ISSN 2186-7437

# NII Shonan Meeting Report

No. 2018-5

# Anonymization methods and inference attacks: Theory and Practice

Hiroaki Kikuchi, professor, Meiji University,
Josep Domingo-Ferrer, professor, Universitat Rovira i Virgili
Prof. Sbastien Gambs, Universit du Qubec  Montral (UQAM)

March 5 - 8, 2018

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

# Anonymization methods and inference attacks: Theory and Practice

Organizers:
Hiroaki Kikuchi (Meiji University)
Josep Domingo-Ferrer (Universitat Rovira i Virgili)
Sébastien Gambs (Université du Québec à Montréal (UQAM)

March 5 – 8, 2018

## Meeting Objectives

The democratization of mobile systems and the development of information technologies have been accompanied by a massive increase of the amount and the diversity of data collected about individuals. For instance, some actors have access to personal data such as social relationships, email content, income information, medical records, credit card and loyalty card usage, pictures taken through public and private cameras, personal files, navigation behaviour, or data issued from quantified self, just to name a few. On the one hand, the analysis of these large scale datasets, often refer to as Big Data, offer the possibility to realize inferences with an unprecedented level of accuracy and details. On the other hand, the massive collection of information raises many privacy issues since most of these large scale datasets contain personal information, which is thus sensitive by nature. As a result, only very few of them are actually released and available. This limits both our ability to analyze such data to derive information that could benefit to the general public and slows down the innovative services that could emerge from such data. It is therefore important to study anonymization mechanisms that can be used to remove the sensitive information or add uncertainty to a dataset before it is released or before further services are developed on it.

Designing an anonymization method that provides strong privacy guarantees while maintaining a high level of utility is known to be difficult task. In particular, pseudonymization is clearly not at alternative as illustrated by infamous examples of privacy failures such as the AOL release or the Netflix challenge. In addition, there is no free-lunch in anonymization and each type of data comes with its own challenges that have to be dealt with. For instance, to address appropriately the particularities of a genomic dataset, mobility traces or a social graph require the development of an anonymization method tailored to the specifics of the data considered. Nonetheless, defining realistic and formally grounded measures of privacy, which are adapted and appropriate for specific contexts, is a challenging task but also a prerequisite both for evaluating the risks and for assessing potential solutions. One of the main difficulties is to be able to design and formalize realistic adversary models, by taking into account the background knowledge of the adversary and his inference capabilities. In particular, many privacy models currently exist in the literature such as $k$-anonymity, and its extensions such as $\ell$-diversity and $t$-closeness, or more recently differential privacy, pan-privacy and empirical privacy. However, these models are not necessarily comparable and what might appear to be the optimal anonymization method in one model is not necessarily the best one for a different model. To be able to assess the privacy risks of publishing a particular anonymized data, it is necessary to practically evaluate the accuracy inference attacks that can be performed by the adversary based on the released data but also on the possible background knowledge that he might have gathered. In addition, of the risk of re-identifying an individual, inference attacks can also target specific attributes. For instance, considering the example of location data, an inference attack can use the mobility data of a user, possibly with some auxiliary information, to deduce other personal data (home and place of work, main interests, social network, etc.), including sensitive data (in the legal sense) such as religion, health condition or business confidential data coming from the users employer.

The main objective of the Shonan meeting is precisely to investigate the strengths and limits of existing anonymization methods, both from theoretical and practical perspective. More precisely, by confronting the points of views of privacy experts coming from diverse background such as databases, cryptography, theoretical computer science, machine learning, quantitative information, graph theory and social sciences, we aim at gaining an in-depth understanding on how to quantify the privacy level provided by a particular anonymization method as well as the achievable trade-off between privacy and utility of the resulting data. The outcomes of the meeting will greatly benefit to the privacy community and one of our objectives is to use them to design an international anonymization competition.

## List of Participants

- Prof. Hiroaki Kikuchi, Meiji University, Japan
- Prof. Josep Domingo-Ferrer, Universitat Rovira i Virgili, Spain
- Prof. Sébastien Gambs, Université du Québec à Montréal (UQAM), Canada
- Dr. Koki Hamada, NTT Secure Platform Laboratories, Japan
- Prof. Kazuhiro Minami, Institute of Statistical Mathematics, Japan
- Dr. Anna Oganyan, U.S. National Center for Health Statistics, USA
- Prof. Tamir Tassa, The Open University of Israel, Israel
- Prof. Vicenç Torra, University of Skovde, Sweden
- Prof. Chiemi Watanabe, University of Tsukuba, Japan
- Prof. Hiroshi Nakagawa, The University of Tokyo, Japan
- Prof. Chris Clifton, Purdue University, USA

- Pror. Bradley Malin, Vanderbilt University, USA
- Prof. Yucel Saygin, Sabanci University, Turkey
- Prof. Goran Lesaja, Georgia Southern University, USA
- Dr. Takao Murakami, AIST, Japan
- Dr. Shogo Masaki, NTT, Japan
- Prof. Yuichi Sei, The University of Electro-Communications, Japan
- Dr. Yusuke Kawamoto, AIST, Japan
- Mr. Shinichi Miyazawa, Secom, Japan
- Dr. Andrew Baker, Privacy Analytics, Canada
- Ms. Santa Borel, Privacy Analytics, Canada
- Mr. Antoine Laurent, Université du Québec à Montréal (UQAM), Canada
- Mr. Antoine Boutet, INSA Lyon/INRIA France
- Mr. Paul Francis, Max Planck Institute for Software Systems, Germany
- Mr. Axel Michel, INSA Centre Val de Loire, France
- Prof. Tristan Allard, Université de Rennes 1, France
- Prof. Jun Sakuma, University of Tsukuba, Japan

## Meeting Schedule

**Check-in Day: March 4, 2018.**

- Welcome Banquet

**Day 1: March 5 (Mon)**

- Session 1: **Introduction**
  3 slides per participant (1 on personal background; 1 on current research; 1 giving the view on the Seminar goal)
- Session 2: **Anonymization 1**
  Chair: Hiroaki Kikuchi
  - Sébastien Gambs (Université du Québec à Montréal): Privacy-preserving WiFi Analytics
- Session 3: **Inference attacks 1**
  Chair: Sébastien Gambs
  - Takao Murakami (National Institute of Advanced Industrial Science and Technology): Expectation-Maximization Tensor Factorization for Practical Location Privacy Attacks
  - Antoine Laurent (Université du Québec à Montréal): Inference attacks on Montreal open data
- Session 4: **Anonymization 2**
  Chair: Josep Domingo-Ferrer
  - Chris Clifton (Purdue University): Partitioning for anonymization
  - Goran Lesaja (Georgia Southern University): A new approach to solving continuous CTA model
  - Tristan Allard (Université de Rennes 1): Towards Using Differential Privacy as a Building Block for Privacy-Preserving Algorithms
  - Paul Francis (Max Planck Institute for Software Systems) : Diffix: Strong Anonymization with Good Utility

**Day 2: March 6 (Tue)**

- Session 5: **Competition 1**
  Chair: Sébastien Gambs
  - Hiroaki Kikuchi (Meiji University): PWSCUP 2017 Report on Anonymization Competition
  - Hiroaki Kikuchi (Meiji University): Plan for Anonymization Competition in 2018
  - Paul Francis (Max Planck Institute for Software Systems): Experiences with the Aircloak Anonymization Bounty Program
- Session 6: **Privacy models 1**
  Chair: Josep Domingo-Ferrer

- Vicenc Torra (University of Skovde): On disclosure risk measures: last results
- Yusuke Kawamoto (AIST): Extension of Differential Privacy to Distribution Obfuscation

- Session 7: **Inference Attacks 2**
  Chair: Hiroaki Kikuchi
  - Koki Hamada (NTT Secure Platform Laboratories): Re-identification with and without knowledge about anonymization algorithm
  - Shogo Masaki (NTT): Towards evaluating anonymity of trajectory data against an adversary model with realistic background knowledge
- Session 8: **Anonymization 3**
  Chair: Sébastien Gambs
  - Josep Domingo-Ferrer (Universitat Rovira i Virgili): Big data anonymization requirements
  - Anna Oganian (National Center for Health Statistics, Centers for Disease Control and Prevention): Synthetic genetic data as an alternative to restricted use genetic data at government institutions
  - Axel Michel (INSA Centre-Val-de-Loire): Optimal personalized $k$-anonymity with constraint clustering
  - Hiroshi Nakagawa (University of Tokyo): Anonymization method for Anonymized Personal Information

**Day 3: March 7 (Wed)**

- Session 9: **Privacy Models 3**
  Chair: Hiroaki Kikuchi
  - Bradley Malin (Vanderbilt University): A Game Theoretic Perspective on Data Privacy
  - Shinichi Miyazawa (SECOM CO.,LTD.): Privacy-Preserving Data Collection for Improving UI/UX
  - Jun Sakuma (University of Tsukuba): Continual counting under local differential privacy
- Session 10: **Anonymization 4**
  Chair: Josep Domingo-Ferrer
  - Andrew Baker (Privacy Analytics): Overlapping Anonymization Projects in Clinical Trial Transparency
  - Tamir Tassa (The Open University of Israel): Anonymizing Graphs
- Excursion and Main Banquet

**Day4: March 8 (Thu)**

- Session 11: **Future plan and open discussion**
  Chair: Sébastien Gambs

- Antoine Boutet (INSA Lyon) : Anonymization of WiFi logs
- Open discussion session to wrap-up the workshop
- Closing

# Overview of Talks

## Privacy-preserving Wi-Fi Analytics

Sébastien Gambs, Université du Québec à Montréal

As communications-enabled devices are becoming more ubiquitous, it becomes easier to track the movements of individuals through the radio signals broadcasted by their devices. Thus, while there is a strong interest for physical analytics platforms to leverage this information for many purposes, this tracking also threatens the privacy of individuals. To solve this issue, we propose a privacy-preserving solution for collecting aggregate mobility patterns while satisfying the strong guarantee of -differential privacy. More precisely, we introduce a sanitization mechanism for efficient, privacy-preserving and non-interactive approximate distinct counting for physical analytics based on perturbed Bloom filters called Pan-Private BLIP. We also extend and generalize previous approaches for estimating distinct count of events and joint events (i.e., intersection and more generally t-out-of-n cardinalities). Finally, we evaluate expirementally our approach and compare it to previous ones on real datasets.

## Expectation-Maximization Tensor Factorization for Practical Location Privacy Attacks

Takao Murakami, AIST

In this talk, we propose EMTF (Expectation-Maximization Tensor Factorization) for practical location privacy attacks, as suggested by the title. While many people use a number of location-based services (LBS) such as map, route finding, and location check-in, the reveal of their locations raises serious privacy concerns. A various kinds of location privacy attacks have been widely studied to understand the risk of location privacy. In particular, a Markov chain model is known as one of the most successful approaches. In this approach, the attacker divides an area into some regions (or extracts some POIs), and partitions time at a fixed interval (such as 30 minutes). Then it trains a transition matrix for each target user. Using these matrices, the attacker can de-anonymize mobility traces or infer actual locations with high accuracy, when the amount of training data is very large.

However, the training data can be sparsely distributed over time in practice. Many users disclose only a small number of locations via SNS, and they disclose not continuously but sporadically. For example, they may use only one or two location check-ins per day, per week, or per month. In such cases, the number of training locations can be small, and there are many missing locations in the training trace. As a result, training a matrix is a very challenging task.

In this work, we show that location privacy attacks can be a threat even in this case. Specifically, we propose EMTF, a method to train transition matrices by incorporating tensor factorization into the EM algorithm. We perform experiments using real datasets, and show that the proposed method outperforms a random guess even when there is only one location composed of ten locations and each location is missing with probability 80%. (This work was published in PoPETs2017).

## Towards Using Differential Privacy as a Building Block for Privacy-Preserving Algorithms

Tristan Allard, Univ Rennes, CNRS, IRISA

Differential privacy has been originally proposed for disclosing agregate information over personal datasets without jeopardizing individuals' privacy. It is usually enforced by carefuly perturbing the agregate functions in order to make the output distributions almost insensitive to the impact of any single individual value. Differential privacy exhibits interesting properties, such as, *e.g.,* self-composability or transformation invariance. In this talk, I advocated the use of differentially private functions as building blocks for designing privacy-preserving algorithms over personal data. Indeed, contrary to encryption schemes, differentially private functions give the opportunity to perform parts of the computation over cleartext information (although perturbed) without giving away sound privacy guarantees. I started by overviewing three recent works that follow this approach - (1) a privacy-preserving distributed k-means algorithm called Chiaroscuro, (2) a privacy-preserving task assignment algorithm, and (3) a privacy-preserving index over encrypted data called PINED-RQ. I then synthesized the lessons learned from these works, with a special focus on the resulting security models and on the privacy/performance/quality tradeoffs. Finally, I concluded the talk by discussing the pros and cons of this approach and by outlining exciting open issues.

## Diffix

Paul Francis, Max Planck Institute for Software Systems

Diffix is a data anonymization mechanism developed as joint research between MPI-SWS and Aircloak. Diffix is designed to be general purpose, easy to use, strongly anonymous, and to provide useful data analytics. Diffix takes an empirical (rather than formal) approach to anonyzation. This opens up the design space and gives Diffix better utility than formal approaches like Differential Privacy and K-anonymity, albiet at the expense of mathematical certainty of the anonymity properties. Diffix combines several new and old mechanisms to achieve anonymity, the new being primarily layered sticky noise. Diffix uses SQL as the query language, but limits query semantics in order to prevent a variety of attacks. The remaining semantics, however, is still useful for a wide variety of analytic tasks, as evidenced by the fact that Diffix is being used in industry as part of the Aircloak anonymization product.

## PWSCUP 2017 Report on anonymization Competition

Hiroaki Kikuchi, Meiji University

PWS Cup is a open-style competition for data anonymization and re-identification risk. It has been held in Japan since 2015 for three times. The purpose of the competition is to develop a reliable algorithm for data anonymization and to evaluate the risk of anonymized data to be re-identified. In this talk, the dataset used in the competition, the basic rule for game, the criteria for evaluating utility and security of the anonymized data in details. In addition to the competition design, the analysis of anonymization strategies based on the submitted data to the past competition is reported.

One of the hot topic in the talk was the concern of the Japanese regulation of data anonymization, which is the amended Act on the Protection of Personal Information has been enforced fully on May 30, 2017. Under the new regulation, a new notion named *De-identified information* (anonymously processed data) was introduced. However, the followings are inconsistent with the notions commonly known as de-identification.

- The algorithm of anonymization is hidden.
- The pseudonymization is approved as one of the methods.
- The definition has some redundant descriptions.
- Biometric (gnomic) data is classified as one of personal data.

## Anonymization Bounty Program

Paul Francis, Max Planck Institute for Software Systems

As part of its research program on practical anonymization, MPI-SWS, working with its research partner Aircloak GmbH, established a bounty program to test the anonymity of our anonymization mechanism Diffix. This talk describes that program and the challenges it faces. The bounty program pays participants to find weaknesses in Diffix. To determine how much to pay, the bounty program developed a measure of anonymity, called the PCK score, that measures the effectiveness of attacks. The amount of payout is tied to the PCK score. The bounty program also defined a set of attacks based on the three criteria proposed by the European Union Working Party 29, namely singling out, linkability, and inference. The payout system is designed to encourge even weak attacks that may well not be a concern in practice, but allows us to learn more about the weaknesses of Diffix.

## Big Data Anonymization Requirements vs Privacy Models

Josep Domingo-Ferrer, Universitat Rovira i Virgili

Big data have come true with the new millennium. Specifically, personally identifiable big data result from the traces that any human activity leaves. To respect the current privacy regulations, and in particular the new European General Data Protection Regulation, personally identifiable information must be anonymized before it is released or exchanged. Anonymized big data should not allow unequivocal reconstruction of any subject's profile. At the same time, anonymized big data that are published should yield results similar to those obtained on the original big data *for a broad range of exploratory analyses.*

Privacy models are *ex ante* privacy guarantees to be met by anonymization procedures. A privacy model for big data should satisfy at least the following conditions:

1. Enforcing it with reasonable privacy parameters should be compatible with preserving exploratory utility;

2. It should be composable, that is, pooling data sources that satisfy the privacy model should yield pooled data that satisfy the model.

3. It should be enforceable with linear or quasi-linear cost.

4. It should allow some degree of linkability between similar individuals across anonymized data sources.

In the first part of this talk, we examine how well the two main privacy models in use ($k$-anonymity and $\epsilon$-differential privacy) satisfy the above requirements:

- For $k$-anonymity to be composable, the controllers sharing subjects must coordinate or follow suitable strategies. There are quasi-linear heuristics for $k$-anonymity. Linkability is possible at least at the $k$-anonymous class level. With some coordination effort, $k$-anonymity is a reasonable option to anonymize big data.

- $\epsilon$-Differential privacy has good composability properties, which may be suitable to anonymize dynamic data. It has also a low computational cost, which may be suitable for very large data sets. However, linkability across differentially private data sets is only feasible if the data sets share unaltered attributes. The main problem with $\epsilon$-differential privacy is that it does not provide significant utility for exploratory analyses unless the $\epsilon$ parameter is quite large.

Thus, none of the above two privacy models is entirely satisfactory, although $k$-anonymity seems more amenable to big data protection.

In the second part of this talk, we examine the connections between the following privacy models: randomized response, post-randomization, $\epsilon$-differential privacy and $t$-closeness (the latter being an extension of $k$-anonymity). They turn out to share common underlying principles, namely deniability and permutation. In particular, deniability is useful to understand the poor privacy guarantees offered by $\epsilon$-differential privacy when large $\epsilon$ values (say $\epsilon > 1$) are taken in quest for utility preservation. Furthermore, the highlighted connections might result in synergies between the analyzed privacy models in order to tackle big data anonymization.

## Extension of Differential Privacy to Distribution Obfuscation

Yusuke Kawamoto, AIST

We propose a privacy notion, called *distribution privacy*, to formally model the privacy of the probability distribution as an extension of differential privacy to distributions. Roughly speaking, a privacy mechanism with the distribution privacy obfuscates a probability distribution so that the attacker cannot significantly gain any information on the distribution by observing the outputs of the privacy mechanism.

We show that when we make two distributions less distinguishable in terms of the distribution privacy with the approximate max divergence $D_\infty^\delta$ (resp. the $f$-divergence $D_f$), then the amount of added noise should be proportional to the $\infty$-Wasserstein metric (resp. the Earth mover's distance) between the distributions. Then we propose a privacy mechanism, called the sampling-then-perturbing (STP) mechanism, that publishes probability distributions while guaranteeing the distribution privacy. Finally, we demonstrate an example of obfuscation of energy consumption distributions.

## Continual counting under local differential privacy

Jun Sakuma, University of Tsukuba

In the talk, two topics are mentioned: continual count reporting with local differential privacy and reconstruction of private training images from deep neural networks. In the first topic, we consider the problem of continual counting under the guarantee of local differential privacy. Suppose each data provider holds a binary state 0 or 1 that varies over time. Continual counting is a problem to calculate the number of data providers having state 1 at each round $t = 1, \ldots, T$ continually. In the local privacy setting, data providers report their state to the data collector via a privacy mechanism at each round with enforcing local privacy. In the talk, we present a local differential privacy mechanism for continual counting, m-shot reporting, and its utility analysis. In m-shot reporting, each user submits users' binary states only at $m(< T)$ rounds instead of reporting at every round. We show that m-shot reporting achieves better utility compared to 1-shot reporting and T-shot reporting when m is appropriately chosen.

In the second topic, we suppose a deep neural network model for image recognition is trained with images that need to be kept private for privacy or confidentiality reasons. In this situation, the problem to be considered is whether or not an adversary can reconstruct the private samples if the model is given to the adversary. With the recent progress of deep neural networks, automatic generation of photorealistic images by generative adversarial networks (GANs) has improved dramatically over the last few years. We introduce PreImage-GAN for this type of reconstruction attack on deep neural networks for image recognition and show that private training images can be reconstructed from the model with a high-level of quality experimentally.

## Anonymizing Graphs

Tamir Tassa, The Open University of Israel

Social networks are structures that describe a set of individuals and the relations between them. In their most basic form they are modeled by a graph, which describes the social relations, but they may include additional information on the individuals in the underlying society. Social networks are of interest to researchers from many disciplines, be it sociology, psychology, market research, or epidemiology. However, the data in such social networks cannot be released as is, since it might contain sensitive information. Therefore, it is needed to anonymize the data prior to its publication in order to address the need to respect the privacy of the individuals whose sensitive information is included in the data.

The first part of the presentation describes the main approaches for anonymizing networks. The methods of anonymizing networks fall into three main categories. The methods of the first category provide k-anonymity via a deterministic procedure of edge additions or deletions. The methods of the second category add noise to the data, in the form of random additions, deletions or switching of edges, in order to prevent adversaries from identifying their target in the network, or inferring the existence of links between nodes. The methods of the third category do not alter the graph data like the methods of the two previous categories; instead, they cluster together nodes into super-nodes of size at least k, where k is the required anonymity parameter, and then publish the graph data in that coarse resolution.

In the lion's part of the presentation we present algorithms for anonymizing network data by means of sequential clustering. We consider social network data in which the nodes are described by some quasi-identifiers (e.g. age, gender, location). The output of our algorithms provides the graph structure over a clustering of the nodes into super-nodes of size at least k, and a corresponding generalization of the quasi-identifiers that are present in each super-node. We offer two variants of an anonymization algorithm which is based on sequential clustering. We present experimental results that demonstrate the advantage offered by our algorithms, in terms of minimizing information loss, with respect to other algorithms. Then, we turn our attention the case in which the network data is distributed between several data holders. The goal is to arrive at an anonymized view of the unified network without revealing to any of the data holders information about links between nodes that are controlled by other data holders. Our study is the first one that considers the problem of privacy-preserving publication of network data in the distributed setting.

## PETS Competition Meeting minutes

Hiroaki Kikuchi, Meiji University

**Room** Room 209, Shonan Village Center

**Date** March 6th, 2018, 20:00-22:00

**Participants** Prof. Hiroaki Kikuchi, Prof. Josep Domingo-Ferrer, Prof. Sebastien Gambs, Dr. Koki Hamada, Prof. Chiemi Watanabe, Prof. Hiroshi Nakagawa, Prof. Chris Clifton, Dr. Bradley Malin, Dr. Takao Murakami, Mr. Antoine Laurent, Mr. Antoine Boutet, Mr. Axel Michel, Prof. Tristan Allard

We pointed out the following issues:

1. Rules  should we adopt PWS Cup rules or revise it?

2. Definition of re-identification

3. Open-source style vs. data only style

4. Pseudonym  do you use pseudonymization?

5. Schedule
6. Platform  PWSCUP platform (SQL, PHP, python, ruby) vs. kaggle
7. Definition of utility  refine list of functions
8. Definition of re-identification  refine the list of sample code to attack
9. Partial knowledge vs. max-knowledge adversary
10. Trust (to participants)  when we used the partial knowledge, some team is suspected to cheat their score
11. Dataset  Online retail dataset vs. Open Montreal trajectory dataset
12. New rule, Submit a part of records with confidence and have penalty for false positive.
13. Contact to PETS committee

Prof. Kikuchi proposed the plan of international competition to be held in PETS 2018. The proposed style is based on that of PWSCup 2017 (Japanese domestic competition) that used the partial knowledge background knowledge with some sampling ratios, 25, 50, 75 and 100%. The participants submit the anonymized data with their algorithm described in some pseudocode (open algorithm without source code). New re-identification based on the fraction of correctly estimated pseudonyms out of the total number of pseudonyms. The partial knowledge model was problematic and suggested to drop.

Prof. Clifton reported past similar competition sponsored by NSF. The dataset is based on the true patient data but swapped and mixed with at least two persons so that the modified dataset preserves the statistics in terms of demographic information. The data is closed within the participants who make some contract with the NSF.

We discussed the pros and cons of the open source-code. Prof. Gambs claims that the source code should be submitted since PETS is academic conference and transparency is significant to make the algorithm to be trusted. If we publish the proceedings of paper at PETS, participants claim their contribution.

We found the following issues in open

- Legal responsibility. If source code is made public, we somehow assure that the code must run and does not include any vulnerability. We must provide the set of open-source license (Creative Commons, UCB, copyleft).
- Paper format. Maximum pages and the format should be defined. For peer reviewing, the criteria to review should be defined, too.
- Timing to reveal evaluation results. If the re-identification rate is given to adversary imidiatetely after he submits the estimated identification, by iterating multiple times and comparing the several results, he eventually identify all records. So, the frequency of resulting re-identification rate should be restricted, e.g., one-time, one per day, one after due day, hidden to the final phase.

Prof. Clifton suggested the following principles for competition.

Dont use text.
Dont try to link real external data.
Do use large data.

Finally, we come to the conclusion as follows:

We use the Online retail dataset as same as PWSCup. The sampling rate should be large enough no to process manually. The number of users to be anonymized is at most about 5,000. The Montreal open trajectory data is reserved as future competition. Participant submits the anonymized data and can submit the source code after he/she agrees the result of evaluation. Participant who discloses the source code has some advantage in terms of final evaluation. Participant must submit the paper that describes the algorithm to anonymize the data and to prevent it to be identified. The paper is reviewed and selected to the final presentation. The program committee review the submitted papers. They may invite some teams to the final presentation based on the evaluation score. Participant tries to estimate the mapping between the true identities and the pseudonyms based on the evaluation score and the paper describing algorithms. They dont submit the source code to identify the user from the anonymized data. The re-identification rate is defined as a fraction of the correctly estimated pseudonyms over the total number of pseudonyms. It is NOT the same definition used in the Japanese version. The selected paper and invited authors give talks at the workshop in PETS 20118. The post-proceedings will be published after the PETS will be held. We dont have the on-site re-identification phrase. We assume the maximum-knowledge model to the background knowledge of the adversaries. Namely, participant tries to identify the anonymized records based on the original dataset. We use any platform for competition. The system used in the Japanese PWS Cup, which uses SQL, PHP, python and ruby, is one of the candidate. The utility of the anonymized data is defined as a maximum of some functions. The simplified versions of PWS Cup are candidates. The privacy of the anonymized data is evaluated as a maximum of some pre-defined re-identification algorithms and all submitted re-identification data. (Option) Participant can submit only records to be identified with confident. Re-identification score is defined considering both true positive (plus) and false positive (minus). We will submit the revised proposal of the competition to PETS board committee. The schedule is as follows:

| | |
|---|---|
| Call for participate | Mar. XX Apr. 27 |
| Submit anonymized data | May XX June 1st |
| Submit source code and paper describing algorithm | May XX June 14 |
| Submit re-identification data | June 15 July 6 |
| Final presentation | July 23th |
| Post-proceeding | September |

So, here is the to-do list toward our competition.

- Survey past competition with similar purpose and styles.
- Revise the proposal of competition and propose to PETS committee
- Prepare the platform to be customized for new rule.
- Organize the program committee to review papers.

## A Privacy-Preserving Mechanism for Requesting Location Data Provider with Wi-Fi Access Points

Antoine Boutet, Insa-Lyon

With the democratization of mobile devices embedding different positioning capabilities, the location of users is now collected to track the location of users. When used for behavioral profiling, this tracking for enhancing raises more and more privacy concerns. Depending on the permissions, mobile applications can get a fine-grained users location from the GPS or a coarse-grained location by requesting location data provider with surrounding Wi-Fi access points for instance. While using the GPS does not rely on external untrusted party, requesting a location data provider clearly exposes the location of users. Whereas location privacy has been an active research field this last decade, most of the contributions are performed on GPS-based data, and it is not clear how to efficiently protect Wi-Fi-based positioning to preserve the users privacy. In this talk, I propose a novel solution to preserve users privacy from curious location data providers when requesting users location from Wi-Fi while supporting high-utility. The key idea behind this online approach is to combine a random sampling (for controlling the quantity of revealed information) and a obfuscation scheme (for ensuring privacy-preserving information disclosure). I report an evaluation of the solution with a real dataset of mobility traces collected through multiple sensors and show that the proposed approach provides a trade-off between privacy (i.e., avoiding to reveal its true location) and utility (i.e., still benefiting from services such as places recommendation) fully controllable by the users.