A New Privacy-Preserving Solution for Clustering Massively Distributed Personal Times-Series
Tristan Allard, Georges Hébrail, Florent Masseglia, Esther Pacitti

To cite this version:
Tristan Allard, Georges Hébrail, Florent Masseglia, Esther Pacitti. A New Privacy-Preserving Solution for Clustering Massively Distributed Personal Times-Series. ICDE: International Conference on Data Engineering, May 2016, Helsinki, Finland. pp.1370-1373, 10.1109/ICDE.2016.7498347 . lirmm-01270268

HAL Id: lirmm-01270268
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01270268
Submitted on 2 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A New Privacy-Preserving Solution for Clustering Massively Distributed Personal Times-Series

Tristan Allard†, Georges Hébrail*, Florent Masseglia‡, Esther Pacitti‡
† IRISA & Univ. Rennes 1, 263 av. Général de Gaulle, 35042 Rennes cedex, France
* EDF R&D, 1 av. Général de Gaulle, BP 408, 92141 Clamart cedex, France
‡ Inria & Lirmm, Univ. Montpellier, Campus St Priest, 860 rue de St Priest, 34095 Montpellier cedex 5, France

first.last@{†irisa.fr, *edf.fr, ‡inria.fr, ‡lirmm.fr}

Abstract—New personal data fields are currently emerging due to the proliferation of on-body/at-home sensors connected to personal devices. However, strong privacy concerns prevent individuals to benefit from large-scale analytics that could be performed on this fine-grain highly sensitive wealth of data. We propose a demonstration of Chiaroscuro, a complete solution for clustering massively-distributed sensitive personal data while guaranteeing their privacy. The demonstration scenario highlights the affordability of the privacy vs. quality and privacy vs. performance tradeoffs by dissecting the inner working of Chiaroscuro - launched over energy consumption times-series -, by exposing the results obtained by the individuals participating in the clustering process, and by illustrating possible uses. Keywords—differential privacy, gossip, k-means, time-series;

I. INTRODUCTION

The ongoing wave of personal sensors is leading to a massive generation of personal time-series related to individuals’ health (e.g., through smart wristbands or smart bed scales) or to inner-home activities (e.g., through electrical smart plugs or smart meters). These time-series are typically sent to their owner’s personal devices, such as his smartphone, his tablet, or his laptop.

The availability of this wealth of data massively distributed over personal devices is an unprecedented opportunity for individuals to learn valuable knowledge. Cluster analysis, also called clustering, aims at forming groups of data (or clusters) such that similar data appear in the same group and dissimilar data appear in different groups. Clustering is widely used in various application domains, e.g., medicine, genetics, marketing, energy, or social sciences. We believe that clustering personal time-series can directly benefit individuals. Assume for example that time-series contain variations of weight. An individual suffering from obesity could benefit from the clustering results by identifying interesting groups of weight time-series - e.g., groups within which weight time-series are similar to his own time-series on a subsequence but exhibit finally a steady decrease – in order to further discover and investigate the associated diets. Similar uses can benefit individuals in many situations (e.g., clustering electrical consumption time-series for identifying the low-consumption groups and discover the equipments that could be replaced to improve the electrical consumption).

A naive approach for clustering these masses of personal time-series could simply consist in copying them from the set of personal devices to a central server which could then apply the usual cluster analysis techniques. However, fine grain personal time-series are sensitive data. A personal time-series indeed conveys a large quantity of information about the individual(s) it concerns, and may lead to arbitrary disclosures [1]. Although entrusting a single entity with large amounts of data may be practical, recurring large scale data breaches show that this is actually hazardous in the real world (see, e.g., http://datalossdb.org). Moreover, the frequent occurrence of data breaches fuels the growing concern of individuals about the systematic centralization of their personal data. As a result, the centralized approach not only introduces an additional risk to data privacy but also an additional obstacle to large-scale personal data analysis.

In this demonstration, we present Chiaroscuro [2], a complete solution for clustering personal time-series that are massively distributed on a large population of honest-but-curious personal devices without jeopardizing their privacy. The challenge addressed by Chiaroscuro arises from the conjunction of the massive distribution of the execution over possibly faulty computing nodes on one side, with the strong privacy guarantees that personal time-series need on the other side. To the best of our knowledge, related works fail in addressing one side of the challenge or the other. This is essentially due either to the use of non fault-tolerant cryptographic techniques or to the presence of out-of-control data disclosures (see [2] for a deeper discussion of related works).

In order to address together the two sides of the challenge, Chiaroscuro relaxes the traditional security desideratum by allowing differentially private intermediate results to be dis-
closed during the execution, any other information remaining encrypted. Demonstrating Chiaroscuro essentially aims at showing that using encryption together with differentially private perturbation can be a key enabler in the design of massively distributed privacy-preserving analytical algorithms. The demonstration highlights the main arguments that support this claim: (1) a high level of privacy can be reached (i.e., probabilistic variant of differential privacy), (2) a high level of quality can be reached (similar to the quality of centralized clustering results), and (3) costs remain affordable given the resources of today’s personal devices.

II. CHIASROSCURO

A. Preliminaries In a Nutshell

Clustering. $k$-means [3] aims at proposing $k$ clusters that optimize an objective quality function1 (e.g., the intra-cluster inertia which measures the homogeneity of the set of time-series within clusters). In general, a cluster can be described by extension by enumerating its content, or by intension with e.g., the average of its time-series, called centroid or profile below. $k$-means is an iterative algorithm that progressively “fits” to data a set of $k$ proposed centroids. It starts the first iteration by choosing the $k$ initial centroids, e.g., at random, and terminates when a termination criterion is satisfied (e.g., the centroids converge, or a given number of iterations is reached). Each iteration follows three steps:

1) Assignment step: For each time-series, get its closest centroid and assign it to the corresponding cluster;
2) Computation step: For each cluster, compute its average time-series, which is its candidate centroid for the next iteration if any (also called its mean below);
3) Convergence step: If the distance between the set of centroids and the set of means is greater than a given threshold, then another iteration starts, taking as input centroids the set of means (go to Step 1), otherwise return the set of means;

Distribution. Gossip aggregation algorithms are lightweight fully decentralized approximate algorithms executed within large sets of participants. They simply consist of periodical point-to-point exchanges between participants, called gossip exchanges in the rest of the document. Each participant holds its own approximation of the global aggregate and updates it, at each exchange, with the one of the communicating participant. The approximation error depends on the number of gossip exchanges per participant and is guaranteed to converge to zero exponentially fast [4].

Privacy. Chiaroscuro’s privacy guarantees rely on two building blocks: an additively-homomorphic encryption scheme on one side and the differentially-private Laplace perturbation scheme on the other side. First, Chiaroscuro is independent of any specific encryption scheme provided that (1) it satisfies the strong semantic-security level, (2) it is additively-homomorphic, and (3) the decryption is performed collaboratively by any subset of participants provided its sufficiently large. The Damgard-Jurik encryption scheme [5] that we use in our implementation is an instance of such schemes.

Second, differential privacy is the current de facto standard for disclosing to untrusted parties aggregated data, such as a sum of a set of time-series for example, while guaranteeing a strong privacy level to individuals having participated in the aggregate. The Laplace perturbation mechanism satisfies the $\epsilon$-differential privacy model [6] by adding random noise to the aggregate to be disclosed. The noise is sampled in a Laplace distribution parameterized according to $\epsilon$ and to the aggregate disclosed. A Laplace random variable can be computed by summing up $n$ terms independently generated based on the gamma distribution, $n$ being fixed beforehand. These terms are called noise-shares. When several aggregates related to the same individuals are perturbed and disclosed, differential privacy is still satisfied (self-composition property) and the global privacy level, seen as a privacy budget, must be divided among the perturbations because it is is the addition of all the privacy levels set for perturbing the various aggregates. Note that because of the inherent approximations of gossip algorithms, Chiaroscuro satisfies a probabilistic variant of $\epsilon$-differential privacy.

B. Overview of Chiaroscuro

Diptych Data Structure. Parallelizability is a crucial property in this massively distributed context, making a clustering algorithm such as $k$-means especially relevant. $k$-means is essentially based on a twofold data structure made of the centroids on one side and of the means on the other side. The centroids are used during Steps 1 and 3 (i.e., assignment and convergence steps) for performing distance comparisons, and the means result from the algebraic computations performed in Step 2 (i.e., computation step). Chiaroscuro brings together additively-homomorphic encryption and differentially-private perturbation in order to shield the two sides of this data structure while enabling the gossip-based distribution of the execution sequence. In Chiaroscuro, Step 2 is performed over additively-homomorphic encrypted means, which support the algebraic operations required by this step, while Step 1 and Step 3, harder to perform over encrypted data, are performed over cleartext centroids perturbed to satisfy differential privacy. The resulting data structure consists thus of the perturbed centroids on one side and of the encrypted means on the other side; it is called Diptych and is key to the execution sequence.

Execution Sequence. Chiaroscuro’s execution sequence revisits the assignment, computation, and convergence steps by distributing and articulating them together based on the diptych data structure. It is iterative, identical for all participants, and proceeds without any global synchronization (the late participants simply synchronize on the latest iteration during their gossip exchanges). It consists in the following:

1) Assignment step (local): pull in a set of perturbed centroids, assign the local time-series to the closest perturbed centroid and initialize (1) the corresponding encrypted mean with the encryption of the local time-series and (2) all the other means with the encryptions of zero-valued time-series;
2) Computation step (distributed): Compute the set of perturbed means:
   a) Gossip computation of the encrypted means;
   b) Gossip computation of the encrypted noises;

---

1Characterizing precisely the algorithms that Chiaroscuro can support is future work.
Chiaroscuro supports the addition of other termination criteria for coping with the impact of the differentially-private perturbation on the convergence of centroids (e.g., monitoring centroids quality).
technical aspects of Chiaroscuro. The GUI is divided into a sequence of screens. After the introductory screens explaining Chiaroscuro and the demonstration platform, and allowing to set the mutable parameters, the GUI displays a set of graphs. The first type of graphs is passive, dedicated to visualize, e.g., the evolutions of the centroids, the noise values, and the quality and cost measures, based on slide bars over the iterations. The second type of graph is interactive, it illustrates the use of the resulting centroids by Bob, for example by displaying Bob’s time-series and allowing the audience to interact with it by selecting a sub-sequence and finding the centroids the closest to the sub-sequence chosen.

IV. ACKNOWLEDGMENT

The authors warmly thank C. Maupetit and M. Simonin from Inria Rennes for having designed and developed the GUI system.

REFERENCES