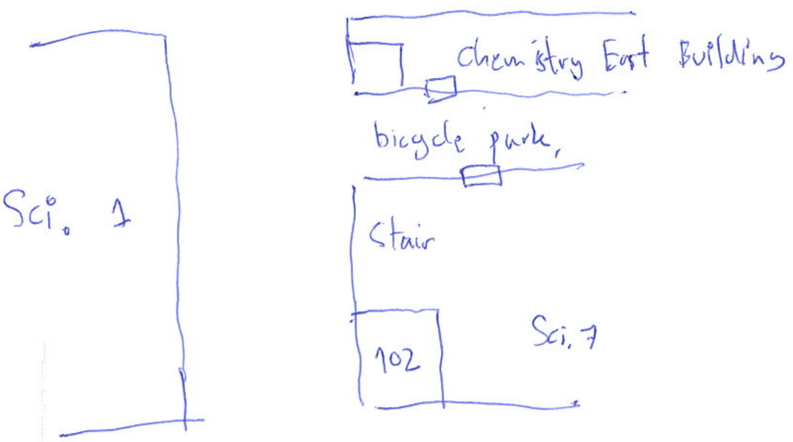
4810-1184 Algorithms for Information Security and Privacy

Instructor: Vorapong Suppakitpaisarn, vorapong@is.s.u-tokyo.ac.jp
(International Center for IST)

I sometimes discuss about our international activities at classes.

Office Hour: Thursday 15:00 - 16:30

How to get at my office? Room 137, Chemistry East Building



First room after you get into the building.
The room has no room nuber at the gate!!!

Class Schedule

9/25 Course Introduction, PAC Learning
10/~~1~~2 Differential Privacy: Laplacian Mechanism
10/~~8~~9 Differential Privacy: Exponential Mechanism
10/~~15~~ 16 Differential Privacy: Composition, Small DB algorithm
10/~~22~~ 23 Differential Privacy: Private PAC Learning
10/~~29~~ 30 Countermeasures for Linking Attack
11/26 Midterm Examination
[30% to grade]

11/13	No class ~~ESI will~~
	[Day for cancelled classes]
11/20	Optional Class: Introduction to Abstract Algebra.
11/27	Calculation on Elliptic Curve Cryptosystem
12/3	Discrete Logarithm Problems
12/10	Elliptic Curve Cryptography Protocol
12/17	Identity-based Cryptosystem
12/25, 1/1	No class. [Holiday]
1/8	Final Examination [70% of credits]

Please inform me if you are not available on 11/6 or 1/8 before 10/9

Relationship with other courses

focus on security

Introduction to cryptography
(Prof. Phong Nguyen)
2018 S

Contemporary cryptography
(Prof. Tsuyoshi Takagi)
2017 A

This course
2018 A

Focus on Speed.

Approximation Algorithms with Applications
2018 S
(Me)

→

Network Optimizations
2018 A

Differential Privacy

Name	Weight
Alice	40
Bob	60
Charles	80
Doe	60

Private Information

$\Downarrow$

Average Weight $= 60$ public information

- Charles does not want to publish his weight, but Alice, Bob, and Doe do publish.
- Average Weight $= \dfrac{w_{Alice} + w_{Bob} + w_{Charles} + w_{doe}}{4}$

$$60 = \frac{40 + 60 + w_{Charles} + 60}{4}$$

$$w_{Charles} = 80$$

information leakage!!!

Idea : Add noise to public information

Average Weight $= 60$ $\Longrightarrow$ Average Weight + noise $= 55$

public information $\qquad$ public information

By the noise, it is impossible to predict Charles' weight.

Linking Attack

Data published by goverment.

Name	Age	Occupation
Alice	25	Student
Bob	30	Student
Charles	30	Banker.

Data that hospitals give to machine learners.

[They want to find the diabete potential of each person.]

Name	Age	Occupation	Diabetes
Alice	25	Student	✓
Bob	30	Student	✗
Charles	30	Banker	✗

Don't publish

- We do not want people to know that Alice has Diabete.
- We know from public information that the only 25-year-old is Alice.
- We know from hospital information that the only 25-year-old has diabete.

$\Downarrow$

Alice has diabete. 

87% of US. citizen can be uniquely identified by sex, birthdate, and city [Sweeney 2002]

Elliptic Curve Cryptography (ECC)

- Alternative choice to RSA
- Used in industrail web services (Google, Facebook, Microsoft)
- Have features that RSA does not have. [forward secrecy]
- Have better security level.

NIST Standard 2010 (National Institute of Standards and Technology)

RSA 1024 bits ——> ECC 160 bits.

security parameter

[larger -> more secure]

[larger -> more memory consumption]

[Bos et al.] ——— at least until 2014

much after that [unlikely before 2028]

ECC is used in computation environment with limited memory.

Public Key Cryptography

Certificate Authority. (CA)

1. May I have Bob's key

Alice.

Bob.

2. put the secret to the box

——> 3. send the box with secret inside

4. Use key that only he has to open the box

Alice has to contact ~~Bob~~ CA at everytime she send a message to a new person.

Ineffient...

Identity-Based Cryptosystem

Alice

Know Bob's e-mail address

1. Create Bob's box using Bob's e-mail address.
2. Put the secret into the box

3. send the box with secret inside

Bob

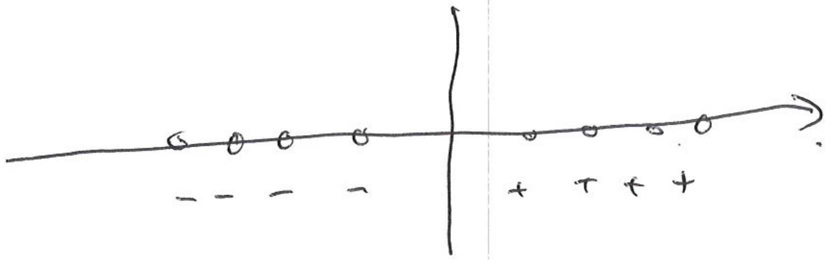
Has a key generated from his e-mail address

4. Use key to open the box

Only task of Certificate Authority is to Generate Bob's key!!!

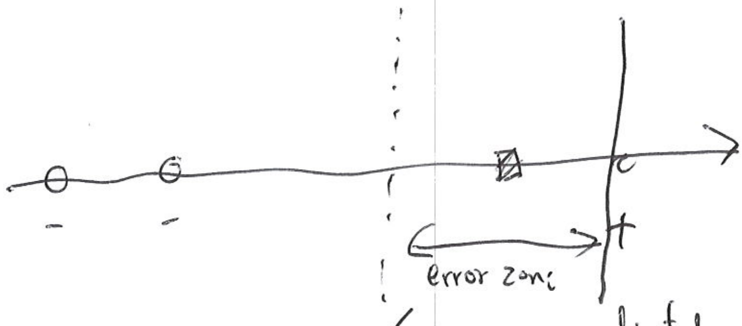
Probably Approximately Correct Learning (PAC Learning)

- Our graduate school is selecting students based on their entrance exam score. (assumption)
- We do not publish or give entrance examination score to students. You have to pay 300 yen to know your score.
- We want to know the borderline.



When we have a lot of information.

Real borderline.



There have not been any of come from this zone.

predicted borderline → smallest score we know that pass to the school.

<u>Question</u>: What is the probability that we have error zone > 10 points. when we conduct a sample <u>independently</u> for m times.

<u>Assumption</u>: The full score is 1000.

Probability that a sample is i is $1/1000$ for all i.

Probability that we have a sample between (real border) and (real border + 10) is 10/1000.

Prob. that we don't is $\frac{99}{100}$

Prob. that we don't for m times is $\left(\frac{99}{100}\right)^m$.

We will have a prediction with $\underbrace{\text{at most 10 points error}}_{\text{Approximatly}}$ is $\underbrace{1-\left(\frac{99}{100}\right)^m}_{\text{Probably}}$.

at most k points error is $1-\left(1-\frac{k}{1000}\right)^m$

Important inequality: $1+x \leq e^x$ for all $x \in \mathbb{R}$

$1-x \leq e^{-x}$

$1-e^{-x} \leq -x$

$1-e^{-x} \approx -x$ when x is small

$$\left(1-\frac{k}{1000}\right)^m \leq e^{-\frac{km}{1000}}$$

$$1-\left(1-\frac{k}{1000}\right)^m \geq 1-e^{-\frac{km}{1000}} \approx \frac{k}{1000} \cdot m.$$

$\frac{k}{1000}$: maximum error in precision.

m: # times we do sampling

~~We will~~

IF we allow machine learning algorithms to have.

- larger error (approximately)
- ~~have~~ more time ε

we have more chances to attain it.

(probably).

Question:

In our discussion, we assume that $\Pr[\text{sample} = i] = \frac{1}{1000}$ for all i.

Show that the whole discussion still works for all probability distribution over i.