Executive summary: k-randomization

Synopsis

Randomized response techniques, such as RAPPOR, are useful for applications where a collection of multivariate client records exists on backend servers for repeated analysis without requiring users to trust the receiving end with their private information. In order to preserve privacy, each client adds random noise to its record before submission, in such a way that reasonable estimates can be computed from the entire collection of client records. Although the privacy requirements of this scheme result in inherently noisy estimates, the k-randomization technique provides a way to increase the precision of RRT-based estimates subject to a specified level of privacy.

Key features

- k-randomization increases precision by having clients submit multiple randomized replicates of their data
- The privacy risk around the identifiability of individual users is calibrated by taking into account the entire noisy collection of client records
- The privacy risk around the collection of multiple randomized copies of client data is mitigated by the using infrastructure to anonymize (unlink) records records from the same client
- A mechanism is provided to adjust the randomization according to the desired tradeoff between randomization noise and estimation precision subject to an overall differential privacy budget

k-randomization should not be thought of as an alternative data collection technique. Rather, in situations where RRT is an appropriate technique for privacy-preserving data collection, k-randomization provides a way to improve on current common methodology such as RAPPOR, subject to set privacy guarantees. It is particularly useful in cases where RAPPOR performs poorly, such as collection of multivariate user records.

Motivation

The motivation for k-randomization is to address the shortcomings of RRT techniques like RAPPOR in collecting multivariate user data. The two main issues are:

- Reliable inference is possible only for a limited set of values occurring most frequently across the collection of user records
- Both precision and privacy guarantees deteriorate rapidly as the dimensionality of the data increases

k-randomization reduces estimation error via replication, and improves on the traditional differential privacy guarantee through a more sophisticated privacy requirement tailored to a collection of independently randomized records.

In general, the goal of k-randomization is to improve on existing techniques within the paradigm of clients reporting individually randomized records. The underlying motivation for working under this paradigm is twofold:

- to be able to measure privacy guarantees using differential privacy, and
- to apply privacy protection before the data record leaves the client.

RAPPOR is currently one of most important techniques in this class. By the nature of the intention to preserve privacy by modifying user records, a certain degree of error is introduced into any inferences drawn from a dataset of such records. Thus, it should be considered to complement, rather than compete with, data collection methods falling outside this class. Homomorphic encryption is one such complementary method which, although lacking any differential privacy guarantee, can be used to compute exact additive aggregates from univariate data.

Summary of Theoretical Background

- We consider user data encoded as a vector of bit values
- Applying traditional differential privacy directly to multitude of multivariate randomized records yields unsatisfactory guarantees since it is based on the worst-case scenario
- Taking advantage of the inherent structure in our application (randomized response applied to multiple bit vectors) lets us reduce the differential privacy bound over all but a negligible set of outcomes
- We obtain a mathematical relationship between the parameters on which the method depends. This can be used to tune parameter values subject to desired tradeoffs. The parameters are:
    - size of the collection ***N*** (number of clients),
    - randomization noise parameter ***q*** (probability of lying),
    - dimensionality of the dataset ***L*** (length of the client vectors),
    - number of replicates ***k*** (number of times clients re-randomize and resubmit their data), and
    - privacy ratio λ, in terms of which the differential privacy criterion is expressed.

$$q \geq \frac{1}{2}\left(1 - \sqrt{1 - \frac{4}{3 + \sqrt[L]{1 + (1 - \frac{1}{\lambda})^2 \cdot \frac{N \cdot k}{9}}}}\right)$$

Calibration

In a realistic data collection setup three restrictions have to be met:

- privacy level needs to be enforced
- adequate precision of estimates needs to be provided
- number of randomizations ***k*** should be small to minimize infrastructure cost.

Precision is determined by the deviation between measurements based on randomized records and those computed from the original values. The error ***e*** (expressed as % of ***N)*** is related to the RRT parameters in the following way:

$$\frac{kN}{pq} > L\left(\frac{3}{e}\right)^2$$

Again, the above formula assumes that full multivariate analysis is performed, in which case the length of the bit-vector has to be considered. For marginal or pair-wise join estimations, or for conditionally independent data sets, ***L*** reduces to a small number.

Given the system of inequalities that govern precision and privacy, the optimal noise parameter ***q*** is chosen to minimize ***k***. The system of inequalities is provided below:

$$\underbrace{q \geq \frac{1}{2}\left(1 - \sqrt{1 - \frac{4}{3 + \sqrt[L]{1 + (1 - \frac{1}{\lambda})^2 \cdot \frac{N \cdot k}{9}}}}\right)}_{privacy}$$

$$\underbrace{\frac{kN}{(1-q)q} > L\left(\frac{3}{e}\right)^2}_{precision}$$

If the data are high-dimensional, the worse case scenario for privacy would require very high randomization noise and a large number of repetitions. In practice, however, datasets do not usually exhibit the worst-case configuration, allowing a significant reduction of randomization noise relative to the worst case. A perfect example is a set of mutually exclusive bit-vectors, where the real length could range in thousands while for the privacy (and precision) bounds ***L*** will be 2. For large user-base, diverse distributions of user data, and possible method enhancements (not included in the summary), we believe ***k*** will be small for a majority of practical use cases.

Applications

Numerous applications require analytical capacity far beyond adding single bits: quantifying and clustering various user audiences with respect to particular feature set, as well as any personalization initiative would require building statistical profiles of user audiences, which is impossible without bit-vectors being preserved. Improving a user search experience by collecting head queries is yet another example of valuable privacy-protecting data extraction impossible unless longer bit-vectors are collected.

Having multivariate datasets enables extraction of associations between variables. Such data-sets are not much different from a regular database table allowing for arbitrary SQL queries and numerical methods. They contains noise, and query results are inexact to protect privacy, but sufficiently precise to meet the business needs.

Publishability

We believe that publishing the research will be beneficial to the privacy community at large and will have significant applications internet-wide.

We have conducted an extensive prior art search and requested pointers to related research from RAPPOR team at Google and Stanford. As we failed to find a treatment of privacy in the case of anonymized collections of randomized user records, we launched an independent research project to investigate the privacy implications of noise reduction through repeated randomization. A large body of privacy research was conducted before 2004 on extracting association rules from randomized collections, but no treatment was given to the differential privacy aspect of it. The differential privacy development (pioneered by C. Dwork in 2006) focuses on randomizing queries and aggregations applied to a database of original records, rather than the effect of randomizing the records individually. Furthermore, the idea that randomization could be repeated without significant privacy loss seems to be unique, as it relies on the idea that the differential privacy criterion can be adjusted to take advantage of the specific structure of our randomized collections. Nevertheless, performing a deeper literature survey focusing on research subsequent to RAPPOR is definitely required.

Novelty

- Formal treatment of differential privacy and implications in the case of synthetic collections, where each record is randomized independently
- Optimization of the differential privacy criterion by taking into account the likelihood of observing the specific randomized collections over which it is computed
- Differential privacy bounds in the case of repeated randomization

- Expression of differential privacy and in terms of the a notion of anonymity provided by being a member of a large randomized collection of records, and its relation to protection against linking replicate records.
- Methodology for tuning parameters in relation to engineering decisions in a practical setting

Required work

- finish differential privacy proofs for ***L***>1 (buckets idea)
- justify a use of unit-vector collection for worse-case scenarios (local DP maximization issue)
- verify privacy-breach ratio derivation, and show that protecting for privacy-breach ratio also protects differential privacy guarantees for high cardinality case
- show merits of non worse-case scenario anonymity and how it reduces privacy bounds
- contact deep prior arts search, especially in RAPPOR domain and underline the novelty of the approach

Further reading

- Comparison of various methods for privacy protection data collection to k-randomization
- Paper draft