Differential Privacy (Part I)
Computing on personal data

Individuals have lots of interesting data

and we would like to compute on it
Which kind of data?
Which computations?

• statistical correlations
  • genotype/phenotype associations
  • correlating medical outcomes with risk factors or events
• aggregate statistics
  • web analytics
• identification of events/outliers
  • intrusion detection
  • disease outbreaks
• data-mining/learning tasks
  • use customers’ data to update strategies
Ok, but we can compute on anonymised data, i.e., not including *personally identifiable information*… that should be fine, right?
First person identified from AOL Data: Thelma Arnold

MICHAEL ARRINGTON

Wednesday, August 9th, 2006

On Sunday the news broke that AOL purposefully released 20 million partially anonymized search queries. On Monday AOL apologized, and later that evening the first web interface to the data went up.

Today the first person was positively identified from the data — Thelma Arnold, a 62-year-old widow who lives in Lilburn, Georgia.

Based on searches ranging from "numb fingers" to "60 single men" to "dog that urinates on everything," the New York Times was able to quickly determine and confirm her identity. Ms Arnold is AOL searcher no. 4417749.

Ms Arnold commented: "My goodness, it's my whole personal life...I had no idea somebody was looking over my shoulder."

AOL replied: "We apologize specifically to her...There is not a whole lot we can do."

Tags: AOL
De-anonymize Netflix data [A. Narayanan and V. Shmatikov, S&P’08]

- Netflix released its database as part of $1 million Netflix Prize, a challenge to the world’s researchers to improve the rental firm’s movie recommendation system
- **Sanitization**: personal identities removed
- Problem, **sparsity of data**: with large probability, no two profiles are similar up to $\epsilon$. In Netflix data, no two records are similar more than 50%
- If the profile can be matched up to 50% to a profile in IMDB, then the adversary knows with good chance the true identity of the profile
- In this work, efficient random algorithm to break privacy
Personally identifiable information

What Information is “Personally Identifiable”? 

Mr. X lives in ZIP code 02138 and was born July 31, 1945.

These facts about him were included in an anonymized medical record released to the public. Sounds like Mr. X is pretty anonymous, right?

Not if you’re Latanya Sweeney, a Carnegie Mellon University computer science professor who showed in 1997 that this information was enough to pin down Mr. X’s more familiar identity -- William Weld, the governor of Massachusetts throughout the 1990s.

Gender, ZIP code, and birth date feel anonymous, but Prof. Sweeney was able to identify Governor Weld through them for two reasons. First, each of these facts about an individual (or other kinds of facts we might not usually think of as identifying) independently narrows down the population, so much so that the combination of (gender, ZIP code, birthdate) was unique for about 87% of the U.S. population. If you live in the United States, there’s an 87% chance that you don’t share all three of these attributes with any other U.S. resident.

Second, there may be particular data sources available (Sweeney used a Massachusetts voter registration database) that let people do searches to bootstrap what they know about someone in order to learn more -- including traditional identifiers like name and address.

In a very concrete sense, “anonymized” or “merely demographic” information about people may be neither. (And a web site that asks “anonymous” users for seemingly trivial information about themselves may be able to use that information to make a unique profile for an individual, or even look up that individual in other databases.)

Many contemporary privacy rules and debates center on the notion of “personally identifiable information” (PII). The PII concept is used by several legal regimes and many organizations’ privacy policies; generally, information that identifies a particular person is considered much more sensitive than information that does not. For instance,

- Federal telecommunications privacy laws use “individually identifiable information” (about a subscriber) as a basis for the category of protected information called Customer Proprietary Network Information (CPNI);
- Federal health privacy regulations use “individually identifiable health information” (about a patient) as a basis for the category called Protected Health Information (PHI);
- Federal financial privacy laws, the EU Data Protection Directive, and state privacy laws all employ similar terms and concepts;

and, in each case, facts deemed “personally identifiable” or “individually identifiable” may receive dramatically higher protections under these laws and regulations.

But research by Prof. Sweeney and other experts has demonstrated that surprisingly many facts, including those that seem quite innocuous, neutral, or “common”, could potentially identify an individual. Privacy law, mainly clinging to a traditional intuitive notion of identifiability, has largely not kept up with the technical reality.

https://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable
Personally identifiable information

From the Facebook privacy policy...

While you are allowing us to use the information we receive about you, you always own all of your information. Your trust is important to us, which is why we don't share information we receive about you with others unless we have:

- received your permission;
- given you notice, such as by telling you about it in this policy; or
- removed your name or any other personally identifying information from it.
Ok, but I do not want to release an entire dataset! I just want to compute some innocent statistics… that should be fine, right?
Actually not!

Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study

Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, Xiaoyong Zhou
Indiana University Bloomington
Bloomington, IN

Abstract

Genome-wide association studies (GWAS) aim at discovering the association between genetic variations, particularly single-nucleotide polymorphism (SNP), and common diseases, which is well recognized to be one of the most important and active areas in biomedical research. Also renowned is the privacy implication of such studies, which has been brought into the limelight by the recent attack proposed by Homer et al. Homer’s attack demonstrates that it is possible to identify a GWAS participant from the allele frequencies of a large number of SNPs. Such a threat, unfortunately, was found in our research to be significantly understated. In this paper, we show that individuals can actually be identified from even a relatively small set of statistics, as those routinely published in GWAS papers. We present two attacks. The first one extends Homer’s attack with a much more powerful test statistic, based on the correlations among different SNPs described by coefficient of determination ($r^2$). This attack can determine the presence of an individual from the statistics related to a couple of hundred SNPs. The second attack can lead to complete disclosure of hundreds of participants’ SNPs, through analyzing the information derived from published statistics. We also found that those attacks can succeed even when the precision of the statistics are low and part of data is missing. We evaluated our attacks on the real human genomes and concluded that such threats are completely realistic.

of genome-wide association study (GWAS) [7], a study that aims at discovering the association between human genes and common diseases. To this end, GWAS investigators determined the genotypes of two groups of participants, people with a disease (cases) and similar people without (controls) in an attempt to use statistical testing to identify genetic markers, typically single-nucleotide polymorphisms (SNP), that are associated to disease susceptibility genes [46]. If the variation of a SNP is found to be significantly higher in the case group than that in the control group, it is reported as a potential marker of the disease. Of great importance to such a study is privacy of the participants, whose sensitive information, personally identifiable genetic markers in particular, should not be leaked out without explicit consent. So far, this has been enforced through an informed consent from participants [9] and an agreement from investigators to ensure proper use of data according to the consent. Unfortunately, while this process prevents explicit misuse of participants’ DNA data, it turns out to be insufficient for deterring information leaks in a more implicit way. Particularly, this paper reports a surprising finding of our research: even the test statistics computed over a small set of SNPs, like those routinely published in GWAS papers, could reveal a substantial amount of genetic information about participants, and even lead to disclosure of their identities.

The inadequacy of privacy protection in current genome research has also been pointed out by other researchers. For example, Ma-
Database privacy

- Ad hoc solutions do not really work
- We need to formally reason about the problem...
What does it mean for a query to be privacy-preserving and how can we achieve that?
Blending into a crowd

- Intuition: “I am safe in a group of $k$ or more”
  - $k$ varies (3...6...100...10,000?)
- Why?
  - Privacy is “protection of being brought to the attention of others” [Gavison]
  - Rare property helps re-identify someone
Clustering-based definitions

- **k-anonymity**: attributes are suppressed or generalized until each row is identical to at least $k-1$ other rows.
  - At this point the database is said to be $k$-anonymous.
- Methods for achieving $k$-anonymity
  - **Suppression** - can replace individual attributes with a *
  - **Generalization** - replace individual attributes with a broader category (e.g., age 26 $\Rightarrow$ age [26-30])
- Purely syntactic definition of privacy
- What adversary does it apply to?
  - Does not consider adversaries with side information
  - Does not consider adversarial algorithm for making decisions (inference)
- Almost abandoned in the literature...
Notations

$D_i = \{d_i | i \in I\}$

$Q(D_i) = R$

$Q$ is the privatized query run on the data set and $R$ is the result released to the public.
What do we want?

- I would feel safe participating in the dataset if
  - I knew that my answer had no impact on the released results
  - I knew that any attacker looking at the published results R couldn’t learn (with any high probability) any new information about me personally [Dalenius 1977]
  - Analogous to semantic security for ciphertexts

- \( Q(D_{(I-me)}) = Q(D_I) \)
- \( \text{Prob}(\text{secret}(me) | R) = \text{Prob}(\text{secret}(me)) \)
Why can’t we have it?

✦ If individuals had no impact on the released results...then the results would have **no utility**!

✦ If R shows there is a strong trend in the dataset (everyone who smokes has a high risk of cancer), with high probability, that trend is true for any individual. Even if she does not participate in the dataset, it is just enough to know that she smokes!

Achieving either privacy or utility is easy, getting a meaningful trade-off is the real challenge!

\[ Q(D_I) = Q(D_{\emptyset}) \]

✦ \( \text{Prob}(\text{secret}(\text{me}) | R) > \text{Prob}(\text{secret}(\text{me})) \)
Even worse, if an attacker knows a function about me that’s dependent on general facts about the population:

- I am twice the average age
- I am in the minority gender

Then releasing just those general facts gives the attacker specific information about me. (Even if I don’t submit a survey!)

\[
\begin{align*}
\text{(age(me) = 2*mean\_age)} & \land \\
\text{(gender(me) \neq top\_gender)} & \land \\
\text{(mean\_age = 14)} & \land \\
\text{(top\_gender = F)} & \Rightarrow \\
\text{age(me)=28} & \land \text{gender(me)=M}
\end{align*}
\]
Impossibility result (informally)

- **Tentative definition:**
  
  For some definition of “privacy breach”,
  
  \( \forall \) distributions on databases,
  
  \( \forall \) adversaries \( A \),
  
  \( \exists \) \( A' \) such that
  
  \[
  Pr(A(San(DB)) = breach) - Pr(A'() = breach) \leq \epsilon
  \]

- **Result:** for reasonable “breach”, if \( San(DB) \) contains information about \( DB \), we can find an adversary that breaks this definition
Proof sketch (informally)

- Suppose DB is drawn uniformly random
- “Breach” is predicting a predicate $g(DB)$
- Adversary knows $H(DB)$, $H(H(DB) \ ; \ San(DB)) \oplus g(DB)$
  - $H$ is a suitable hash function
- By itself, the attacker’s knowledge does not leak anything about DB
- Together with $San(DB)$, it reveals $g(DB)$
Disappointing fact

- We can’t promise my data won’t affect the results
- We can’t promise that the attacker won’t be able to learn new information about me, given proper background information

What can we do?
One more try...

The chance that the sanitised released result will be $R$, is nearly the same whether or not I submitted my personal information.
Differential privacy

- Proposed by Cynthia Dwork in 2006
- **Intuition**: perturb the result (e.g., by adding noise) such that the chance that the perturbed result will be \( C \) is nearly the same, whether or not you submit your info
- **Challenge**: achieve privacy while minimising the utility loss
Differential privacy (cont’d)

A query mechanism $M$ is $\epsilon$-differentially private if, for any two adjacent databases $D$ and $D'$ (differing in just one entry) and $C \subseteq \text{range}(M)$

$$\Pr(M(D) \in C) \leq e^{\epsilon} \cdot Pr(M(D') \in C)$$
Sequential composition theorem

Let $M_i$ each provide $\epsilon_i$-differential privacy. The sequence of $M_i(X)$ provides $(\sum_i \epsilon_i)$-differential privacy.

- Privacy losses sum up
- **Privacy budget** = maximum tolerated privacy loss
- If the privacy budget is exhausted, then the server administrator acts according to the policy
  - answers the query and reports a warning
  - does not answer further queries
Sequential composition theorem

Let $M_i$ each provide $\epsilon_i$-differential privacy. The sequence of $M_i(X)$ provides $(\sum_i \epsilon_i)$-differential privacy.

- Result holds against active attacker (i.e., each query depends on the previous ones’ result)
- Result proved for a generalized definition of differential privacy [McSharry, Sigmod’09]
- $\oplus$ denotes symmetric difference

A query mechanism $M$ is differentially private if, for any two databases $D$ and $D'$ and $C \subseteq \text{range}(M)$

$$\Pr(M(D) \in C) \leq e^{\epsilon \cdot |D \oplus D'|} \cdot \Pr(M(D') \in C)$$
Parallel composition theorem

Let \( M_i \) each provide \( \epsilon \)-differential privacy. Let \( D_i \) be arbitrary disjoint subsets of the input domain \( \mathcal{D} \). The sequence of \( M_i(X \cap D_i) \) provides \( \epsilon \)-differential privacy.

- When queries are applied to disjoint subsets of the data, we can improve the bound.
- The ultimate privacy guarantee depends only on the worst of the guarantees of each analysis, not on the sum.
What about group privacy?

- Differential privacy protects one entry of the database
- What if we want to protect several entries?
- We consider databases differing in $c$ entries
- By inductive reasoning, we can see that the probability dilatation is bounded by $e^{c\epsilon}$ instead of $e^\epsilon$, i.e.,

$$
\Pr(M(D) \in C) \leq e^{c\cdot\epsilon} \cdot \Pr(M(D') \in C)
$$

- To get $\epsilon$-differential privacy for $c$ items, one has to protect each of them with $\epsilon/c$-differential privacy
- Exercise: prove it
Achieving differential privacy

- So far we focused on the definition itself
- The question now is, how can we make a certain query differentially private?
- We will consider first a generally applicable sanitization mechanism, the Laplace mechanism
The sensitivity of a function \( f : \mathcal{D} \to \mathbb{R} \) is defined as:

\[
\Delta f = \max_{D, D'} |f(D) - f(D')|
\]

for all adjacent \( D, D' \in \mathcal{D} \)

- **Sensitivity** measures how much the function amplifies the distance of the inputs.
- **Exercises:** what is the sensitivity of
  - counting queries (e.g., “how many patients in the database have diabetes”)?
  - “How old is the oldest patient in the database?”
Laplace distribution

- Denoted by Lap(b)
- Increasing b flattens the curve

\[ pr(z) = \frac{e^{-|z|/b}}{2b} \]

Variance = \(2b^2\)

Standard deviation \(\sigma = \sqrt{2b}\)
Laplace mechanism [Dwork et al., TCC’06]

Let \( f : \mathcal{D} \rightarrow \mathbb{R} \) be a function with sensitivity \( \Delta f \). Then \( g = f(X) + \text{Lap}(\frac{\Delta f}{\epsilon}) \) is \( \epsilon \)-differentially private.

- General sanitization mechanism
  - we have just to compute the sensitivity of the function
- Noise depends on \( f \) and \( \mathcal{E} \), not on the database!
- Remember how the Laplace distribution looks like: smaller sensitivity (and/or less privacy) means less distortion
- Exercise: how much noise do we have to add to sanitize the following question?
  - “How many people in the database are female?”