Dept. of Informatics
University of Piraeus
Piraeus, Greece
dkopanak@unipi.gr

Yannis Theodoridis
Dept. of Informatics
University of Piraeus
Piraeus, Greece
ytheod@unipi.gr

ABSTRACT

Mobility data sources feed larger and larger trajectory databases nowadays. Due to the need of extracting useful knowledge patterns that improve services based on users' and customers' behavior, querying and mining such databases has gained significant attention in recent years. However, publishing mobility data may lead to severe privacy violations. In this paper, we present Private-HERMES, an integrated platform for applying data mining and privacy-preserving querying over mobility data. The presented platform provides a two-dimension benchmark framework that includes: (i) a query engine that provides privacy-aware data management functionality of the in-house data via a set of auditing mechanisms that protect the sensitive information against several types of attacks, and (ii) a progressive analysis framework, which, apart from anonymization methods for data publishing, includes various well-known mobility data mining techniques to evaluate the effect of anonymization in the querying and mining results. The demonstration of Private-HERMES via a real-world case study, illustrates the flexibility and usefulness of the platform for supporting privacy-aware data analysis, as well as for providing an extensible blueprint benchmark architecture for privacy-preservation related methods in mobility data.

General Terms

Algorithms, Management

Keywords

Privacy, Anonymity, Mobility Data, Trajectories, Data Mining

1. INTRODUCTION

Due to the explosion of mobile devices and positioning technologies, moving object data are collected in a massive scale and are becoming increasingly rich, complex and ubiquitous, thus making data analysis a major challenge. The popularity of Moving Object Databases (MOD) nowadays, poses great research opportunities to the data management and mining research community [4], whose primary objective is to efficiently analyze mobility datasets in order to reveal interesting and useful patterns. However, the collection and the disclosure of personal mobility information increase the risk of individuals' privacy violations.

Although a variety of anonymization algorithms that enable the privacy-aware publishing of personal mobility data have been recently proposed in the literature (e.g. [1] [2]), no systemic approach has been taken so far to integrate these algorithms under a common, benchmark-oriented framework. Moreover, until recently, there was no extensible privacy–aware query engine proposed in the research literature that would be able to act as the counterpart of the anonymous data publishing approach, allowing the maintenance of the mobility data in-house and their privacy-protection through carefully designed auditing mechanisms. Along this direction, this paper demonstrates Private-HERMES, a benchmark framework that gives end-users the ability to:

(i) pose regular queries (e.g. range, nearest-neighbor) on a state-of-the-art MOD engine, HERMES [14], and respective privacy-enhanced queries on its recently proposed privacy-aware extension, HERMES++ [15], which protects sensitive information from several types of attacks; the returned answers preserve the privacy of the individuals whose movement is recorded in the dataset, through the generation of some well-crafted, realistic fake trajectories;

(ii) apply popular mobility anonymization algorithms on the original data ([1] [2]), moreover, being able to compare and evaluate the results between the original and the privacy-protected data through a variety of querying and data mining techniques;

(iii) design and execute benchmarks to evaluate the effectiveness of the anonymization and auditing mechanisms, using different workloads of queries; such benchmarks can be used to measure the utility of the anonymized data either by applying data mining techniques and comparing the patterns extracted from original and anonymized (or the fake) data or by posing queries to original and anonymized (or fake) data (or patterns);

EDBT'12, March 26–30, 2012, Berlin, Germany.

(iv) interactively and progressively repeat the previous three analysis operations in subsets of mobility data, which are either the results of MOD queries or the mapping of discovered mobility mining models (e.g. a cluster or a set of frequent sequential patterns) to data;

(v) automatically build the profile of the end-user based on his or her queries to the database, allowing the data administrator to identify suspicious user behavior [5].

The Private-HERMES platform adopts the above ideas, proposes solutions, and demonstrates their implementation. To our knowledge this is the first benchmark framework that includes a complete set of state-of-the-art mobility anonymization algorithms and mobility data mining techniques, which have been integrated with both a query engine and a privacy-aware query engine.

The rest of this paper is organized as follows. Section 2 provides a brief description of the techniques that have been integrated into the benchmark platform. In Section 3, we illustrate the architecture of the proposed framework. Section 4 describes demo specifications through different application scenarios.

2. A PRIVACY BENCHMARK FOR MOD

The first dimension of an envisioned benchmark w.r.t. privacy issues involves in-house stored data and privacy–aware query answering. Private-HERMES incorporates HERMES [14], a query engine based on a powerful query language for trajectory databases, which enables the support of aggregative queries. HERMES supports a variety of well-known queries such as range, nearest neighbor, topological, directional queries, etc. On top of this functionality, we recently introduced HERMES++ [15], a privacy-preserving trajectory query engine that allows subscribed users to gain restricted access to the trajectory database to accomplish various analysis tasks. HERMES++ audits queries for trajectory data to block potential attacks to user privacy, supports the most popular spatiotemporal queries (range, distance, k-NN) and preserves user privacy by generating carefully crafted, realistic fake trajectories.

In particular for attack handling, which is the main objective for such a system, HERMES++ protects user privacy by blocking three types of attacks that malevolent users may try to pursue in the database: *user identification attack* (one may associate an individual to his or her trajectory by submitting ad hoc queries involving overlapping spatio-temporal regions), *sensitive location tracking attack* (one may reveal a user's identity by identifying points of interest, i.e. starting or ending points of user trajectories, the address of a house or a betting office, and then map-matching them with users' *sensitive* location information; such locations are called *sensitive* for a specific user, as they should not be disclosed to the attackers), and *sequential tracking attack* (one may disclose a user's habits, which is privacy violation, by "following" the trajectory of a user in the system, i.e. by posing a set of focused queries on regions that are spatially and temporally close to each other, thus learning sensitive places that the user has visited).

The second dimension w.r.t. privacy that is supported by our benchmark involves privacy-preserving MOD publishing. The objective of mobility data anonymization is to sanitize a dataset so that a malevolent user can no longer match the recorded movement of an object to a specific individual. The algorithms that have been integrated in Private-HERMES to help anonymize trajectories are NWA [1] and W4M [2]. NWA introduces the concept of *(k,δ)-anonymity*, where δ represents the location impression. The method is based on trajectory clustering and spatial translation in order to make a trajectory lie within the anonymity cylinder that contains at least K-1 other trajectories. In order to achieve space-time translation, the authors proposed W4M [2], which uses a different distance measure that allows time-warping. Both algorithms take as input a set of trajectories and publish atomic anonymized trajectories suitable for trajectory data mining applications. The objective of Private-HERMES is to support the evaluation of such anonymization techniques and to study their effect in the utility of the sanitized data, when compared with queries into the original MOD. Private-HERMES further supports the comparison of the query results posed also on fake trajectories produced by the auditing mechanism, as discussed earlier [15].

Private-HERMES also gives the ability to users to evaluate the utility either of the fake or the sanitized trajectories via a variety of well-known mobility data mining algorithms, i.e. various types of clustering, frequent sequential patterns, etc. The idea is that by adding fake trajectories (that affect the cardinality of the MOD), as well as perturbating original ones (that affects the shape of the MOD) should not destroy the patterns hidden in the original MOD. Such an evaluation can be done by using clustering and frequent pattern mining techniques, appropriate for mobility data. Towards this goal, Private-HERMES incorporates three state-of-the-art trajectory clustering algorithms, namely TRACLUS [7], T-Optics [8] and CenTR-I-FCM [16]. K-medoids [6] and Bisecting K-medoids [19] are also included as representative examples of traditional clustering techniques that can applied in MOD with the special feature that the user can choose different distance functions between the trajectories (i.e. grouping only by their starting or destination point, without taking into account the whole route) [13]. As for frequent pattern mining, Private-HERMES incorporates the T-pattern mining technique [3], which models sequences of visited regions, frequently visited in the specified order with similar transition times, out of trajectory databases.

In particular for trajectory clustering, a useful requirement is to extract a compact representation of the clusters found, in terms of "representative" or "typical" trajectories that effectively represent the cluster sets. To achieve this, Private-HERMES supports CenTra "centroid" trajectories [16] and TRACLUS "typical" trajectories [7]. Last but not least, the user may also work on large datasets by first sampling the initial MOD by appropriate methods that preserve the hidden patterns by simultaneously covering the whole space [17].

The above presented functionality is integrated in HERMES MOD engine by appropriately extending the query language with new constructs, in a fashion originally proposed in [12]. This allows users to progressively analyze the MOD and interchange between querying and mining operations.

3. SYSTEM IMPLEMENTATION

The main advantage of the Private-HERMES platform is that it offers users the ability to perform different processes on mobility data, as shown in Figure 1. The user interacts with a GUI with 3D rendering capabilities developed in Java and based on the Swing GUI widget toolkit [11]. The results from the operations that the program supports are visualized in the 3D globe provided by NASA World Wind [9]. To draw the charts reporting performance results, the JFreeChart library was used [10]. Every component and library used during the development process is open source.

Through the provided GUI, the user is able to setup his/her benchmark or, more generally, his/her analysis scenario. Private-HERMES retrieves the necessary data by calling the HERMES

MOD engine. The supported mobility data mining and anonymization algorithms have been incorporated as modules of the extensible DAEDALUS's MO-DMQL [12], while both of these sets of algorithms exchange data (i.e. real / fake / anonymized trajectories and mining models) directly with the database layer.

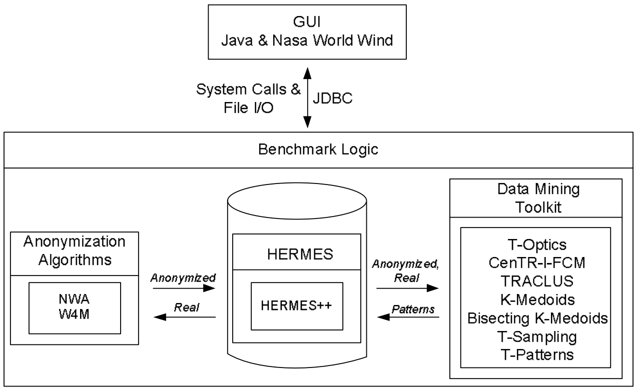

Figure 1. System Architecture

Zooming in at the database layer, HERMES++ exploits on the trajectory storage functionality and the spatiotemporal query processing capabilities of HERMES for providing privacy-aware queries to end-users. More specifically, HERMES defines a trajectory data type and a collection of operations as an Oracle data cartridge, which is further enhanced by the TB-tree access method [18] for efficient querying on trajectory data. HERMES++ directly utilizes this functionality at the ORDBMS level to store real and fake trajectories, as well as any historical information of all the users' queries (and the corresponding responses), in order to avoid different types of tracking attacks (e.g., sequential tracking). It succeeds so by the embedded auditing module, which invokes the HERMES queries and the fake trajectory generator algorithm [15]. Since the entire framework is built at the ORDBMS level, end-users are also able to pose their queries through PL/SQL (i.e. not only via the GUI). As such, from an architectural point of view, HERMES++ acts as a wrapper over the HERMES query engine and not as a secure middleware. Figure 2 illustrates the HERMES++ architectural framework.

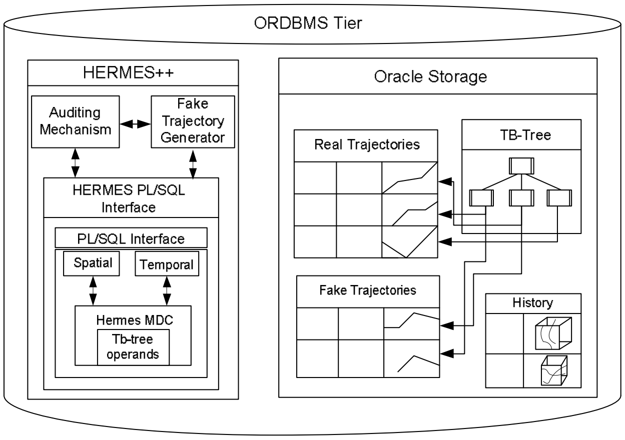

Figure 2. The Architecture of HERMES++ [15]

4. DEMO SPECIFICATIONS

Throughout the demonstration, users will be able to test the system by using a real dataset consisting of GPS traces of cars moving in the city of Milan, Italy. The different functionalities that are currently supported by the Private-HERMES platform are:

- *Querying and mining operations on HERMES:* the platform is capable of executing simple queries on HERMES, such as range and NN queries. Queries are posed via the GUI, which provides essential capabilities, including query predicate selection, parameters selection and results projection. Graphical map user-interaction for predicate definition is also supported. Moreover, there is extensive support for the most representative mobility data mining algorithms for clustering (T-Optics [8], CenTR-I-FCM [16], TRACLUS [7], K-medoids [6], Bisecting K-medoids [19]) frequent pattern mining (T-Patterns [3]) and sampling purposes (T-Sampling [17]).
- *Privacy-aware queries on HERMES++*: the user has the ability to run the aforementioned queries using HERMES++ [15]. HERMES++ is able to protect users whose movements are recorded in the database from privacy attacks, i.e. user identification attack, sensitive location tracking attack and sequential tracking attack, issued by malevolent end-users. The data owner requires that at least a certain number of trajectories are returned to the end-users in response to their queries, for all different types of supported queries. The result consists of a set of carefully crafted, realistic fake trajectories aiming to preserve the trend of the original user trajectories.
- *Comparison / evaluation of anonymization algorithms*: the platform integrates two well-known anonymization algorithms, namely NWA [1] and W4M [2]. Both algorithms take as input trajectories which may have been extracted from a query posed to HERMES, and transform them into anonymous equivalents, subsequently stored in the MOD. An advantage of the platform is its ability to design and execute benchmarks that evaluate the results from the application of anonymization algorithms regarding the distortion over real user trajectories. The incorporated data mining techniques can be applied, and patterns steaming from original data with patterns resulting from anonymized data can be compared. This can be achieved by executing queries in the original and in the anonymized data (or patterns), and comparing the attained results.
- *Profiling end-user's behavior to identify malevolent users*: The platform supports query auditing techniques [5], which can be used to monitor the behavior of the end-users and build user profiles. These user profiles can be subsequently analyzed by the data owner, as explained in [5], to help him or her identify suspicious behavior of end-users in the system.

In Figures 3-6 are some representative snapshots of the Private-HERMES GUI. In Figure 3, original data have been extracted using a range query, while in Figure 4 the data have been anonymized using NWA. From these outputs, a user can compare the distortion that has been caused to the trajectory database after the application of the anonymization algorithm. In Figure 5, the result from the application of T-Optics [8] on the original data is depicted, while Figure 6 presents the result from applying T-Optics on the anonymized data. The extracted patterns can be visually compared.

A preview of Private-HERMES including more screenshots from the GUI, experimental results and an accompanying video is available at: http://infolab.cs.unipi.gr/pubs/edbt2012/

5. ACKNOWLEDGMENTS

Research partially supported by the EU FP7/ICT/FET Projects MODAP (Mobility, Data Mining and Privacy; 2009-12) and DATASIM (Data Science for Simulating the Era of Electric

Vehicles; 2011-14), both funded by the European Union. The implementation of the tool was done by the University of Pireaus.

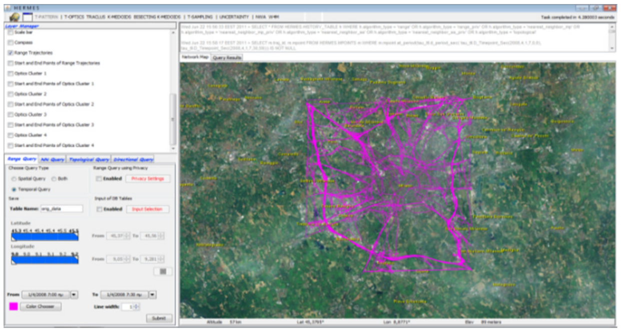
Range Query
Choose Query Type
○ Spatial Query ○ Both
● Temporal Query
Range Query using Privacy
□ Enabled Privacy Settings
Save
Table Name: orig_data
Input of DB Tables
□ Enabled Input Selection
From 1/4/2008 7:00 πμ To 1/4/2008 7:30 πμ
Color Chooser Line width: 1
Submit

Figure 3. Original Data from a Range Query

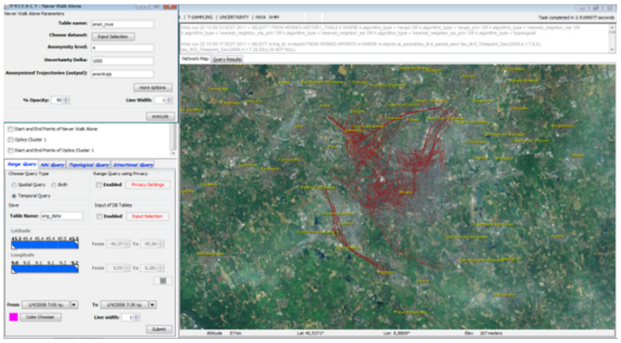
PRIVACY - Never Walk Alone
Never Walk Alone Parameters
Table name: anon_nwa
Choose dataset: Input Selection
Anonymity level: 4
Uncertainty Delta: 1000
Anonymized Trajectories (output): anontraja
more options
% Opacity: 90 Line Width: 1
execute
□ Start and End Points of Never Walk Alone
□ Optics Cluster 1
□ Start and End Points of Optics Cluster 1
Range Query
Choose Query Type
○ Spatial Query ○ Both
● Temporal Query
Range Query using Privacy
□ Enabled Privacy Settings
Save
Table Name: orig_data
Input of DB Tables
□ Enabled Input Selection
From 1/4/2008 7:00 πμ To 1/4/2008 7:30 πμ
Color Chooser Line width: 1
Submit

Figure 4. NWA-based Anonymized Data

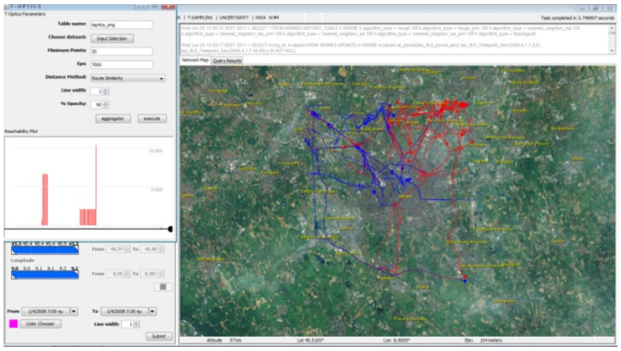
T-OPTICS
T-Optics Parameters
Table name: toptics_orig
Choose dataset: Input Selection
Minimum Points: 20
Eps: 7000
Distance Method: Route Similarity
Line width: 1
% Opacity: 90
aggregator execute
Reachability Plot
From 1/4/2008 7:00 πμ To 1/4/2008 7:30 πμ
Color Chooser Line width: 1
Submit

Figure 5. T-Optics applied on the original MOD

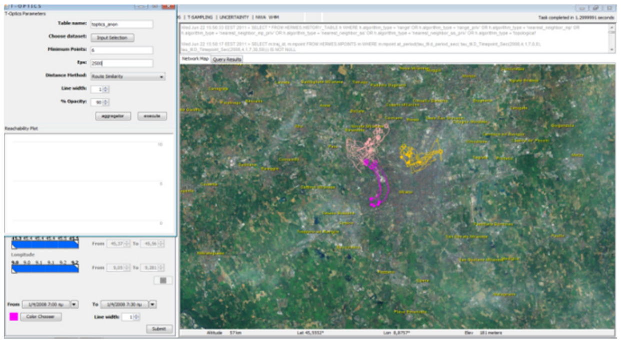
T-OPTICS
T-Optics Parameters
Table name: toptics_anon
Choose dataset: Input Selection
Minimum Points: 6
Eps: 2500
Distance Method: Route Similarity
Line width: 1
% Opacity: 90
aggregator execute
Reachability Plot
From 1/4/2008 7:00 πμ To 1/4/2008 7:30 πμ
Color Chooser Line width: 1
Submit

Figure 6. T-Optics applied on the Anonymized MOD

6. REFERENCES

[1] Abul, O., Bonchi, F., and Nanni, M. (2008). Never walk alone: Uncertainty for anonymity in moving objects databases. In *Proceedings of ICDE.*

[2] Abul, O., Bonchi, F., and Nanni, M. (2010). Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884-910.

[3] Giannotti, F., Nanni, M., Pedreschi, D., and Pinelli, F. (2007). Trajectory Pattern Mining, In *Proceedings of SIGKDD*.

[4] Giannotti, F. and Pedreschi, D. (2008). *Mobility, Data Mining and Privacy, Geographic Knowledge Discovery*. Springer-Verlag.

[5] Gkoulalas-Divanis, A. and Verykios, V. S. (2008). A privacy–aware trajectory tracking query engine. *SIGKDD Explorations*, 10(1):40-49.

[6] Kaufman, L., Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, NY.

[7] Lee, J.-G., Han, J., and Whang, K.-Y. (2007). Trajectory clustering: a partition-and-group framework. In *Proceedings of SIGMOD*.

[8] Nanni, M. and Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects, *Journal of Intelligent Information Systems*, 27(3):267-289.

[9] NASA, World Wind Java SDK. URL: http://worldwind.arc.nasa.gov/java. (accessed: 19 Jan. 2012)

[10] Object Refinery, the JFreeChart project. URL: http://www.jfree.org/jfreechart. (accessed: 19 Jan. 2012)

[11] Oracle, The Swing Tutorial. URL: http://download.oracle.com/javase/tutorial/uiswing. (accessed: 19 Jan. 2012)

[12] Ortale, R., Ritacco, E., Pelekis, N., Trasarti, R., Costa, G., Giannotti, F., Manco, G., Renso, C., and Theodoridis, Y. (2008). The DAEDALUS Framework: Progressive Querying and Mining of Movement Data. In *Proceedings of ACM GIS*.

[13] Pelekis, N., Andrienko, G., Andrienko, N., Kopanakis, I., Marketos, G., and Theodoridis, Y. (2011). Visually Exploring Movement Data via Similarity-based Analysis”, *Journal of Intelligent Information Systems*, online first.

[14] Pelekis, N., Frentzos, E., Giatrakos, N., and Theodoridis, Y. (2008). HERMES: Aggregative LBS via a trajectory DB engine. In *Proceedings of SIGMOD*.

[15] Pelekis, N., Gkoulalas-Divanis, A., Vodas, M., Kopanaki, D., and Theodoridis, Y. (2011). Privacy-Aware Querying over Sensitive Trajectory Data. In *Proceedings of CIKM*.

[16] Pelekis, N., Kopanakis, I., Kotsifakos, E., Frentzos, E. and Theodoridis, Y. (2011). Clustering uncertain trajectories. *Knowledge and Information Systems*, 28(1):117-147.

[17] Pelekis, N., Panagiotakis, C., Kopanakis, I., and Theodoridis, Y. (2010). Unsupervised trajectory sampling. In *Proceedings of ECML PKDD*.

[18] Pfoser, D., Jensen, C. S., and Theodoridis, Y. (2000). Novel approaches to the indexing of moving object trajectories. In *Proceedings of VLDB.*

[19] Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of document clustering techniques. In *Proceedings of KDD Workshop on Text Mining*