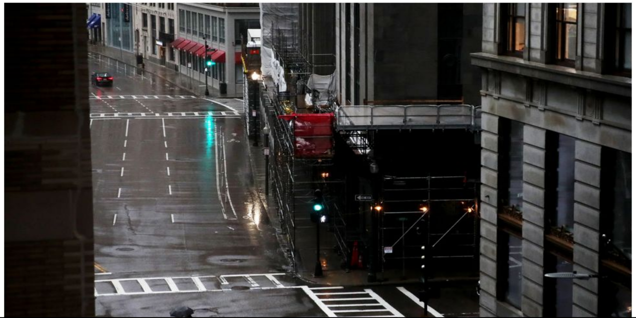
Monitoring Coronavirus and the need of privacy

- https://www.pepp-pt.org/
- https://tech.newstatesman.com/security/pepp-pt-vs-dp-3t-the-coronavirus-contact-tracing-privacy-debate-kicks-up-another-gear
- https://github.com/DP-3T/documents/blob/master/Security%20analysis/PEPP-PT_%20Data%20Protection%20Architechture%20-%20Security%20and%20privacy%20analysis.pdf
- https://tech.newstatesman.com/security/pepp-pt-vs-dp-3t-the-coronavirus-contact-tracing-privacy-debate-kicks-up-another-gear

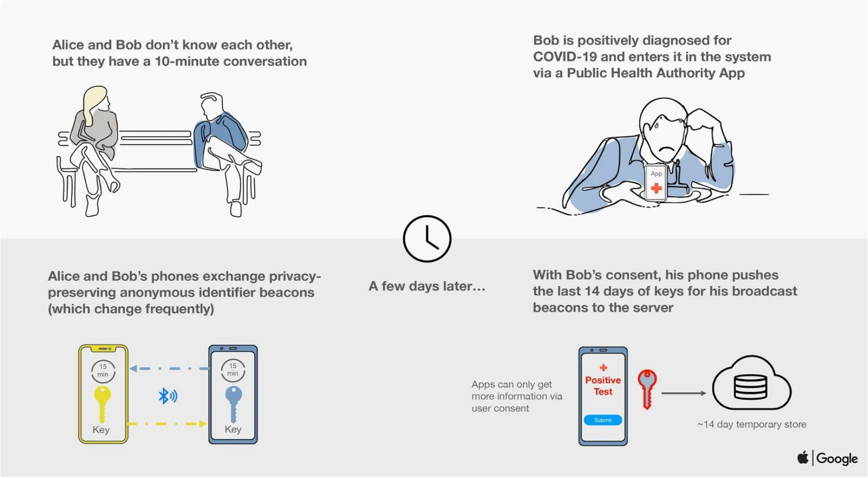
ANDY GREENBERG SECURITY 04.10.2020 03:37 PM

How Apple and Google Are Enabling Covid-19 Contact-Tracing

The tech giants have teamed up to use a Bluetooth-based framework to keep track of the spread of infections without compromising location privacy.

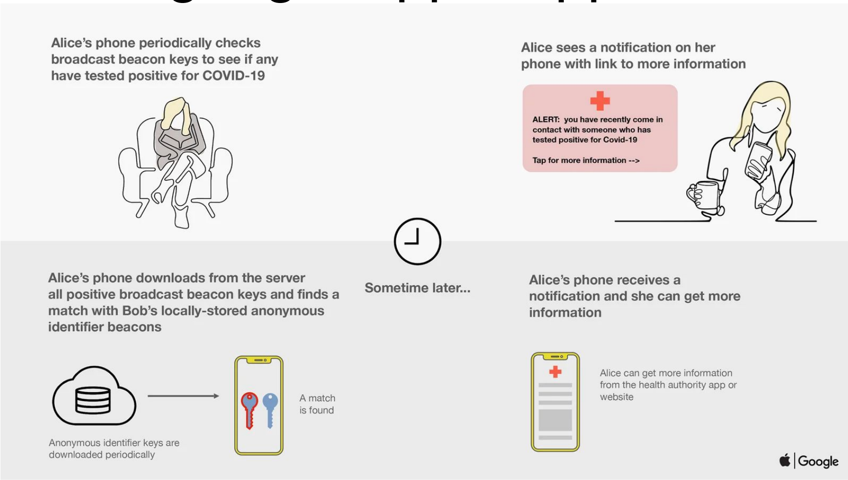


The google-apple approach



Source:

The google-apple approach



Source:

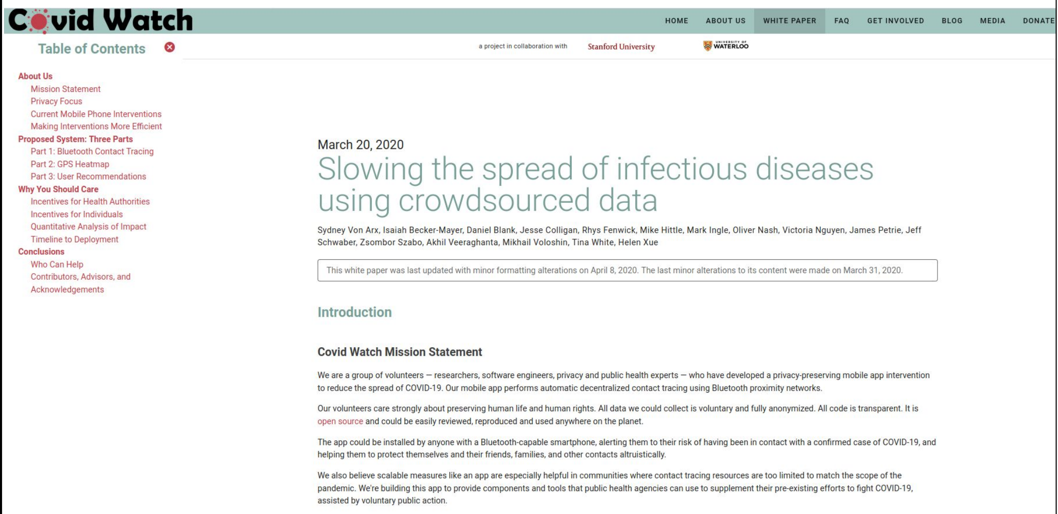

Covid Watch
HOME ABOUT US WHITE PAPER FAQ GET INVOLVED BLOG MEDIA DONATE

Table of Contents

a project in collaboration with Stanford University WATERLOO

About Us
Mission Statement
Privacy Focus
Current Mobile Phone Interventions
Making Interventions More Efficient
Proposed System: Three Parts
Part 1: Bluetooth Contact Tracing
Part 2: GPS Heatmap
Part 3: User Recommendations
Why You Should Care
Incentives for Health Authorities
Incentives for Individuals
Quantitative Analysis of Impact
Timeline to Deployment
Conclusions
Who Can Help
Contributors, Advisors, and Acknowledgements

March 20, 2020

Slowing the spread of infectious diseases using crowdsourced data

Sydney Von Arx, Isaiah Becker-Mayer, Daniel Blank, Jesse Colligan, Rhys Fenwick, Mike Hittle, Mark Ingle, Oliver Nash, Victoria Nguyen, James Petrie, Jeff Schwaber, Zsombor Szabo, Akhil Veeraghanta, Mikhail Voloshin, Tina White, Helen Xue

This white paper was last updated with minor formatting alterations on April 8, 2020. The last minor alterations to its content were made on March 31, 2020.

Introduction

Covid Watch Mission Statement

We are a group of volunteers — researchers, software engineers, privacy and public health experts — who have developed a privacy-preserving mobile app intervention to reduce the spread of COVID-19. Our mobile app performs automatic decentralized contact tracing using Bluetooth proximity networks.

Our volunteers care strongly about preserving human life and human rights. All data we could collect is voluntary and fully anonymized. All code is transparent. It is open source and could be easily reviewed, reproduced and used anywhere on the planet.

The app could be installed by anyone with a Bluetooth-capable smartphone, alerting them to their risk of having been in contact with a confirmed case of COVID-19, and helping them to protect themselves and their friends, families, and other contacts altruistically.

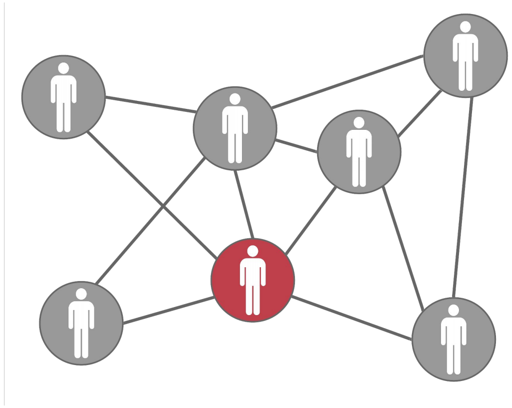
We also believe scalable measures like an app are especially helpful in communities where contact tracing resources are too limited to match the scope of the pandemic. We're building this app to provide components and tools that public health agencies can use to supplement their pre-existing efforts to fight COVID-19, assisted by voluntary public action.

Covid watch

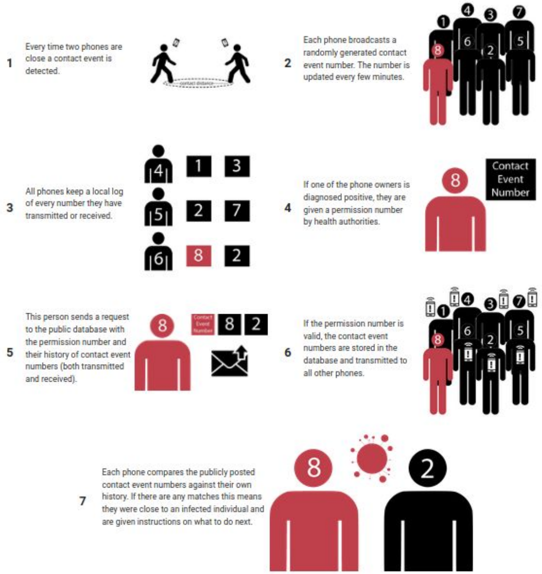
- Automated contact tracing at scale using anonymized Bluetooth proximity sensing
- Heatmap informed by epidemiological models using anonymized GPS data to warn users about high risk areas
- Recommendations from local health authorities and risk-aware suggestions about when to get

https://www.covid-watch.org/article

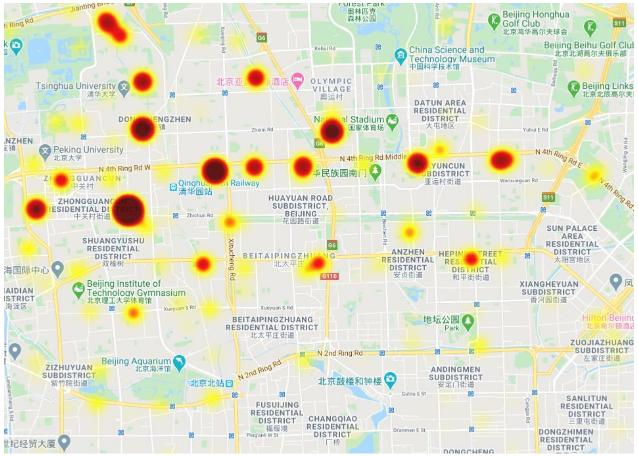
Bluetooth proximity



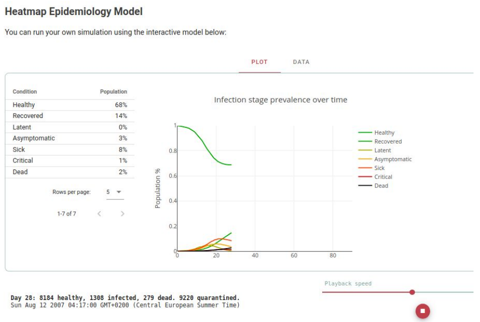
Workflow

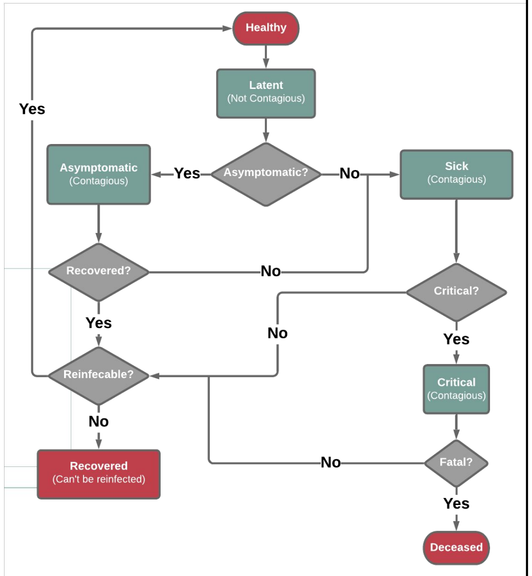


HEATMAP



The model






Life & Arts
+ Add to myFT

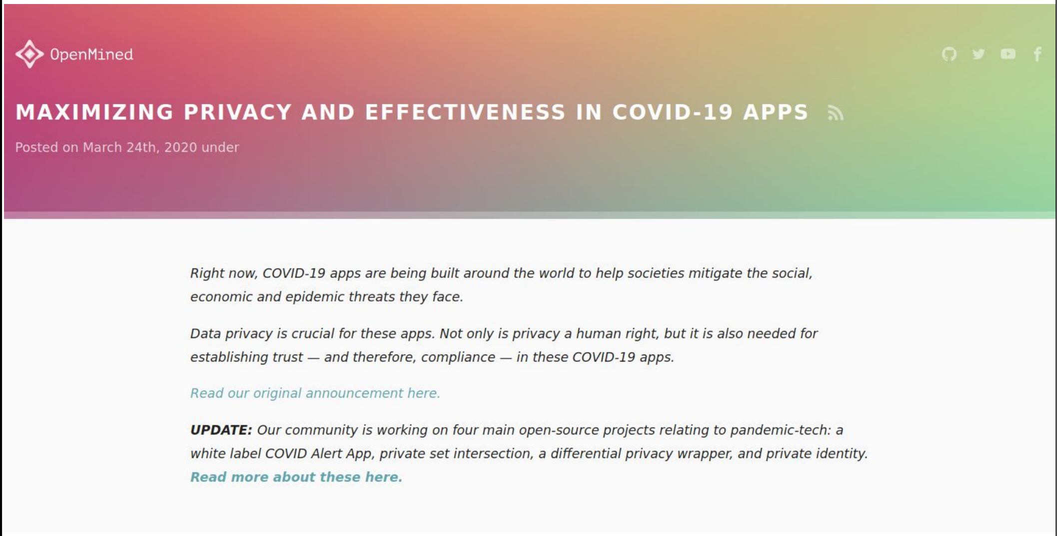
Yuval Noah Harari: the world after coronavirus | Free to read

This storm will pass. But the choices we make now could change our lives for years to come

https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75

takeaway

“Asking people to choose between privacy and health is, in fact, the very root of the problem. Because this is a false choice. We can and should enjoy both privacy and health. We can choose to protect our health and stop the coronavirus epidemic not by instituting totalitarian surveillance regimes, but rather by empowering citizens.”



Needs

- Historical and Current Absolute Location: where are you, where have you been, and have you (are you) quarantining?
- Historical and Current Relative Location: with whom have you been in close physical proximity in the last 2 weeks. This can sometimes be inferred from your absolute location, but the difference between 50 feet and 5 feet matters in terms of probability of

The group of people who have been tested for COVID-19
The group of people who have NOT been tested for COVID-19
Of those tested, the group of people who tested positive
Of those tested, the group of people who tested negative
The group of people who are self isolating
The group of people who are not self isolating
The group of people over 60
The group of people under 60
The group of people with underlying conditions

Private Set Intersection (PSI)

- Why?
  - To find out the people that met infected subjects
  - “… any analytic where you want to compare a user’s data (on a phone) with a patient’s data (in the cloud), use PSI to avoid centralizing a massive amount of data to a single location, which is a prime target for hacking and intentional or accidental mis-use”

PSI (toy example from wikipedia)

Set A

Name	Hash
Alice	64489c85dc2fe0787b85cd87214b3810
Eve	d3f791f59cbeff0ec06afb94bb23e772

Set B

Name	Hash
Eve	d3f791f59cbeff0ec06afb94bb23e772
Bob	2fc1c0beb992cd7096975cfebf9d5c3b

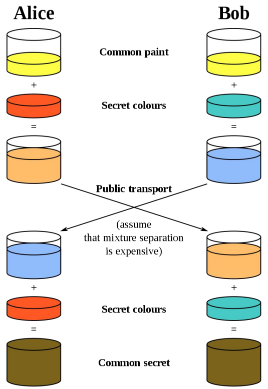
EXCHANGE

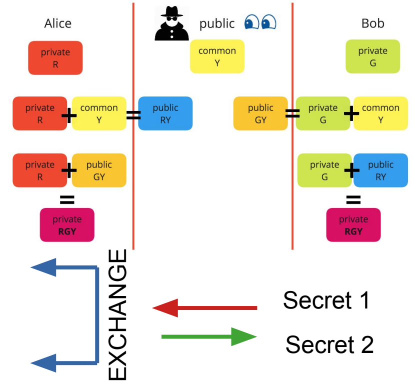
Eve has been in both places, but you don't know anything about Bob and Alice

Man-in-the-middle attack to obtain the encrypted data → one record matched (the intersection), although the name of shared record cannot be learned.

PSI (toy example from wikipedia)

Man-in-the-middle attack to obtain the encrypted data → one record matched (the intersection), although the name of shared record cannot be learned.





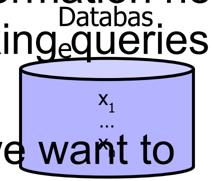
More details on WebSecurity and Privacy

Prof. Marchetti-Spaccamela

• Problem definition

A user makes query to a DB

- He is allowed to make statistical queries (privacy reasons)
- He is smart so he is able to infer information he should not know by repeteatedly making queries q
- to allow him to query but we want to preserve privacy of the DB



K-Anonimity

The information for each person contained in the release cannot be distinguished from at least k − 1 individuals whose

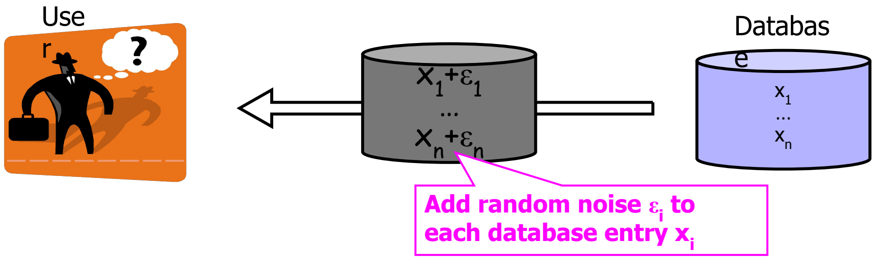
ID	Age	Zipcode	Diagnosis
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13068	Cancer

k-anonymization

ID	Age	Zipcode	Diagnosis
1	[20-30]	130**	Heart Disease
2	[20-30]	130**	Heart Disease
3	[20-30]	130**	Viral Infection
4	[20-30]	130**	Viral Infection
5	[40-60]	148**	Cancer
6	[40-60]	148**	Heart Disease
7	[40-60]	148**	Viral Infection
8	[40-60]	148**	Viral Infection
9	[30-40]	13***	Cancer
10	[30-40]	13***	Cancer
11	[30-40]	13***	Cancer
12	[30-40]	13***	Cancer

• Input Perturbation

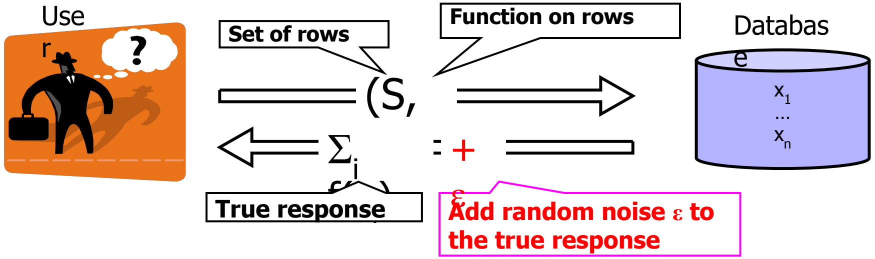
- Reveal entire database, but randomize entries



For example, if distribution of noise has
mean 0, user can compute average

• Output Perturbation

- Randomize response to each query



• Limits of Output Perturbation

- Let n be the size of the database (# of entries)
- If $O(n^{½})$ perturbation applied, adversary can extract entire database after poly(n) queries
- …but even with $O(n^{½} \log n)$ perturbation, it is unlikely that user can

Revealing Information while Preserving Privacy

Irit Dinur Kobbi Nissim*
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
{iritd,kobbi}@research.nj.nec.com

ABSTRACT

Keywords

1. INTRODUCTION

1.1 A Brief Background

A more practical definition of Privacy

IMPOSSIBLE

Privacy means that anything that can be learned about a respondent from the statistical database **can be learned without access to the database**

SECOND APPROACH (DIFFERENTIAL PRIVACY)

Whatever is learned about one respondent A

Example

- I know you have a salary 10 times higher of the average salary
- From the database I know the average salary
- I know your salary

Differential Privacy

- Promise: an individual will not be affected, adversely or otherwise, by allowing his/her data to be used in any study or analysis, no matter what other studies, datasets, or information sources, are available
- Paradox: learning nothing about an individual while learning useful statistical information about a population

Differential Privacy: ε parameter

For every pair of inputs D1 D2 that differ in one row ← presence or absence of a record

For every output O

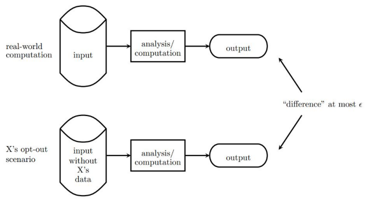
Controls the degree to which D1 and D2 can be distinguished.

Smaller ε gives more privacy and worse utility

If algorithm A satisfies differential privacy then

$$\frac{\Pr[A(D1) = O]}{\Pr[A(D2) = O]} < \exp(\varepsilon) \quad (\varepsilon > 0)$$

Intuition: adversary should not be able to use output O to distinguish between any D1 and D2



A randomized algorithm K gives *ε-differential privacy* if for all data sets D and D′ differing on at most one row, and any S ⊆ Range(K),

$$Pr[K(D) \in S] \leq exp(\varepsilon) \times Pr[K(D') \in S]$$

These are 2 quantities must be considered in DP algorithms are:

- Epsilon (ε): A metric of privacy loss at a differentially change in data (adding, removing 1 entry). The smaller the value is, the better privacy protection.
- Accuracy: The closeness of the output of DP algorithms to the pure output. In the *Private Machine Learning with PATE* part, I will use the classification accuracy on the test set as a statistic for evaluating accuracy.

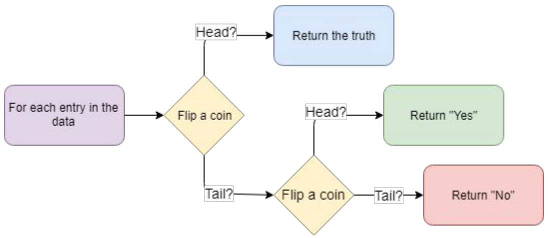
Calculate the average age of a group of people.
Instead of having each person send you their true age, you have them send you their true age + random number between -100 and 100. So, if someone was 42, they might send you 42 + (-50) = -8.
we can generate random numbers so that if you average over enough of them, they cancel each other out. Thus, if 10,000 people all add a random number (pulled from a distribution with a mean of 0) to their age before reporting it, the average age reported will still be similar to the underlying raw data despite the fact that nobody revealed their true age.

The bigger the random numbers (on average), the more privacy protection we give people, but the larger the group of people we need to average over before we can get aggregate statistics

This approach is useful for allowing app users to transform their local data in a way that protects it so that the central server can collect useful statistics without the central server being able to reverse engineer any specific person's personal data.

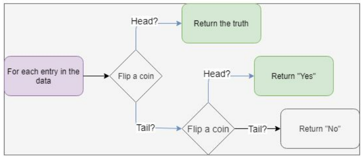
Differential Privacy

https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a

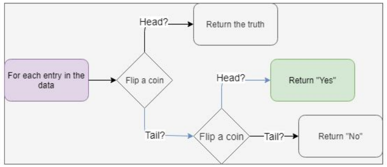


Assume you want to infer the percentage of innocents in the population (p_innocent) from that noisy data.

- Compute probability of returning "yes" given that individual isn't an innocent: P("yes"|not innocent) = *p_head+(1-p_head)*p_head.*



- Compute probability of returning "yes" given that individual is an innocent: P("yes"|innocent) = (1-*p_head*)**p_head.*



- Compute *p_innocent*:
p_innocent = (P("yes")-P("yes"|not innocent))/(P("yes"|innocent)-P("yes"|not innocent)) = 1-(P("yes")-(1-*p_head*)**p_head*)/*p_head.*
(proof by deduction)

Note: Above result is an asymptotic result guaranteed by Law of Large Numbers.

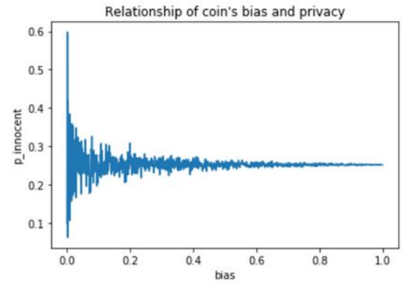
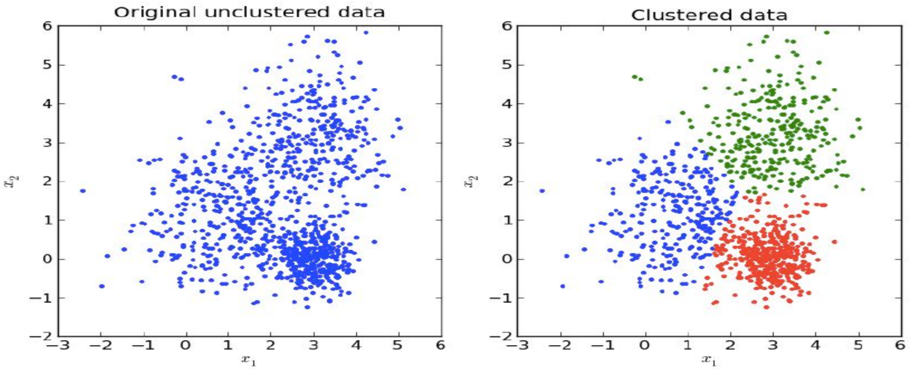


Figure 4: Compute on a simulated survey data of 20,000 individuals. Low bias coins add more noise to the data. A consequence of adding noise is the decrease in privacy loss.

- What does DP tell us?

  As you can see in Figure 4, the variance of *p_innocent* increases dramatically and approaches infinity when *p_head* approaches 0, lead to a rapid decrease in privacy loss. DP also gives us the same conclusion. Thus, when *p_head* is 0, the distribution of returned result is identical, no matter an individual is an innocent or not (the distance of 2 distributions is P("yes"|not innocent)-P("yes"|innocent)=*p_head*, the bias). If the number of innocents participates in the data changed, it does not lead to any changes in information in the noisy returned data. It means that there is no private information in the noisy returned data.

Case Study: K-means Clustering



K-means Clustering

- Partition a set of points x1, x2, …, xn into k clusters S1, S2, …, Sk such that the following is minimized:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|_2^2$$

Mean of the cluster S_i

K-means Clustering

Algorithm (Lloyd):

- Initialize a set of k centers
- Repeat
  - Assign each point to its nearest center
  - Recompute the set of centers

  Until convergence …
- Output final set of k centers

Differentially Private K-means

- Suppose we fix the number of iterations to T
- In each iteration (given a set of centers):
  - 1. Assign the points to the new center to form clusters
  - 2. Noisily compute the size of each cluster
  - 3. Compute noisy sums of points in each cluster

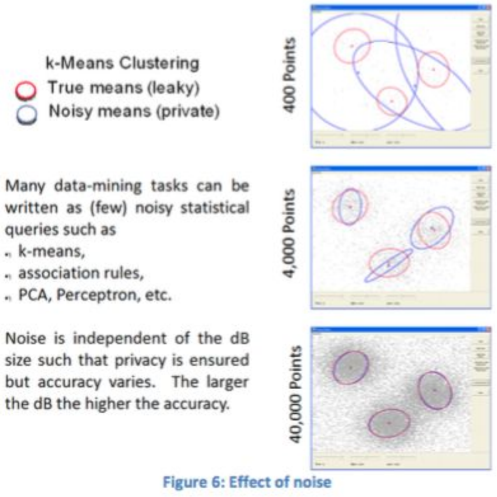


Figure 6: Effect of noise

https://dennyglee.com/2007/09/24/analyzing-data-while-protecting-privacy-a-case-study/

Differentially Private Kmeans

- Suppose we fix the number of iterations to T
- In each iteration (given a set of centers):
  - 1. Assign the points to the new center to form clusters
  - 2. Noisily compute the size of each cluster
  - 3. Compute noisy sums of points in each cluster

Each iteration uses ε/T privacy budget, total privacy loss is ε

Differentially Private Kmeans

- Question: Which of these steps expands privacy budget?

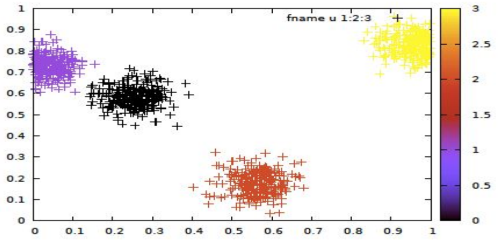
- In each iteration (given a set of centers):
  - 1. Assign the points to the new center to form clusters
  - 2. Noisily compute the size of each cluster
  - 3. Compute noisy sums of points in each cluster

Differentially Private Kmeans

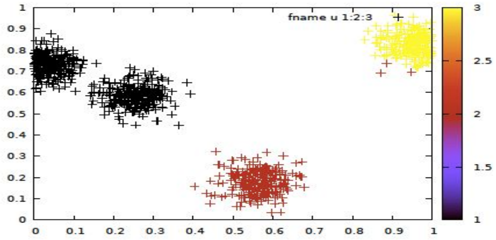
- In each iteration (given a set of centers):
  - 1. Assign the points to the new center to form clusters
  - 2. Noisily compute the size of each cluster
  - 3. Compute noisy sums of points in each cluster
- If each cluster has large size then we expect that the error in computing size is small

Differentially Private Kmeans

Original K-means alg.



Laplace Kmeans



- Even though we noisily compute centers, Laplace k-means can distinguish clusters that are far apart.
- Since we add noise to the sums with sensitivity proportional to |dom|.

Private Identity Server

- Prove an attribute about themselves in a privacy preserving way is a challenge.
- Namely, having someone simply enter “I am a doctor” in an app is really no guarantee at all that it’s true.

You must trust the PIS to be honest.

- A private identity server is a neutral third-party which will login to online services on behalf of an app user (via SSO) and verify certain claims. In this way, it is a “digital witness” of literally anything which you yourself could look up about yourself online.
- For example, let’s say that an app user wants to

Other resources

- https://www.technologyreview.com/s/615329/coronavirus-south-korea-smartphone-app-quarantine/
- http://theconversation.com/coronavirus-south-koreas-success-in-controlling-disease-is-due-to-its-acceptance-of-surveillance-134068
- https://www.ansa.it/sito/videogallery/italia/2020/03/19/coronavirus-controlli-celle-telefoniche-cosa-sono-e-come-funzionano_3610d248-e14d-495f-aefe-9d03333362b5.html
- http://dangrover.com/blog/2020/04/05/covid-in-ui.html?fbclid=IwAR318H6XUAdxOSJjjwBdI6nDpOqAx_EcJ_AmSkNAtssFuC3pt5jG0y5NqNg