Personal Privacy vs Population Privacy: Learning to Attack Anonymization

Graham Cormode, AT&T Labs–Research

Overview

There has been much recent work on anonymizing data:

- *Syntactic Privacy* [1998–]: k-anonymity, ℓ-diversity etc. [XT06].
  Much initial interest, but only informal security guarantees.
  Attacks exist to allow inference of private information, e.g. [Kif09]
- *Semantic Privacy* [2005–]: Differential privacy [Dwo06].
  Gives a stronger guarantee: the output distribution is insensitive to the information of any one individual.

This work studies an "attack" on differential privacy: we show how to infer information about an individual with high probability.

This does not violate any properties of differential privacy.

Rather, it shows that in some cases, the guarantees are insufficient.

The main message is that just providing differential privacy does not cure all privacy problems.

Differential Privacy

Data Model. The input is a table of tuples T, each tuple $t \in T$ represents an individual.

We distinguish one *sensitive attribute* (SA) e.g. disease, salary band.

The remaining m attributes are *quasi-identifiers* (QI) e.g. age, sex.

Differential Privacy. For any 2 inputs T, T' differing in one individual (so $\|T - T'\|_1 = 1$) and any subset of outputs S, an algorithm A is ε-differentially private (ε-DP) if:

$$\Pr[A(T) \in S] \leq e^{\varepsilon} \Pr[A(T') \in S]$$

Laplace Mechanism is a simple algorithm for ε-DP on count queries.

A query $q(T)$ asks for k counts of subsets of T.

Define sensitivity $\sigma(q)$ as $\max_{T,T':\|T-T'\|_1=1} \|q(T) - q(T')\|_1$
— the maximum impact one individual can have on the query.

Outputting $q(T) + \mathrm{Lap}^k(\sigma(q)/\varepsilon)$ is ε-differentially private,
where $\mathrm{Lap}^k(\lambda)$ is the k-dimensional Laplace dbn with param λ.

Attacks and Inference

Attacks aim to infer the SA for an individual given their QI.

The success of the attack is measured as the fraction of tuples for which the inference is correct.

Naive Bayes Classifier. Given QI values t, we predict its SA $\hat{s}(t)$ via:

$$\hat{s}(t) = \arg\max_{s \in SA} \Pr[s] \prod_{i=1}^{m} \Pr[t_i|s]$$

The conditional probabilities can be estimated via simple queries:

$$\Pr[t_i|s] = \frac{\Pr[t_i \cap s]}{\Pr[s]} \approx \frac{|\{r \in T : r_i = t_i \cap r_s = s\}|}{|\{r \in T : r_s = s|\}}$$

Differentially Private Classifiers

We build a classifier to predict an SA, as in [BS08, Kif09].

We choose Naive Bayes, due to its simplicity, and its low sensitivity.

To instantiate the classifier, we need counts of values present in T^i, the ith column of table T:

$$\forall s \in SA, 1 \leq i \leq m, v \in T^i : |\{r \in T : r_i = v \cap r_s = s\}|$$
$$\text{and} \quad \forall s \in SA : |\{r \in T : r_s = s\}|$$

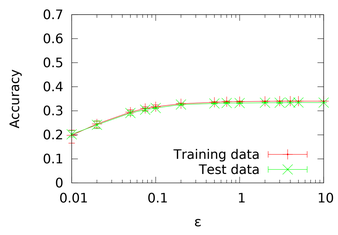
We can obtain noisy counts via the Laplace mechanism.
Observe that sensitivity σ is very low: just $m + 1$.

We round-up zero counts, and apply a standard correction to obtain:

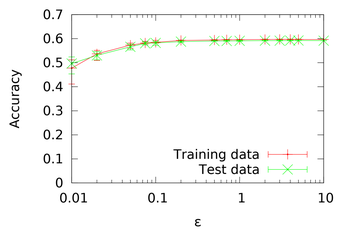
$$\widehat{\Pr}[t_i|s] := \frac{1 + \max(0, |\{r \in T : r_i = t_i \cap r_s = s\}|)}{\sum_{t \in T_i} 1 + \max(0, |\{r \in T : r_i = t \cap r_s = s\}|)}.$$

Effectiveness of Classifier

We apply this attack to the UCI 'Adult' dataset (a common target).



Accuracy on 'occupation'



Accuracy on 'marital-status'

Results. The "naive" attack is very successful at inferring SAs.

Prior work uses $\varepsilon = 0.1$–1. Attack is accurate even at $\varepsilon = 0.01$.

Other results show that the attack is over 80% accurate on those tuples where its confidence is above 80% (up to 25% of tuples).

The attack is trivial to implement: full running time < 3 seconds.

Attack Discussion

The attack proceeds by learning properties of a population, not any one individual.

These properties are hard to hide, robust to addition of a lot of noise.

The population model is very good at predicting an individual's SA!

An individual does not need to contribute to the data to be at risk: accuracy on withheld test data was similar to on training set.

Should we worry? The "attack" only learns true global facts.

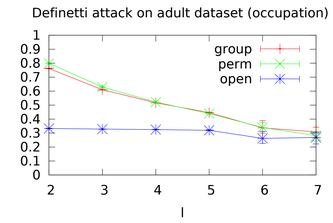
Consider a costly, complex medical study: releasing the results can let an attacker to draw conclusions about your health "for free".
The attacker could not do this without the released data.

Can vary the choice of classifier:
— Group multiple attributes and discretize continuous attributes
— Choose other classifiers–but must have low sensitivity

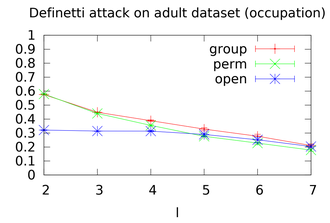
Comparison to deFinetti Attack

deFinetti attack on ℓ-diverse data also builds a classifier [Kif09].
We can compare accuracy to that on differential privacy.

Three ways to use the classifier on data published in groups:
open — only use the classifier directly on the SA values
group — only consider SA values in the same group as the target
perm — find most likely permutation mapping in each group.



Small groups



Larger groups

Results. Accuracy degrades fast as group size increases.
Threat is actually quite similar to the ε-DP case.
Extra power comes from using the group structure of the output.

Discussion

Attacks are possible by learning coarse properties of a population.

This affects both syntactic and semantic privacy to a similar degree.

Tough to separate privacy from utility: to some this knowledge threatens privacy, to others it is the reason for releasing the data!

Example: "people who do X are at high risk for disease Y"
— *helpful* if you want to modify behavior to reduce your risk.
— *worrying* if insurance refuses to cover you because you do X!

One subtlety: early attacks on syntactic data gave deterministic facts.
Now, attacks give only probabilistic inference.
What accuracy probability are you happy with: 50%? 10%? 1%?

Sometimes, false inferences are just as dangerous as correct ones!
Can lead to e.g. higher insurance rates or denial of coverage.

Conclusions: Differential privacy still operates as promised, but we must ensure we understand fully what those promises are.

It is certainly not sufficient to just apply differential privacy and then claim that all privacy problems are now solved!

References

[BS08] Justin Brickell and Vitaly Shmatikov. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *ACM SIGKDD*, 2008.

[Dwo06] Cynthia Dwork. Differential privacy. In *ICALP*, 2006.

[Kif09] Dan Kifer. Attacks on privacy and deFinetti's theorem. In *ACM SIGMOD*, 2009.

[XT06] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, 2006.