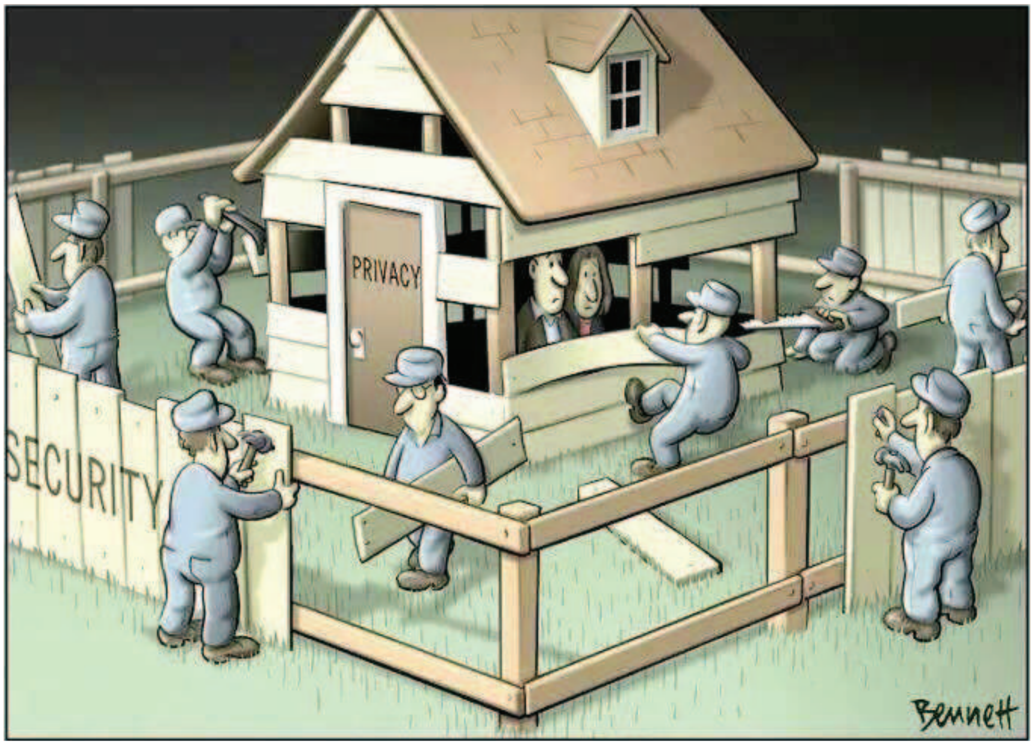
FACULTY OF COMPUTER SCIENCE
CS4413/6413: Foundations of Privacy
Professor: Dr. Rongxing Lu
Office: GE 114
Email: rlu1@unb.ca
Phone: 451-6966
Winter 2019 Tutorial Time: F 3:30-4:20

Tutorial 0

1) What is the personal data?
   - Answer: Any information that can be used to identify a living person - directly and indirectly, or that relates to them.
     - This could be: name, an identification number, or location data, like an IP address.
     - It could also include other information that leads to an individual being identified (which could be: physical, genetic or cultural).
     - More care needs to be taken with sensitive personal data, e.g., health data, religious beliefs.
2) Why data privacy matters to us?
   - Answer:
     - We care - we are responsible for handling people's most personal information.
     - This is an opportunity to make privacy central to what we do.
     - By not handling personal data properly we could put individuals at risk and the entitys reputation at stake.
     - Getting it wrong could result in significant fines.
     - We need robust systems and processes in place to make sure we use personal information properly and comply.
3) Briefly discuss the General Data Protection Regulation (GDPR), which has come in to force on 25 May 2018.
   - Answer: GDPR is a European law that will replace the current Data Protection Act and the UK government will still implement the rules after Brexit. The aim is to strengthen and unify personal data protection for all individuals living in the European Union. In addition, the Information Commissioners Office (ICO) will lead on GDPR in the UK and will hand out penalties for organisations who are in breach of the new law.



Tutorial 1

1) There are four key security issues that we need to consider in database, including availability, authenticity, integrity, and confidentiality. Please give a brief definition for each of them.

   ● Answer:

   - Availability: Data should be available at all necessary times and only the appropriate users. In addition, data should be able to track who has access to and he/she has accessed what data.
   - Authenticity: Data should be edited by an authorized source and the database system should be accessed by the users who they say they are. Authenticity should verify that all report requests are for authorized users and any outbound data are going to the expected receiver.
   - Integrity: Integrity is to verify that any external data has the correct formatting and other metadata, and all input data is accurate and verifiable. At the same time, data is following the correct work flow rules for the institution/company. All data changes and who authored them to ensure compliance with corporate rules and privacy laws should be able to report.
   - Confidentiality: Confidential data is only available to correct people, and entire database is secure from external and internal system breaches. Confidentiality should provide for reporting on who has accessed what data and what they have done with it. Mission critical and legal sensitive data must be highly secure at the potential risk of lost business and litigation.

2) One of privacy threats to database privacy is knowledge discovery data mining (KDDM). Please give a brief description of KDDM.

   ● Answer: KDMM is to extract information from database, and suggest a pattern regarding the data stored in the database. It is used to discover patterns that classify individuals into categories, revealing in the way confidential personal information with certain probability.

3) There are three main types of authentication techniques, including what you know, what you have, and what you are. Please give a brief description for each of them.

   ● Answer:

   - What you know? We know the most well known form of authentication is user/password verification, which is used to gain access. There are two problems and security issues considerations: (i) potential for password to be written down on paper; (ii) password can be guessed or broken easily.
   - What you have? Physical object is needed in order to gain access, such as key card, ID card, token etc. At the same time, the most common uses includes gaining access to buildings or restricted areas. There are two problems or security issues: loss of object and replication.
   - What you are? We have biometric information and the unique biological characteristics can be used for identification. There are two problems or security issues: (i) the biological information may change with aging, injury and some other biological or environmental conditions; (ii) the identification system is not fully accurate, i.e., there are false-acceptance rate and false-rejection rate.

4) What are the models of anonymity?

   ● Answer:

   - Interactive model: This model is similar to statistical database. Data owner acts as "gatekeeper" to data. Users submit queries in some agreed language and gatekeeper gives an (anonymized) answer, or refuses to answer.
   - "Send me your code" model: In this model, data owner executes code on their system and reports result, but it cannot be sure the code is not malicious.
   - Offline model or "publish and be damned" model: Data owner somehow anonymizes data set and publishes the results to the world and retires.

5) What are the goals of anonymity?

   ● Answer: The goals are as follows.

   - Prevent (high confidence) inference of associations. In other words, it aims to prevent linking sensitive information to an individual.
   - Prevent inference of presence of an individual in the data set.
   - Have to model what knowledge might be known to attacker, including the background knowledge and domain knowledge.

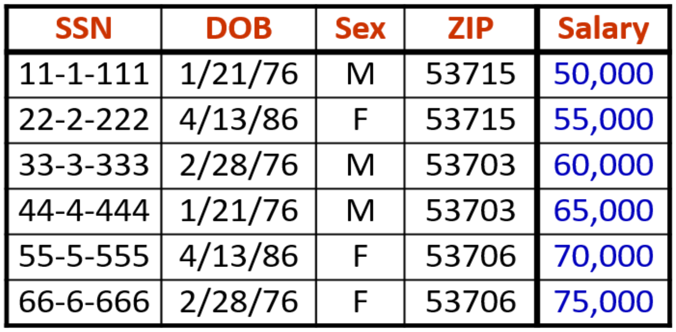
6) What are identifiers, quasi-identifiers, and sensitive attributes in database?

   ● Answer:

   - Identifiers: uniquely identity, e.g., Social Security Number
   - Quasi-Identifiers: combine several attributes together to be a new identity, which is enough to partially identify an individual in a database. For example, DOB+Sex+ZIP are unique for 87% of US residents.

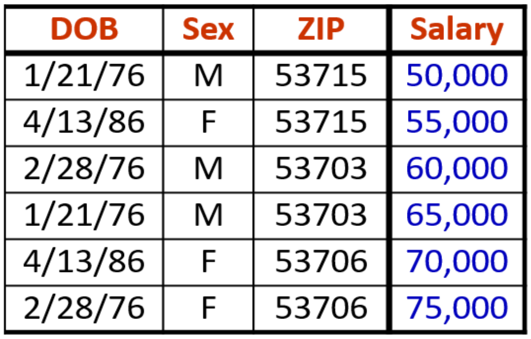
- Sensitive attributes: the associations we want to hide, e.g., the salary in the “census”.

7) Use an example to explain the De-Identification, and Linking Attack for Tabular Data.
   - Answer:
     - De-Identification:
       We take the census data as an example to explain the de-identification. The following table contains some census data recording incomes and demographics, where SSN is an identifier and salary is a sensitive attribute. At the same time, salary association violates individual’s privacy.



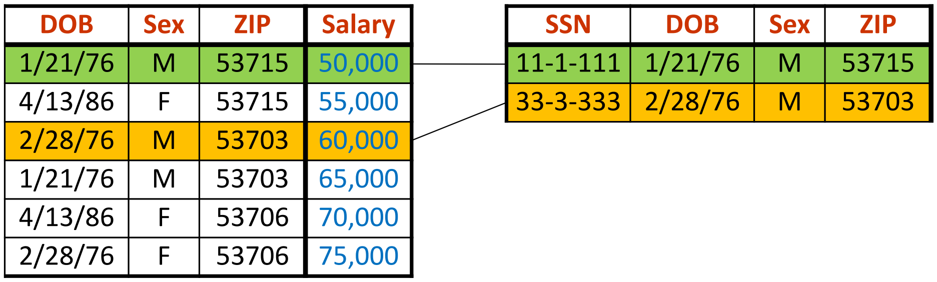
SSN	DOB	Sex	ZIP	Salary
11-1-111	1/21/76	M	53715	50,000
22-2-222	4/13/86	F	53715	55,000
33-3-333	2/28/76	M	53703	60,000
44-4-444	1/21/76	M	53703	65,000
55-5-555	4/13/86	F	53706	70,000
66-6-666	2/28/76	F	53706	75,000

       We can remove SSN to create a de-identified table.



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

     - Linking Attack: linking de-identified private data and public available data is likely to achieve identification and access some sensitive data. For example, in the following table, the de-identified census data table and public available data table can be combined to identify users and their salary information.



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

SSN	DOB	Sex	ZIP
11-1-111	1/21/76	M	53715
33-3-333	2/28/76	M	53703

8) What is k-Anonymization in in database privacy?
   - Answer: k-anonymity: Table T satisfies k-anonymity with regards to quasi-identifier QI if each tuple in (the multiset) T[QI] appears at least k times.

9) What is homogeneity attack in k-anonymity?
   - Answer: k-anonymity requires each tuple in (the multiset) T[QI] to appear at least k times. In this case, if (almost) all sensitive attribute values in a QI group are equal, privacy is lost. This is homogeneity attack.

10) What are l-Diversity and t-Closeness?
    - Answer:
      - l-diversity: Table T satisfies l-diversity with regards to quasi-identifier QI if each of its QI groups contains at least l well-represented values for the sensitive attributes.
      - t-closeness: Table T satisfies t-closeness with regards to quasi-identifier QI if the distance between the distribution of sensitive attribute values in the group and in the whole table is no more than threshold t.

11) What is Differential Privacy (DP) technique?

• Answer: The DP can guarantee that the privacy risk should not substantially increase as a result of participating in a statistical database. For example, there are two datasets X and X', and X is a neighbor of X' because they differ in one row. However, from the released statistics, it is hard to distinguish X and X'.

12) Discuss why the sensitivity of the function F =HISTOGRAM is 2, i.e., $S(F) = 2$.

• Answer: The sensitivity of a function F is the maximum (absolute) change over all possible adjacent inputs

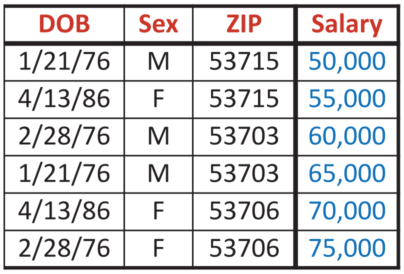
$$S(F) = Max_{D_1,D_2:|D_1-D_2|=1}|F(D_1) - F(D_2)|.$$

For the HISTOGRAM, removing any record affects one bin and changing any record affects two bins in the worst case. Therefore, $S(F) = 2$.

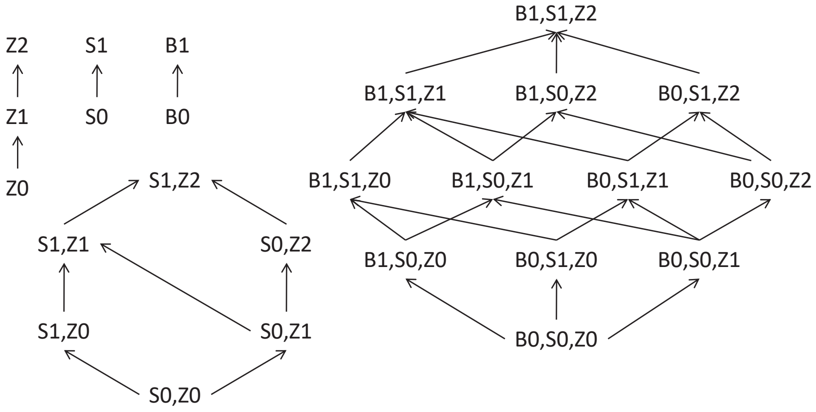
13) Incognito is one of approaches to implement the k-anonymity. Given a table below, the full-domain generalizations described by "domain vectors" are represented as follows.

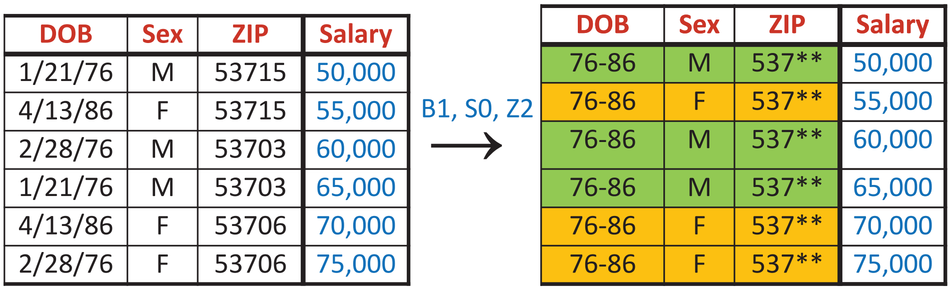
- $B_0 = \{1/21/76, 2/28/76, 4/13/86\} \rightarrow B_1 = \{76 - 86\}$
- $S_0 = \{M, F\} \rightarrow S_1 = \{*\}$
- $Z_0 = \{53715, 53710, 3706, 53703\} \rightarrow Z_1 = \{5371*, 5370*\} \rightarrow Z_2 = \{537**\}$

Based on the Lattice of Domain Vectors, we can achieve the k-anonymity for the original table. For example, with (B_1, S_0, Z_2) generalization, we can achieve $k = 3$-anonymity.



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000





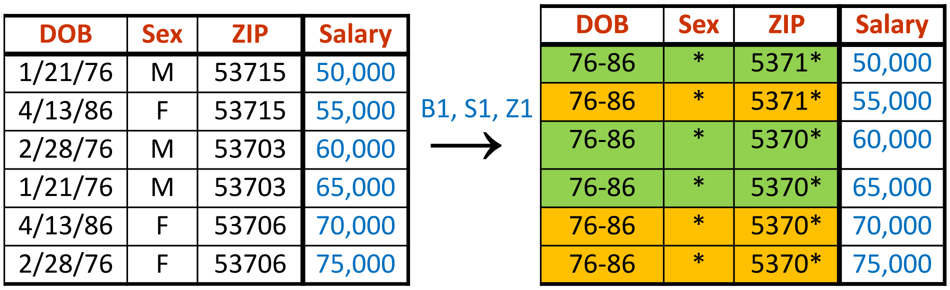
DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

B1, S0, Z2 →

DOB	Sex	ZIP	Salary
76-86	M	537**	50,000
76-86	F	537**	55,000
76-86	M	537**	60,000
76-86	M	537**	65,000
76-86	F	537**	70,000
76-86	F	537**	75,000

Please follow the example to apply (B_1, S_1, Z_1), (B_1, S_1, Z_2) to generalize the original table, and discuss what is the value of k in each generalized case?

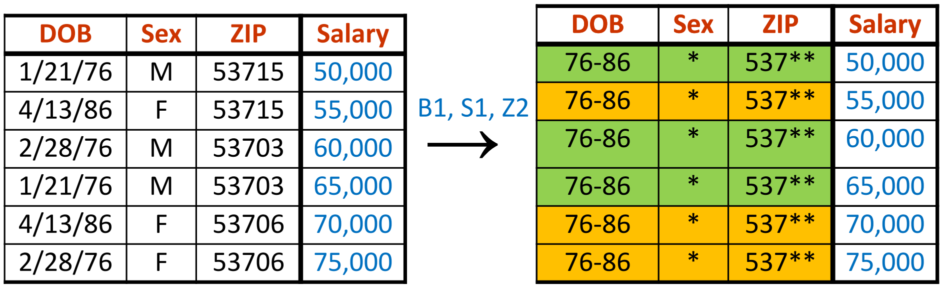
• Answer: The generalized tales are as follows.



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

B1, S1, Z1 →

DOB	Sex	ZIP	Salary
76-86	*	5371*	50,000
76-86	*	5371*	55,000
76-86	*	5370*	60,000
76-86	*	5370*	65,000
76-86	*	5370*	70,000
76-86	*	5370*	75,000



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

DOB	Sex	ZIP	Salary
76-86	*	537**	50,000
76-86	*	537**	55,000
76-86	*	537**	60,000
76-86	*	537**	65,000
76-86	*	537**	70,000
76-86	*	537**	75,000

When applying (B_1, S_1, Z_1) to generalize the original table, $k = 2$. Similarly, when applying (B_1, S_1, Z_2) to generalize the original table, $k = 6$.

14) Let $F(x)$ be the true answer on input x, and $Lap(\lambda)$ be the noise sampled from Laplace distribution with parameter $\lambda = S(F)/\epsilon$. Please prove that the release of $F(x) + Lap(\lambda) = F(x) + Lap(S(F)/\epsilon)$ can obtain ϵ-DP guarantee.

Proof: Suppose that $A = F(x) + Lap(\lambda)$, and D_1 and D_2 are any two adjacent DBs. Thus, $A(D_1) = F(D_1) + x_1$ and $A(D_2) = F(D_2) + x_2$, where x_1 and x_2 are $Lap(\lambda)$ distributed. Since $\lambda = S(F)/\epsilon$, the probability density for x_1 is proportional to $e^{-||x_1||_1(\frac{\epsilon}{S(F)})}$ and . Similarly, the probability density for x_1 is proportional to $e^{-||x_2||_1(\frac{\epsilon}{S(F)})}$. Therefore, for any $T \in range(A)$

$$\frac{Pr[A(D_1) = T]}{Pr[A(D_2) = T]} = \frac{Pr[F(D_1) + x_1 = T]}{Pr[F(D_2) + x_2 = T]} = \frac{Pr[x_1 = T - F(D_1)]}{Pr[x_2 = T - F(D_2)]} = \frac{e^{-||T-F(D_1)||_1(\frac{\epsilon}{S(F)})}}{e^{-||T-F(D_2)||_1(\frac{\epsilon}{S(F)})}}$$

$$= e^{(-(||T-F(D_1)||_1 - ||T-F(D_2)||_1)(\frac{\epsilon}{S(F)})} \leq e^{-||F(D_2)-F(D_1)||_1(\frac{\epsilon}{S(F)})}$$

where the inequality follows form the triangle inequality. By the definition of sensitivity,

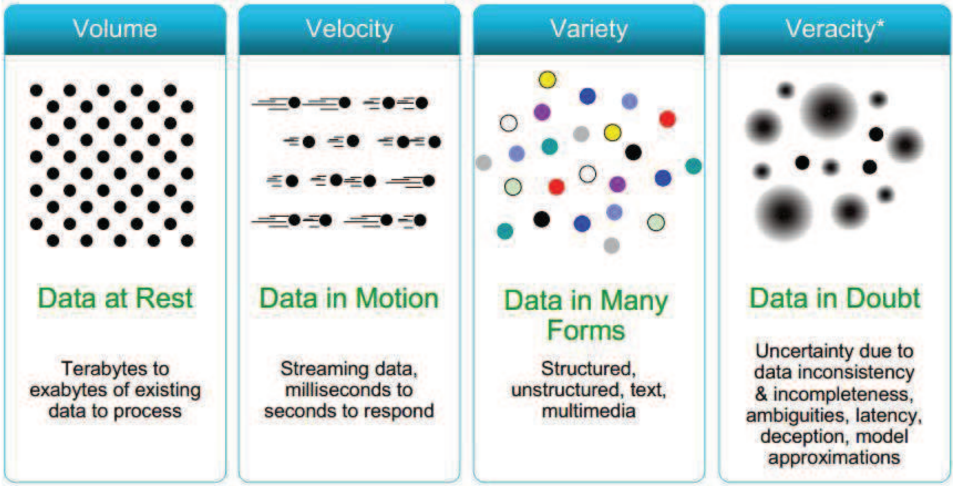
$$S(F) = Max_{D_1, D_2 : |D_1 - D_2| = 1} |F(D_1) - F(D_2)|$$

.

Thus, $e^{-||F(D_2)-F(D_1)||_1(\frac{\epsilon}{S(F)})} \leq e^{-\epsilon}$. The ration is bounded by $e^{-\epsilon}$, yields ϵ-Differential Privacy.

Tutorial 2

1) What are big data, its challenges and opportunities?
   - Answer:
     - Big data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. (Gartner 2012) In addition, Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. (Wikipedia)
     - Challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
     - Opportunities: The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data. Compared to separate smaller sets with the same total amount of data, allowing correlations can find "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."
2) Briefly explain 4V characteristics in big data.
   - Answer:
     - Volume: Data volume is increasing exponentially, e.g., 44x increase from 2009 to 2020 and from 0.8 zettabytes to 35zb.
     - Velocity: Data are generated fast and need to be processed fast.
     - Variety: Different types of data are involved, including relational data, text data, graph data, etc., which become more complex. All these types of data need to be linked together.
     - Veracity: The data inconsistency and incompleteness, ambiguities etc. will bring some uncertainty in big data. Veracity will consider some security issues in big data in order to protect data veracity.

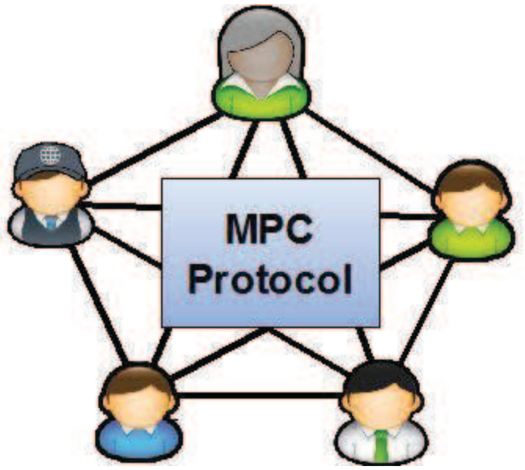

Volume
Velocity
Variety
Veracity*
Data at Rest
Terabytes to exabytes of existing data to process
Data in Motion
Streaming data, milliseconds to seconds to respond
Data in Many Forms
Structured, unstructured, text, multimedia
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

3) Briefly describe some typical security issues often discussed in big data.
   - Answer:
     - Secure data collection: It aims to make the collection of data private as well as authenticated.
     - Secure data filtration: In order to reduce useless data, set the data filtering criteria for being useful and keep the filtering criteria secret even if it is executed at the source.
     - Data integrity and poisoning concerns: Only conduct data computing on authenticated data.
     - Proof of data storage: Check the file or data is actually stored in the cloud.
     - Secure outsourcing of computation: A weak client outsources a computation to the provider. After the provider returns the result and a "proof" that the computation was carried out correctly, the client can use less computation cost to verify the proof is true or not.
4) What are data aggregation, suppression, swapping, synthesis techniques?
   - Answer:
     - Aggregation: With aggregation, privacy is protected by aggregating individual records within a report-based and summarized format before release. Aggregation reduces disclosure risks by turning records at risk into less-risky records.
     - Suppression: In suppression, not all the data values are released. Some values are removed, withheld, or disclosed. Typically, data agencies remove sensitive values from the released dataset. They may select to suppress entire variables

or just at-risk data values. However, suppression may lead to inaccurate data mining and analysis as important data values are suppressed and missing.

- Swapping: In swapping, data values of selected records are swapped to hide the true owner of the records, thereby making the matching inaccurate. Agencies may choose to select to have high rate of data swapping in which a large percentage of records are selected for swapping, or low rate in which only a small percentage of records are selected for swapping.
- Synthesis techniques: In this way, the values of sensitive variables are replaced with synthetic values generated by simulation. In addition, the synthetic values are basically random values generated by a probability distribution function simulator.

5) What is multi-party secure computation in big data?



• Answer:

Generally, MPSC allows the players to perform an arbitrary on-going computation during which new inputs can be provided and security in MPSC means that the players' inputs remain secret (except for what is revealed by the intended results of the computation) and that the results of the computation are guaranteed to be correct.

Tutorial 3

1) What are ciphertext-only attack and known-plaintext attack on symmetric encryption?
   - • Answer:
     - Ciphertext-only attack: An adversary only has several ciphertexts and he/she tried to recover the decryption key or plaintext of a given ciphertext.
     - Known-plaintext attack: An adversary has several pairs of (plaintext, ciphertext) and he/she tried to recover the decryption key or plaintext of a given ciphertext.
2) What is the Kerckhoffs principles?
   • Answer: The security of algorithms should depend on the confidentiality of the key. All other things except the key can be public.
3) What is the block cipher? What is the goal of padding in block cipher?
   • Answer:
     - Block ciphers break messages into fixed length blocks, and encrypt each block using the same key.
     - The goal of padding in block cipher: when the plaintext message is broken into blocks, the last block may be short of a whole block and needs padding.
4) What are advantages and disadvantages of symmetric encryption in terms of efficiency and key management?
   • Answer:
     - Efficiency: Symmetric encryptions are generally very fast.
     - Key management is a problem. When there are N users, the total number of required keys are $\frac{N(N-1)}{2}$.
5) What is the birthday attack?
   • Answer: Birthday attack is to find two people with the same birthday, which is the same thing as finding a collision for a particular hash function. For example, in a group of 23 randomly chosen people, at least two will share a birthday with probability at least 50%. If there are 30, the probability is around 70%.
6) Let $P(N,q)$ denote the probability of at least one collision when we throw $q \geq 1$ balls at random into $N \geq q$ buckets. Then, we have

$$P(N,q) \leq \frac{q(q-1)}{2N}, \quad P(N,q) \geq 1 - e^{\frac{-q(q-1)}{2N}}$$

when $1 \leq q \leq \sqrt{2N}$,

$$P(N,q) \geq 0.3 \cdot \frac{q(q-1)}{N}$$

Please prove these birthday bounds.
Proof: Let C_i be the event that the i-th ball collides with one of the previous ones. Then, $Pr[C_i]$ is at most $\frac{i-1}{N}$, since when the i-th ball is thrown in, there are at most $i-1$ different occupied slots and the i-th ball is equally likely to land in any of them. Thus,

$$P(N,q) = Pr[C_1 \vee C_2 \vee \cdots \vee C_q] \leq Pr[C_1] \vee Pr[C_2] \vee \cdots \vee Pr[C_q] \leq \frac{0}{N} + \frac{1}{N} + \cdots + \frac{q-1}{N} = \frac{q(q-1)}{2N}$$

Thus, this proves the upper bound.
Then, let D_i be the event that there is no collision after having thrown in the i-th ball. If there is no collision after throwing in i balls then they must all be occupying different slots, so the probability of no collision upon throwing in the $(i+1)$-th ball is exactly $\frac{N-i}{N}$. That is, $Pr[D_{i+1}|D_i] = \frac{N-i}{N} = 1 - \frac{i}{N}$.
Noth that $Pr[D_1] = 1$. The probability of no collision at the end of the game can now be computed via

$$1 - P(N,q) = Pr[D_q] = Pr[D_q|D_{q-1}] \cdot Pr[D_{q-1}] = \cdots = \prod_{i=1}^{q-1} Pr[D_{i+1}|D_i] = \prod_{i=1}^{q-1}(1 - \frac{i}{N})$$

Note that $\frac{i}{N} < 1$, we can use the inequality $1 - x \leq e^{-x}$ for each term of the above expression. This means the above is not more than

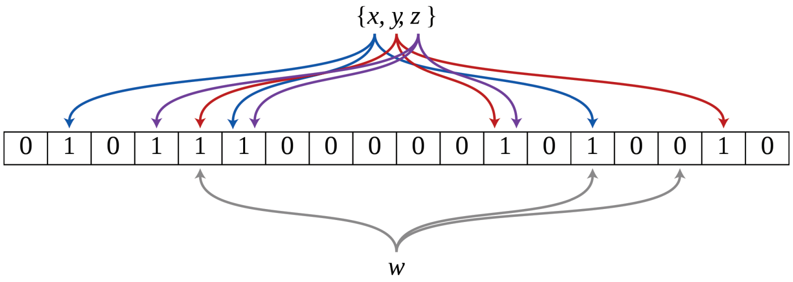
$$\prod_{i=1}^{q-1}(e^{-\frac{i}{N}}) = e^{-\frac{1}{N}-\frac{2}{N}-\cdots-\frac{q-1}{N}} = e^{-\frac{q(q-1)}{2N}}$$

Therefore, $P(N,q) \geq 1 - e^{-\frac{q(q-1)}{2N}}$.
Thus, this proves the lower bound.

When $1 \leq q \leq \sqrt{2N}$, we know $\frac{q(q-1)}{2N} \leq 1$. Then, we can use the inequality $(1-\frac{1}{e}) \cdot x \leq 1 - e^x$ to get $P(N,q) \geq (1-\frac{1}{e})\frac{q(q-1)}{2N}$.

Thus, $P(N,q) \geq 0.3 \cdot \frac{q(q-1)}{N}$.

7) Alice wants to send $m = 10$ files to Bob, how to use the optimal Bloom Filter technique to assure the data integrity of these files?



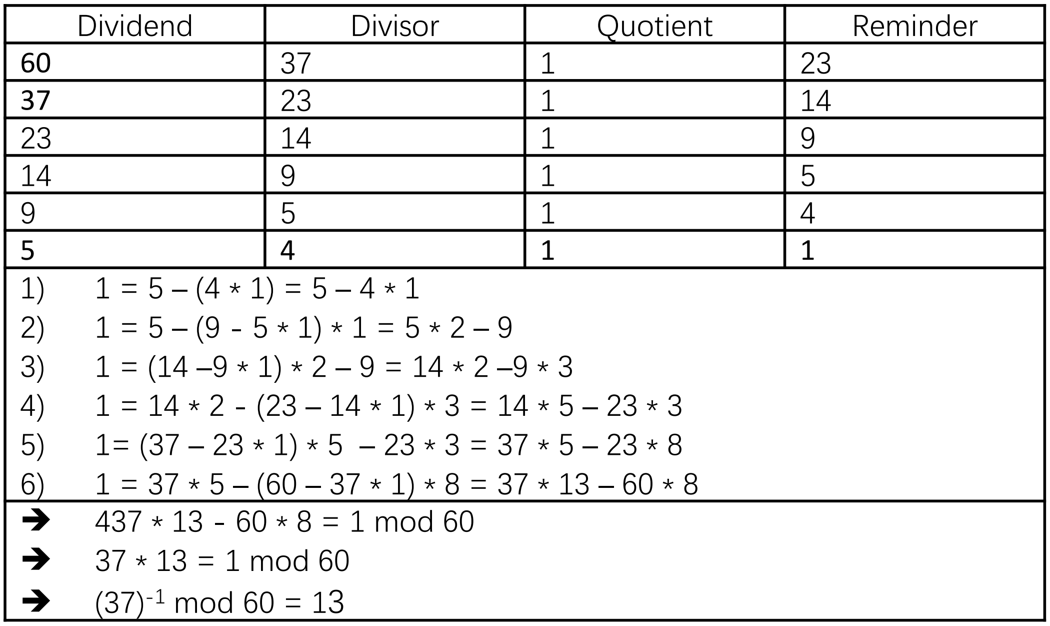
• Answer: Alice has $m = 10$ files and the number of hash functions is $k = 3$. In order to optimize the Bloom Filter technique and reduce the probability of false positive, it should satisfy that

$$k = \ln 2 \cdot \frac{n}{m}.$$

Then $n = k * m / \ln 2 = 3 * 10 / \ln 2 = 44$, so Alice can use a Bloom Filter array with size n to store these data.

8) Use the extended Euclidean algorithm to compute the value of $x = 37^{-1} \bmod 60$.

• Answer:



Dividend	Divisor	Quotient	Reminder
60	37	1	23
37	23	1	14
23	14	1	9
14	9	1	5
9	5	1	4
5	4	1	1

1) 1 = 5 – (4 * 1) = 5 – 4 * 1
2) 1 = 5 – (9 - 5 * 1) * 1 = 5 * 2 – 9
3) 1 = (14 –9 * 1) * 2 – 9 = 14 * 2 –9 * 3
4) 1 = 14 * 2 - (23 – 14 * 1) * 3 = 14 * 5 – 23 * 3
5) 1= (37 – 23 * 1) * 5 – 23 * 3 = 37 * 5 – 23 * 8
6) 1 = 37 * 5 – (60 – 37 * 1) * 8 = 37 * 13 – 60 * 8

➔ 437 * 13 - 60 * 8 = 1 mod 60
➔ 37 * 13 = 1 mod 60
➔ $(37)^{-1}$ mod 60 = 13

9) Prove the following results.

a) If $p|10a - b$, $p|10c - d$, then $p|ad - bc$.

Proof: From $p|10a - b$, we know $p \cdot k_1 = 10a - b$ for some k_1. Then,

$$p \cdot k_1 \cdot c = (10a - b) \cdot c$$

Let $k_2 = k_1 \cdot c$, we have $p \cdot k_2 = (10a - b) \cdot c = 10ac - bc$.

Similarly, we have $p \cdot k_4 = (10c - d) \cdot a = 10ca - ad$ for some k_4. Then,

$$p \cdot k_2 - p \cdot k_4 = 10ac - bc - 10ac + ad \Longrightarrow p(k_2 - k_4) = ad - bc$$

Therefore,

$$p|ad - bc$$

b) if n is odd, then $3|2^n+1$.
Proof: Since $2+1 \equiv 0 \mod 3$, we have $2 \equiv -1 \mod 3$. Then,

$$2^n \equiv (-1)^n \mod 3$$

Because, n is odd, we have

$$2^n - (-1)^n \equiv 0 \mod 3 \Longrightarrow 2^n + 1 \equiv 0 \mod 3 \Longrightarrow 3|2^n+1$$

c) $k = 0,1,2,\cdots$ for $n \in Z$, we have $2n+1|1^{2k+1}+2^{2k+1}+\cdots+(2n)^{2k+1}$.
Proof: For each $i = 1,2,\cdots,n$, we have

$$i + (2n+1) - i \equiv 2n+1 \equiv 0 \mod 2n+1$$

$$i \equiv -(2n+1-i) \mod 2n+1 \Longrightarrow i^{2k+1} \equiv [-(2n+1-i)^{2k+1}] \mod 2n+1$$

Since, $2k+1$ is odd, we have

$$i^{2k+1} + (2n+1-i)^{2k+1} \equiv 0 \mod 2n+1$$

$$\sum_{i=1}^{n}[i^{2k+1} + (2n+1-i)^{2k+1}] \equiv 0 \mod 2n+1$$

Therefore,

$$2n+1|1^{2k+1}+2^{2k+1}+\cdots+(2n-1)^{2k+1}+(2n)^{2k+1}$$

d) if $m-p|mn+pq$, then $m-p|mq+np$.
Proof: Since

$$(m-p)|(m-p)(n-q) \Longrightarrow (m-p)|mn+pq-(mq+np)$$

Because $m-p|mn+pq$, we have $m-p|mq+np$.

e) if $x \equiv 1 \bmod m^k$, then $x^m \equiv 1 \bmod m^{k+1}$.
Proof: Since $x \equiv 1 \mod m^k$, we have $x = 1 + k \cdot m^k = 1 + (k \cdot m^{k-1}) \cdot m$. Thus,

$$m^k|x-1, \qquad x \equiv 1 \mod m$$

From $x \equiv 1 \mod m$, we have $x^i \equiv 1^i \mod m$ for $i = 0,1,\cdots,m-1$. Then,

$$\sum_{i=1}^{m-1} x^i \equiv \sum_{i=1}^{m-1} 1^i \mod m$$

$$1 + x + x^2 + \cdots + x^{m-1} \equiv m \mod m \Longrightarrow 1 + x + x^2 + \cdots + x^{m-1} \equiv 0 \mod m$$

Then,

$$m|(1 + x + x^2 + \cdots + x^{m-1})$$

Finally, we have

$$m^k \cdot m|(x-1)(1 + x + x^2 + \cdots + x^{m-1}) \Longrightarrow m^{k+1}|x^m - 1 \Longrightarrow x^m \equiv 1 \mod m^{k+1}.$$

10) Prove the Fermat's Little Theorem, i.e., $a^{p-1} \equiv 1 \bmod p$, where p is a prime, and $a < p$.
Proof: Suppose that $Z_p^* = \{1,2,3\cdots,p-1\}$ and $B = \{a*1, a*2, a*3\cdots, a*(p-1)\}$ for any $a \in Z_p^*$. We need to prove $|Z_p^*| = |B|$ or there is not redundant element in B, so the p-1 multiples of a in B are distinct and nonzero.
By contradiction, if $i \neq j$, $a*i \equiv a*j \mod p$, $a*(i-j) \equiv 0 \mod p$. Then, $a \equiv 0 \mod p$ or $i-j \equiv 0 \mod p$.
As we know, $gcd(a,p) = 1$, thus $a \neq 0 \mod p$ and $i-j \equiv 0 \mod p$. Thus, $i = j$.
Now we know that $|Z_P^*|$ and B have the same number of elements, and try to calculate the multiplication of element in Z_p^* and B.

$$\prod_{x_i \in Z_p^*} x_i \equiv \prod_{x_i \in Z_p^*} a * x_i \mod p$$

$$\prod_{x_i \in Z_p^*} x_i \equiv 1 * 2 * 3 * \cdots * (p-1) \mod p$$

$$\prod_{x_i \in Z_p^*} a * x_i \equiv (a*1)*(a*2)*\cdots*(a*(p-1)) \equiv a^{p-1}*(1*2*3*\cdots*(p-1)) \mod p$$

Assume, $1*2*3*\cdots*(p-1) \mod p = \beta$. Then,
Then, $\beta \equiv (a^{p-1}-1)*\beta \mod p$, so $(a^{p-1}-1)*\beta \equiv 0 \mod p$. Thus, $a^{p-1} \equiv 1 \mod p$.

11) What's the difference between the public key encryption and the symmetric encryption?
• Answer: In public key encryption, there are two different keys, i.e., public key and private key. In the encryption algorithm, public key is published and used for encryption. Private key is kept secret and used for decryption. However, in symmetric key, one same key is used for both encryption and decryption.

12) In a public-key system using RSA, you intercept the ciphertext $C = 8$ sent to a user whose public key is $e = 5, n = 35$. What is the plaintext M?
• Answer: Since $n = 35$, we have $p = 5$ and $q = 7$.
Then,

$$\phi(n) = (p-1)(q-1) = 24$$

As we know $ed \equiv 1 \mod \phi(n)$, so

$$d \equiv e^{-1} \mod \phi(n)$$

$$d \equiv 5^{-1} \equiv 5 \mod 24$$

According to the RSA decryption algorithm, $M \equiv C^d \equiv \mod n$.
Therefore,

$$M \equiv 8^5 \equiv 8 \mod 35$$

13) Use the Chinese Remainder Theorem (CRT) to solve x, where

$$\begin{cases} x & \equiv & 2 \text{ mod } 3 \\ x & \equiv & 3 \text{ mod } 5 \\ x & \equiv & 4 \text{ mod } 7 \end{cases}$$

• Answer: Let $m_1 = 3, m_2 = 5, m_3 = 7$, $a_1 = 2, a_2 = 3, a_3 = 4$. We have $M = m_1 \cdot m_2 \cdot m_3 = 3 \times 5 \times 7 = 105$.
$M_1 = M/m_1 = 35, M_2 = M/m_2 = 21, M_3 = M/m_3 = 15$.
$\alpha_1 = M_1^{-1} \mod m_1 = 35^{-1} \mod 3 = 2, \alpha_2 = M_2^{-1} \mod m_2 = 21^{-1} \mod 5 = 1, \alpha_3 = M_3^{-1} \mod m_3 = 15^{-1} \mod 7 = 1$.
Therefore,

$$x = a_1 \cdot \alpha_1 \cdot M_1 + a_2 \cdot \alpha_2 \cdot M_2 + a_3 \cdot \alpha_3 \cdot M_3 = 2 \times 2 \times 35 + 3 \times 1 \times 21 + 4 \times 1 \times 15 \mod M = 53$$

14) Consider a Diffie-Hellman scheme with a common prime $q = 11$ and a primitive root $\alpha = 2$.
- If user A has public key $Y_A = 3$, what is A's private key X_A?
  • Answer: $X_A = 8$, because $2^8 \mod 11 = 3$.
- If user B has public key $Y_B = 8$, what is the secret key K shared with A?
  • Answer: Since $K = Y_B^{X_A} \mod q$, we have $K = 8^8 \mod 11 = 5$.

15) Consider an Elgamal encryption scheme with a common prime $q = 11$ and a primitive root $\alpha = 2$. If B has public key $Y_B = 3$ and A choose the random integer $k = 2$, what is the ciphertext of $M = 9$?
• Answer: (4,4)
Since,

$$C_1 = \alpha^k = 2^2 = 4 \mod 11 = 4$$

$$C_2 = M \cdot Y_B^k = 9 \cdot 3^2 = 4 \mod 11 = 4$$

16) What is the Man-In-The-Middle attack? How to deal with the Man-In-The-Middle attack in the Diffie-Hellman key exchange protocol?
Answer:
- Man-In-The-Middle attack: When Alice and Bob run the Diffie-Hellman key exchange protocol. An attacker Darth can attack the protocol as follows.
  - Darth prepares for the attack by generating random numbers X_D, X'_D and then computing the corresponding public key Y_D, Y'_D.
  - Alice transmits Y_A to Bob.

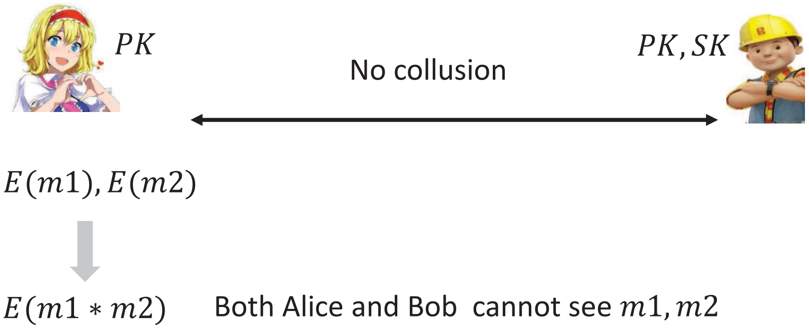
  - Darth intercepts Y_A and transmits Y_D to Bob.
  - Bob receives Y_D and calculates $K_2 = (Y_D)^{X_B} \mod q$.
  - Bob transmits Y_B to Alice.
  - Darth intercepts Y_B and calculates $K_2 = (Y_B)^{X_D} \mod q$.
  - Darth transmits Y'_D to Alice. Darth also calculates $K_1 = (Y_A)^{X'_D} \mod q$.
  - Alice receives Y'_D and calculates $K_1 = (Y'_D)^{X_A} \mod q$. Then, Darth and Alice share the key K_1. At the same time, Darth and Bob share the key K_2.

- Solution: When running the key exchange protocol, each party signs its own DH value to prevent man-in-the-middle attack. The specific protocol is as follows.
  - Alice transmits ID_A, Y_A to Bob.
  - Bob receives ID_A, Y_A and calculates $K = (Y_A)^{X_B}$. Then, he/she returns $ID_B, Y_B, SIG_B(ID_A, ID_B, Y_A, Y_B)$ to Alice.
  - Alice verifies the received message and calculates $K = (Y_B)^{X_A}$. Then, he/she sends $SIG_A(ID_A, ID_B, Y_A, Y_B)$ to Bob.
  - Bob verifies the identity of Alice. Then, Alice and Bob share the key K.

Tutorial 4

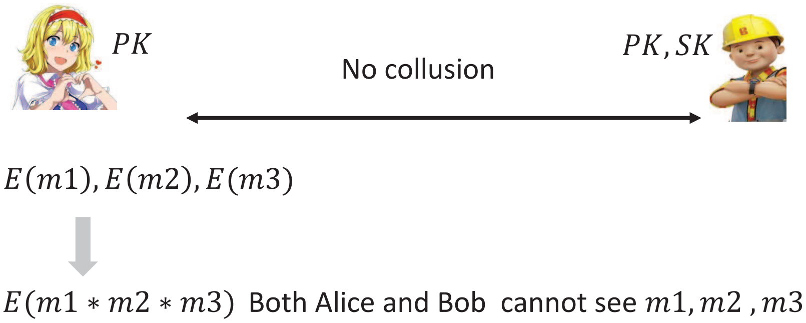
1) What is the homomorphic encryption technique?
   - Answer: Homomorphic encryption is a form of encryption that allows computation on ciphertexts, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the plaintexts.

2) Let $E()$ be a Paillier homomorphic encryption scheme. Assume Bob has the public/private key pair (pk, sk) of $E()$, and Alice only has the public key pk of $E()$. Consider there is no collusion between Alice and Bob. When Alice has two ciphertexts $E(m_1)$ and $E(m_2)$, please design a protocol run between Alice and Bob. With the protocol, Alice can finally obtain the ciphertext $E(m_1 \cdot m_2)$ while both Alice and Bob have no idea on the plaintexts m_1 and m_2.



   - Answer:
     - Alice chooses two random numbers r_1 and r_2. Then, it can compute $E(m_1+r_1) = E(m_1) * E(r_1)$, $E(m_2+r_2) = E(m_2) * E(r_2)$. At the same time, it sends $E(m_1+r_1)$ and $E(m_2+r_2)$ to Bob.
     - On receiving $E(m_1+r_1)$ and $E(m_2+r_2)$, Bob recovers the plaintext m_2+r_2 with private key sk. Then, Bob computes $E((m_1+r_1)*(m_2+r_2)) = E(m_1+r_1)^{m_2+r_2}$ and sends it to Alice.
     - Alice receives $E((m_1+r_1)*(m_2+r_2))$. Because $(m_1+r_1)(m_2+r_2) = m_1*m_2 + m_1*r_2 + m_2*r_1 + r_1*r_2$, we have $m_1*m_2 = (m_1+r_1)(m_2+r_2) - m_1*r_2 - m_2*r_1 - r_1*r_2$. Then, $E(m_1*m_2) = E((m_1+r_1)*(m_2+r_2)) * (E(m_1)^{r_2})^{-1} * (E(m_2)^{r_1})^{-1} * E(r_1*r_2)^{-1}$. Thus, Alice obtain ciphertext $E(m_1*m_2)$.

3) Let $E()$ be a BGN homomorphic encryption scheme. Assume Bob has the public/private key pair (pk, sk) of $E()$, and Alice only has the public key pk of $E()$. Consider there is no collusion between Alice and Bob. When Alice has three ciphertexts $E(m_1)$, $E(m_2)$ and $E(m_3)$, please design a protocol run between Alice and Bob. With the protocol, Alice can finally obtain the ciphertext $E(m_1 \cdot m_2 \cdot m_3)$ while both Alice and Bob have no idea on the plaintexts m_1, m_2, and m_3.



PK
No collusion
PK, SK
$E(m1), E(m2), E(m3)$
$E(m1 * m2 * m3)$ Both Alice and Bob cannot see $m1, m2, m3$

   - Answer:
     - Alice chooses two random numbers r_1 and r_2. Then, it can compute $E(m_1+r_1) = E(m_1) * E(r_1)$, $E(m_2+r_2) = E(m_2) * E(r_2)$. At the same time, it sends $E(m_1+r_1)$ and $E(m_2+r_2)$ to Bob.
     - On receiving $E(m_1+r_1)$ and $E(m_2+r_2)$, Bob recovers the plaintext m_2+r_2 with private key sk. Then, Bob computes $E((m_1+r_1)*(m_2+r_2)) = E(m_1+r_1)^{m_2+r_2}$ and sends it to Alice.
     - Alice receives $E((m_1+r_1)*(m_2+r_2))$. Because $(m_1+r_1)(m_2+r_2) = m_1*m_2 + m_1*r_2 + m_2*r_1 + r_1*r_2$, we have $m_1*m_2 = (m_1+r_1)(m_2+r_2) - m_1*r_2 - m_2*r_1 - r_1*r_2$. Then, $E(m_1*m_2) = E((m_1+r_1)*(m_2+r_2)) * (E(m_1)^{r_2})^{-1} * (E(m_2)^{r_1})^{-1} * E(r_1*r_2)^{-1}$. Thus, Alice obtain ciphertext $E(m_1*m_2)$.
     - Then, Alice calculates $e(E(m_1*m_2), E(m_3)) = E'(m_1*m_2*m_3)$. Thus, Alice can finally obtain the ciphertext $E'(m_1 \cdot m_2 \cdot m_3)$.

Tutorial 5

1) What is the Tor protocol? How can it enable the anonymous communication?

   • Answer: Tor is an Internet networking protocol designed to anonymize the data relayed across it. Using Tor makes it more difficult to trace Internet activity to the user. Tor's intended use is to protect the personal privacy of its users, as well as their freedom and ability to conduct confidential communication by keeping their Internet activities from being monitored.

   The Tor network runs through the computer servers of thousands of volunteers spread throughout the world. Your data is bundled into an encrypted packet when it enters the Tor network. Then, unlike the case with normal Internet connections, Tor strips away part of the packet's header, which is a part of the addressing information that could be used to learn things about the sender such as the operating system from which the message was sent. Finally, Tor encrypts the rest of the addressing information, called the packet wrapper. Regular Internet connections don't do this.

2) What is the Onion routing?

   • Answer: Onion routing is like an advanced form of proxy routing. Instead of routing through a single unprotected server, it uses a network of nodes that constantly encrypt your data packets at every step. Only at the end of this chain of onion nodes, your data become decrypted and sent to the final destination. In fact, only this exit node has the power to decrypt your message, so no other node can even see what you're sending.

3) What is the message frequency attack in anonymous communication system?

   • Answer: These types of attacks are similar to the other contextual attacks in that they exploit the fact that some communication are easy to distinguish from others. If a participant sends a non-standard (i.e., unusual) number of messages, a passive external attacker can spot these messages coming out of the mix-networks. In fact, unless all users send the same number of messages, this type of attack allows the adversary to gain non-trivial information.

4) What is the k-Anonymous message transmission?

   • Answer: k-Anonymous message transmission means that the adversary is able to learn something about the origin or destination of a particular message, but cannot narrow down its search to a set of less than k participants.

5) What is the Dining cryptographers network?

   • Answer: A Dining cryptographers network is an anonymous communication network. It consists of a set of participants that have mutually agreed on a secret and send messages to each other in every round. All round messages are summed together and the result is the message content sent by one of the participants. The sender is not identifiable in the set of participants (sender anonymity). Since the message content is broadcasted, all participants are receivers (receiver anonymity)

6) What is the "randomized" attack in k-Anonymous message transmission?

   • Answer: The attacker does not follow the protocol. After receiving shares $S_{1,i}, \cdots, S_{k,i}$, instead of broadcasting S_i, it generates a random value r and broadcasts that instead. This will randomize the result of the DC-Net protocol.

7) What is the Pedersen Commitment Scheme? How to apply it to implement an electronic soccer lottery?

How many $5 plays do you want?
Each $5 play includes your 7 number selection plus 2 Quick Pick selections
PLAY

   • Answer: The Pedersen commitment scheme runs between a committer C, holding a secret message $m \in Z_q$ to commit to, and a receiver R. It consists of three algorithms, i.e., Setup, Commit and Reveal.

   - Setup: Two large primes p, q are selected such that $q|p-1$. Z_q^* is a group and g is a generator of Z_p^*. Choose a random value a from Z_q and let $h = g^a \mod p$. Then, publish p, q, g, h and keep a as a secret.
   - Commit: To commit to some $x \in Z_q$, sender chooses random $r \in Z_q$ and sends $c = g^x h^r \mod p$ to receiver.
   - Reveal: To open the commitment, sender reveals x and r, and receiver verifies that $c = g^x h^r \mod p$.

The electronic soccer lottery can use Pedersen commitment scheme to implement. When a user intends to buy the lottery x, he/she selects a random number and makes a commitment c by running Commit algorithm. Then, he/she sends the commitment c to agency. When the user wins the prize, he/she has to provide x and r with the agency. Agency verifies whether the user actually wins the prize by comparing the commitment c and $g^x h^r \mod p$. If it is same, verification succeeds. Otherwise, verification fails.

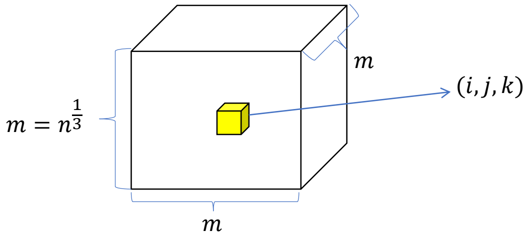
Tutorial 6

1) What is private information retrieval (PIR)?

   • Answer: In cryptography, a private information retrieval (PIR) protocol is a protocol that allows a user to retrieve an item from a server in possession of a database without revealing which item is retrieved.

2) How to design an Information-Theoretic 2-Server PIR protocol with communication costs $O(n^{1/3})$?

   • Answer: Suppose that each server holds n bit strings. Let $m = n^{\frac{1}{3}}$ and the string can be represented as follows. Each bit in the string has 3 coordinates, i.e., (i, j, k). If a user wants to retrieve item (i, j, k) with privacy, it can do as follows.

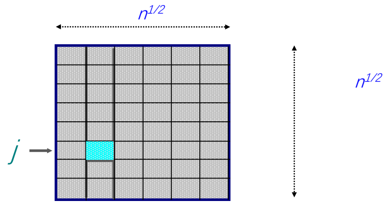


   - First, user randomly chooses $Q_{1x} \in \{0,1\}^m, Q_{1y} \in \{0,1\}^m, Q_{1z} \in \{0,1\}^m$.
   - Second, user sends (Q_{1x}, Q_{1y}, Q_{1z}) to server S_1 and sends $(Q_{2x} = Q_{1x} + 2^i, Q_{2y} = Q_{1y} + 2^j, Q_{2z} = Q_{1z} + 2^k)$ to server S_2.
   - For each $x = 1, 2, \cdots, m$, server S_1 computes sums all nodes where $y \in Q_{1y}, z \in Q_{1z}$. Similarly, S_1 sums all nodes for each $y = 1, 2, \cdots, m$ where $x \in Q_{1x}, z \in Q_{1z}$ and sums all nodes for each $z = 1, 2, \cdots, m$ where $x \in Q_{1x}, y \in Q_{1y}$. Then, S_1 returns the result to user and let S_{xyz} denote the result.
   - S_2 does the same computation as S_1. At the same time, S_2 returns the computation result to user and let S'_{xyz} denote the result.
   - When receiving the results from S_1 and S_2, user can compute the item (i, j, k) as follows.
     - Choose $x = i$ and compute $S_{iyz} \oplus S'_{iyz}$.
     - Choose $y = j$ and compute $S_{xjz} \oplus S'_{xjz}$.
     - Choose $z = k$ and compute $S_{xyk} \oplus S'_{xyk}$.
     - Compute $X(i, j, k) = (S_{iyz} \oplus S'_{iyz}) \oplus (S_{xjz} \oplus S'_{xjz}) \oplus (S_{xyk} \oplus S'_{xyk})$

   The communication cost from user to two servers is $3m + 3m = 6m$. At the same time, the communication cost from two servers to user is also $3m + 3m = 6m$. Thus, the communication cost in total is $O(m)$, i.e., $O(n^{\frac{1}{3}})$.

3) How to design a computational 1-server PIR protocol with communication costs $O(n^{1/2})$?

   • Answer: The 1-server PIR protocol with communication costs $O(n^{1/2})$ can be achieved by Paillier homomorphic encryption technique. Suppose that the server holds n bit strings. Let $m = n^{\frac{1}{2}}$ and the string can be represented as follows. Each bit in the string has 2 coordinates, i.e., (i, j). If a user wants to retrieve item (i, j) with privacy, it can do as follows.

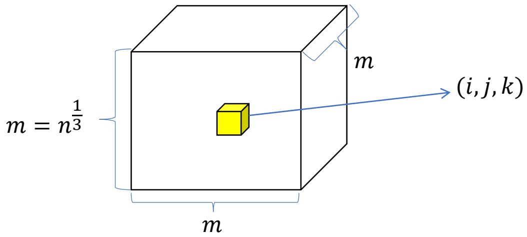


   - User sends $\{C_k = E(0), C_i = E(1) | k = 1, 2, \cdots, m, (k \neq i)\}$ to server.
   - Sever computes a "bit" for each row and returns to user. Suppose that data in the r-th row is $\{x_1, x_2, \cdots, x_m\}$. Then, server will compute one bit for the r-th row as $b_r = \prod_{k=1}^{m} C_k^{x_k}$. Finally, server returns $\{b_r | r = 1, 2, \cdots, m\}$ to user.
   - User recovers the item (i, j) from b_j by decryption.

   The communication cost from user to two servers is m. At the same time, the communication cost from two servers to user is also m. Thus, the communication cost in total is $O(m)$, i.e., $O(n^{\frac{1}{2}})$.

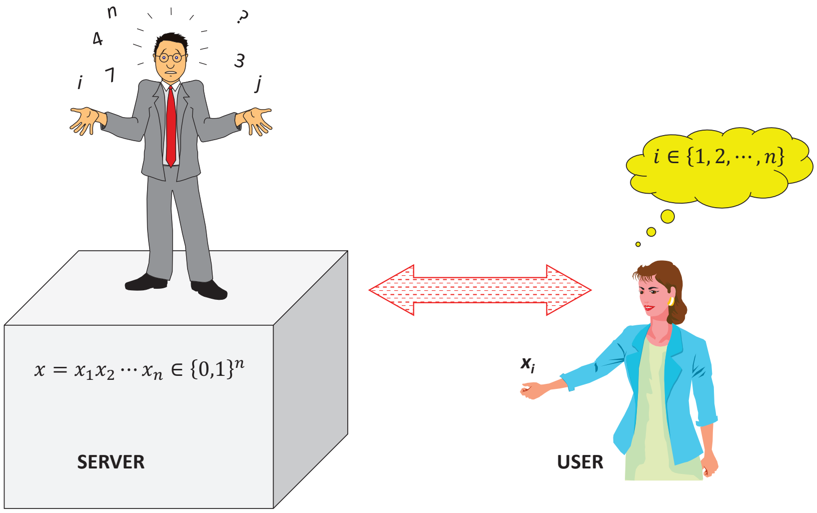
4) How to design a computational 1-server PIR protocol with communication costs $O(n^{1/3})$?

   • Answer: The 1-server PIR protocol with communication costs $O(n^{1/3})$ can be achieved by BGN homomorphic encryption technique, which satisfies $e(E(a), E(b)) = E_T(a \cdot b)$. Suppose that a server holds n bit strings. Let $m = n^{\frac{1}{3}}$ and the string can be represented as follows. Each bit in the string has 3 coordinates, i.e., (i, j, k). If a user wants to retrieve item (i, j, k) with communication costs $O(n^{1/3})$, it can do as follows.



- User sends $Str_1 = \{C_k = E(0), C_i = E(1)|k = 1, 2, \cdots, m, (k \neq i)\}$ and $Str_2 = \{C_k = E(0), C_j = E(1)|k = 1, 2, \cdots, m, (k \neq j)\}$ to server.
- For $z = 1, 2, \cdots, m$, sever computes a "bit" value. Let V_{xyz} denote the value in (x, y, z). Then, for each $z = 1, 2, \cdots, m$, server will compute one bit $b_z = \prod_{x,y \in 1,2,\cdots,m} e(C_x, C_y)^{V_{xyz}}$. Finally, server returns $\{b_z|z = 1, 2, \cdots, m\}$ to user.
- User recovers the item (i, j, k) from b_k by decryption.

The communication cost from user to server is $2m$. At the same time, the communication cost from server to user is also m. Thus, the communication cost in total is $O(m)$, i.e., $O(n^{\frac{1}{3}})$.


n
?
4
3
i
7
j
$i \in \{1, 2, \cdots, n\}$
$x = x_1 x_2 \cdots x_n \in \{0,1\}^n$
x_i
SERVER
USER

Tutorial 7

1) What is Oblivious Transfer (OT) protocol?

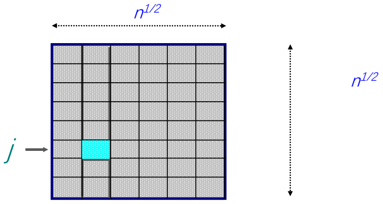
   • Answer: Oblivious Transfer (OT) protocol is a protocol by which sender sends information to receiver, but remains oblivious as to what is received.

2) What is the difference between OT protocol and PIR protocol?

   • Answer: In both OT and PIR, server cannot know x_i. For the user, it can only x_i in OT, but it can know x_i and others in PIR.

3) How to use the Paillier encryption to design 1-out-n OT with $O(n^{1/2})$ Communication costs?

   • Answer: Suppose that Alice holds n messages and has Paillier's public key pk and private key sk, and Bob only has public key pk. Let $l = n^{\frac{1}{2}}$ and these messages can be represented as follows. Each message is denoted by $x_{i,j}$. The OT protocol can be run as follows.

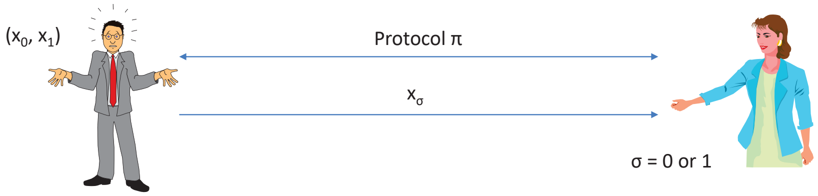


   - Bob selects $i, j \in \{1, 2, \cdots, l\}$. Then, he/she computes $S = \{C_k = E(0), C_j = E(1) | k = 1, 2, \cdots, m, (k \neq j)\}$ and returns S to Alice.
   - After receiving message from Bob, Alice first computes a value for each row. In specific, for the k-th row, the value $E(x_{k,j})$ can be computed as $E(x_{k,j}) = \prod_{t=1}^{l} C_t^{x_{k,t}}$. Then, Alice will have l values $\{E(x_{1,j}), E(x_{2,j}), \cdots, E(x_{l,j})\}$, and return them to Bob.
   - Bob chooses a random number w and computes $E(x_{i,j} + w) = E(x_{i,j}) \cdot E(w)$ and returns it to the Alice.
   - Alice recovers $x_{i,j} + w$ by decryption and returns it to the Bob.
   - Bob obtains $x_{i,j} = x_{i,j} + w - w$.

   Thus, the communication cost in total is $O(l)$, i.e., $O(n^{\frac{1}{2}})$.

4) What is the weakness in the above Paillier based OT protocol? How to design a stronger 1-out-n OT with $O(n^{1/2})$ Communication costs?

   • Answer:



Weakness: In the above Paillier based OT protocol, the weakness is that Bob must honestly follow the protocol. Otherwise, he/she can obtain other values except $x_{i,j}$. For example, when Bob returns $E(x_{i,j} + w)$ to Alice, he/she may return $E(x_{i,j} + 10000x_{k,j} + w)$. In this case, Bob can compute $x_{i,j} + 10000x_{k,j}$ on receiving the decryption result from Alice, and then he/she can obtain $x_{i,j} = x_{i,j} + 10000x_{k,j} \mod 10000$ and $x_{k,j} = \frac{x_{i,j} + 10000x_{k,j} - x_{i,j}}{10000}$.

Stronger scheme: We can use BGN homomorphic encryption scheme to design a stronger 1-out-n OT with $O(n^{1/2})$ communication cost. The protocol is as follows.

   - Bob computes $S_i = \{(C_{i,1}, W_{i,1}), (C_{i,2}, W_{i,2}), \cdots, (C_{i,l}, W_{i,l})\}$, where $C_{i,k} = E(0)(1 \leq k \neq i \leq l)$ and $C_{i,i} = E(1)$. At the same time, $W_{i,k}$ is used to proof that $C_{i,k}$ is either $E(0)$ or $E(1)$, where $k = 1, 2, \cdots, l$. In specific, suppose that if $C_{i,k} = g^m h^r$, the proof of $W_{i,k}$ is $g^{1-2m} h^{-r}$. Then, the statement of whether $C_{i,k}$ is either $E(0)$ or $E(1)$ can be checked by verifying whether $e(C_{i,k}, gC_{i,k}^{-1})$ is equal to $e(h, W_{i,k})$. In addition, Bob needs to compute a proof $W_{i,l+1}$ for $\prod_{k=1}^{l} C_{i,k}$, which is used to proof that there is only one $E(1)$ in S_i. With the same method, Bob can compute $S_j = \{(C_{j,1}, W_{j,1}), (C_{j,2}, W_{j,2}), \cdots, (C_{j,l}, W_{j,l})\}$ and $W_{j,l+1}$ as a proof for $\prod_{k=1}^{l} C_{j,k}$. Then, he/she will send S_i, $W_{i,l+1}$, S_j and $W_{j,l+1}$ to the Alice.
   - After receiving message from Bob, Alice first verifies each value in S_i and S_j is either $E(0)$ or $E(1)$, and there is only one $E(1)$ in both S_i and S_j. Then, Alice computes $E'(x_{i,j}) = \prod_{k=1}^{l} \prod_{t=1}^{l} e(C_{i,k}, C_{j,t})^{x_{i,j}}$ and returns $E'(x_{i,j})$ to Bob.
   - Bob recovers $x_{i,j}$ by decryption.

Tutorial 8

1) What is zero knowledge (ZK) proof?

   • Answer: A zero-knowledge proof protocol allows one party, usually called PROVER, to convince another party, called VERIFIER, that PROVER knows some facts (a secret, a proof of a theorem,...) without revealing to the VERIFIER any information about his knowledge (secret, proof,...).

2) Please formally describe the Fiat-Shamir identification scheme, and show why it is ZK?

   • Answer: In the identification protocol, the verifier wishes to authenticate the identity of the prover, which is claimed to have a public key I. Thus, he requests the prover to convince him that he knows the secret key S corresponding to I. The scheme is as follows.

   - The prover chooses a random value $1 < R < n$ and computes $X = R^2 \mod n$. Then, the prover sends X to the verifier.
   - The verifier requests from the prover one of the following requests at random: R or $RS \mod n$.
   - The prover sends the requested information to the verifier.
   - The verifier verifies that he received the correct answer by checking whether: $R^2 \equiv ?X \mod n$ or $(RS)^2 \equiv ?IX \mod n$.
   - If the verification fails, the verifier concludes that the prover does not know S, and thus he is not the claimed party.
   - This protocol is repeated t times, and if in all of them the verification succeeds, the verifier concludes that the prover is the claimed party.

   In the following, we prove that the Fiat-Shamir scheme is zero knowledge by giving a simulator for the problem.

   The input for the simulator are numbers I, N, which the prover claims to know the square root of $I \mod N$. The output of the simulator is a forged transcript of a proof. A transcript for the problem is of the form:

   $$(I, N)(X_1, i_1, M_1), \cdots (X_n, i_n, M_n)$$

   where M_k is either the square root of X_k in case $i_k = 1$, or the square root of IX_i in case $i_k = 2$.

   The transcript generated by the simulator is from the same probability distribution as the protocol runs. Thus, it is ZK.

3) Why is Parallel Fiat-Shamir identification not ZK?

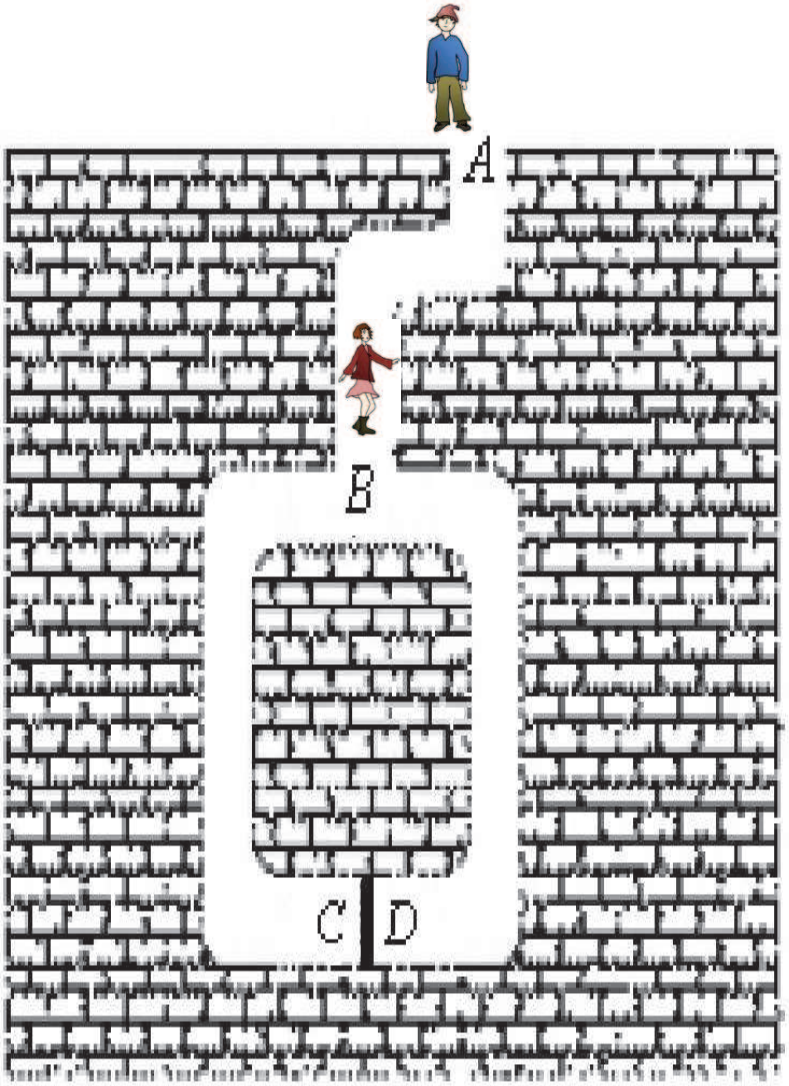
   • Answer: In parallel Fiat-Shamir identification, all rounds can be run in parallel.

   - The prover chooses random values $R_1, R_2, \cdots, R_t$ and computes $X_1 = R_1^2 \mod n, \cdots, X_t = R_t^2 \mod n$. Then, the prover sends $X_1, \cdots, X_t$ to the verifier.
   - The verifier sends t random bits $i_1, i_2, \cdots, i_t$, where $i_k = 0, 1$ mean the verifier requests R_k and $R_k S$ respectively.
   - The prover sends the requested information $Z_1, Z_2, \cdots, Z_t$ to the verifier.
   - The verifier accepts if $Z_k^2 \equiv ?X_k I^{i_k} \mod n$.

   However, this protocol is no longer ZK. Consider the V^* such that V^* chooses $i_1, i_2, \cdots, i_t$ to be the first t bits of $H(X_1, X_2, \cdots, X_t)$, where H is a cryptographic hash function. One can no longer generate an indistinguishable transcript.

4) How to design non-interactive zero-knowledge proof?

   • Answer: For the non-interactive zero-knowledge proof, the prover can publish the proof and anyone can verify the proof. In a normal ZK proof, the prover first issues a bunch of commitments, then the verifier issues challenges that the prover complies with. This proves anything only as long as the verifier is assumed to issue challenges normally without any prior understanding with the prover. In a non-interactive ZK proof, the verifier can be replaced by a hash function (or something similar) which is computed over the whole set of commitments: the hash function result is the challenge. If the hash function is really a random oracle then the prover cannot guess its output before trying it, i.e. before having produced his commitments, and that's where the security comes from.


A
B
C
D

Tutorial 9

1) What is the secret handshake?

   • Answer: The secret handshake considers a CIA agent who wants to authenticate herself to a server, but does not want to reveal her CIA credentials unless the server is a genuine CIA outlet. It also considers that the CIA server does not want to reveal its CIA credentials to anyone but CIA agents C not even to other CIA servers.

2) How to use the Paillier encryption to implement one privacy-preserving friend match protocol?

   • Answer: Suppose Alice has an interests list $A = \{a_1, a_2, \cdots, a_N\}$ and Bob has an interests list $B = \{b_1, b_2, \cdots, b_N\}$. If they have more than or equal to t common interests, they can run the friend match protocol to make friends as follows.

   - Alice computes $c_i = E(a_i)$ for $i = 1, 2, \cdots, N$, and sends $\{c_i | i = 1, 2, \cdots, N\}$ to Bob.
   - Bob calculates $D = \prod_{i=1}^{N} c_i^{b_i}$ and returns it to Alice.
   - Alice decrypts D. If the result is more than or equal to t, make friend. Otherwise, stop.

3) How to design a non OT-based Private Equality Test protocol?

   • Answer: The homomorphic encryption techniques can be used to achieve private equality test protocol. Suppose that Alice and Bob have values x and y respectively. Alice can test the equality of x and y as follows.

   - Suppose that there is a homomorphic public encryption algorithm. Alice has public key pk and private key sk. Bob only has the public key pk.
   - Alice sends $E(x)$ and $E(y)$ to Bob.
   - Bob computes $E(r(x-y)) = (\frac{E(x)}{E(y)})^r$ using the homomorphic property, where r is a random number. Then, Bob returns $E(r(x-y))$ to Alice.
   - Alice recovers $r(x-y)$. If it is equal to 0, x is equal to y. Otherwise, x is not equal to y.

   Note that in this protocol, except the comparison result, Alice and Bob do not obtain any other information.

Tutorial 10

1) What is cloud computing?

   • Answer: Cloud computing is a general term used to describe a new class of network based computing that takes place over the Internet, and it has the following features.

   - basically a step on from Utility Computing.
   - a collection/group of integrated and networked hardware, software and Internet infrastructure (called a platform).
   - using the Internet for communication and transport provides hardware, software and networking services to clients.

2) What are the characteristics of cloud computing?

   • Answer:

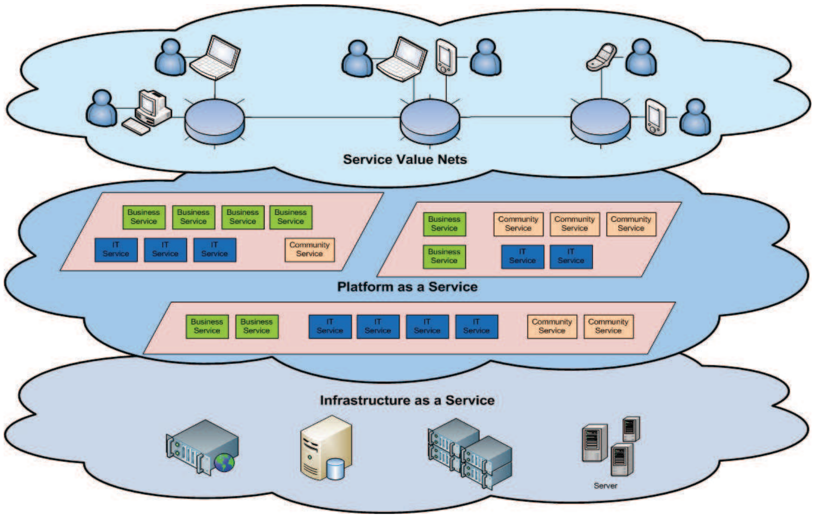
   - On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each services provider.
   - Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
   - Resource pooling and multi-tenant model: There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.
   - Rapid elasticity: Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.
   - Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts).

3) What are the security challenges in cloud computing?

   Answer: Most security challenges stem from consumer's loss of control, lack of trust in the cloud and multi-tenancy issues in the cloud.

   - Consumer's loss of control
     - Data, applications, resources are located with provider.
     - User identity management is handled by the cloud.
     - User access control rules, security policies and enforcement are managed by the cloud provider.
     - Consumer relies on provider to ensure data security, privacy, resource availability and monitoring and repairing of services/resources.
   - Lack of trust in the cloud: trusting a third party requires taking risks.
   - Multi-tenancy issues in the cloud: Multiple independent users share the same physical infrastructure. Thus, an attacker can legitimately be in the same physical machine as the target.

4) What are the key privacy concerns in cloud computing?



   Answer: The key privacy concerns in cloud computing include storage, retention, destruction, auditing, monitoring and risk management.

- Storage: Information from different organizations may be commingled together when they use the same CSP. At the same time, the aggregation of data raises new privacy issues. Some governments may decide to search through data without necessarily notifying the data owner, depending on where the data resides. It is also important to control the storage access of the cloud provider, i.e., whether the cloud provider itself has any right to see and access customer data or not. In addition, some services today track user behaviour for a range of purposes, from sending targeted advertising to improving services.
- Retention: It concerns how long the personal information (that is transferred to the cloud) retained and the related policy that governs the data. At the same time, there are other challenges, e.g., whether the organization or the CSP own the data, who enforces the retention policy in the cloud, how exceptions to this policy are managed, etc.
- Destruction: This concerns several problems. The first one is how the cloud server destroys PII at the end of the retention period. The second one is how organizations ensure that their PII is destroyed by the CSP at the right point and is not available to other cloud users. The third one is how the organizations know the CSP really destroyed the data or just makes it inaccessible to the organizations, how the organizations know the CSP does not retain additional copies and whether the CSP keeps the information longer than necessary so that it can min the data for its own use.
- Auditing, monitoring and risk management: This concerns several problems. The first is how organizations monitor their CSP and provide assurance to relevant stakeholders that privacy requirements are met when their PII is in the cloud. The second one is whether they are regularly audited. The third one is if business-critical processes are migrated to a cloud computing model, internal security processes need to evolve to allow multiple cloud providers to participate in those processes, as needed. These include processes such as security monitoring, auditing, forensics, incident response, and business continuity.

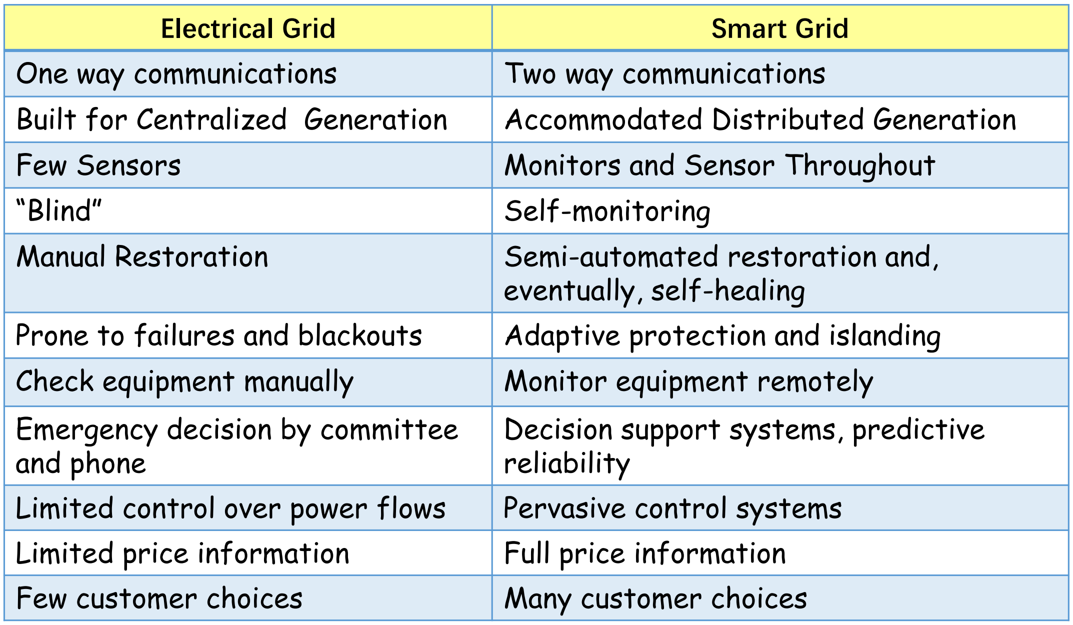
Tutorial 11

1) What is smart grid?

   ● Answer: Smart grid is an intelligent power system, which involves advanced technology, digital communications, sensing, measurement and control technologies etc. In addition, it integrates renewable sources and electric vehicles to achieve above goals.

2) What are the characteristics of smart grid?

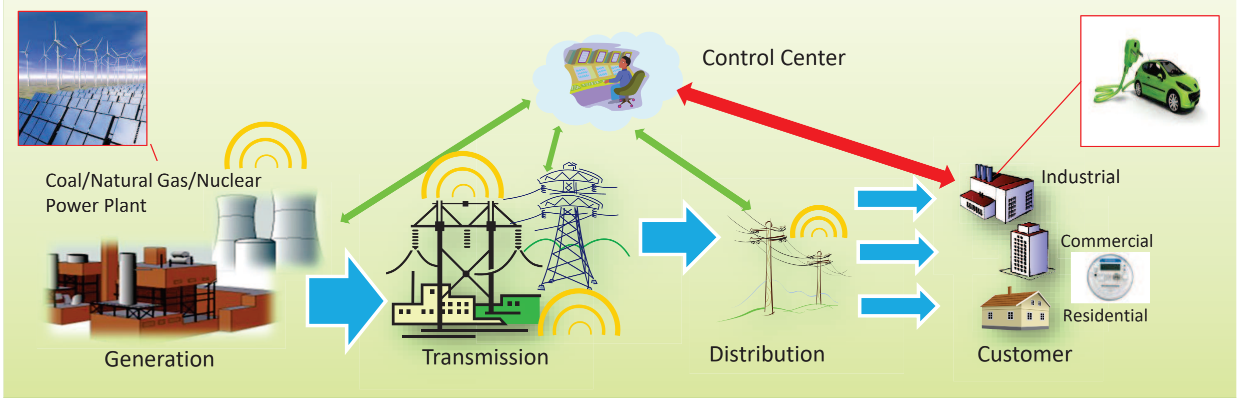
   ● Answer: Compared with the electrical grid, smart grid has the following characteristics.



Electrical Grid	Smart Grid
One way communications	Two way communications
Built for Centralized Generation	Accommodated Distributed Generation
Few Sensors	Monitors and Sensor Throughout
"Blind"	Self-monitoring
Manual Restoration	Semi-automated restoration and, eventually, self-healing
Prone to failures and blackouts	Adaptive protection and islanding
Check equipment manually	Monitor equipment remotely
Emergency decision by committee and phone	Decision support systems, predictive reliability
Limited control over power flows	Pervasive control systems
Limited price information	Full price information
Few customer choices	Many customer choices

3) What are the security and privacy challenges in smart grid?
   Answer:

   - Security challenges:
     - The information flow is vulnerable to the Cyber attacks.
     - Former CIA Director James Woolsey said the federal government's oversight of grid security is inadequate and attacks on the grid are entirely possible.
     - Smart grid is a double-edged sword.
   - Privacy challenges: Personal information can be inferred:
     - When you are at home.
     - Which appliances you use.
     - When you eat.
     - Whether you arrive late to work.



University of New Brunswick
Faculty of Computer Science
CS4413/6413: Foundations of Privacy
Theory Homework Assignment 1, ***Due Time, Date*** *5:00 PM, February 28, 2019*

Student Name: ____________________________ Matriculation Number: ________________

Instructor: Rongxing Lu
The marking scheme is shown in the left margin and [100] constitutes full marks.

[**30**] 1. Please answer the following questions.

[5] (a) Why data privacy matters to us? Please elaborate your view as detailed as possible in terms of the General Data Protection Regulation (GDPR).

• Answer:

- We care - we are responsible for handling people's most personal information.
- This is an opportunity to make privacy central to what we do.
- By not handling personal data properly we could put individuals at risk and the entitys reputation at stake.
- Getting it wrong could result in significant fines.
- We need robust systems and processes in place to make sure we use personal information properly and comply.

[5] (b) What are k-anonymity, l-diversity, and t-closeness in database privacy?

• Answer:

- k-anonymity: Table T satisfies k-anonymity with regards to quasi-identifier QI if each tuple in (the multiset) $T[QI]$ appears at least k times.
- l-diversity: a table T satisfies l-diversity with regards to quasi-identifier QI if each of its $T[QI]$ group contains at least l well-represented values for the sensitive attributes.
- t-closeness: a table T has t-closeness with regards to quasi-identifier QI if the distance between the distribution of sensitive attribute values in each of its $T[QI]$ group is no more than threshold t.

[5] (c) What is differential privacy technique? Please describe the steps on how to add the proper Laplace noise to obtain the desirable privacy for the released dataset.

• Answer: The differential privacy technique can guarantee that the privacy risk should not substantially increase as a result of adding or deleting operations in a statistical database. For example, there are two datasets X and X', and X is a neighbor of X' because they differ in one row. However, from the released statistics, it is hard to distinguish X and X'.

Adding Laplace noise follows the following steps.

- According to the statistics function $F(D)$, compute sensitivity of function $F(D)$, i.e., SF.
- Choose the privacy level of the database ε and set the parameter of Laplace distribution λ to be $\frac{SF}{\varepsilon}$.

- When computing the statistics function $F(D)$, add a random noise x to $F(D)$, i.e., $F(D)+x$, where x is from $Lap(\lambda)$ distribution.

[5] (d) Describe the Big Data 4V's characteristics, including volumn, velocity, variety, and veracity, as detailed as possible.

- Answer:

  - Volume: Data volume is increasing exponentially, e.g., 44x increase from 2009 to 2020 and from 0.8 zettabytes to 35zb.
  - Velocity: Data are generated fast and need to be processed fast.
  - Variety: Different types of data are involved, including relational data, text data, graph data, etc., which become more complex. All these types of data need to be linked together.
  - Veracity: The data inconsistency and incompleteness, ambiguities etc. will bring some uncertainty in big data. Veracity will consider some security issues in big data in order to protect data veracity.

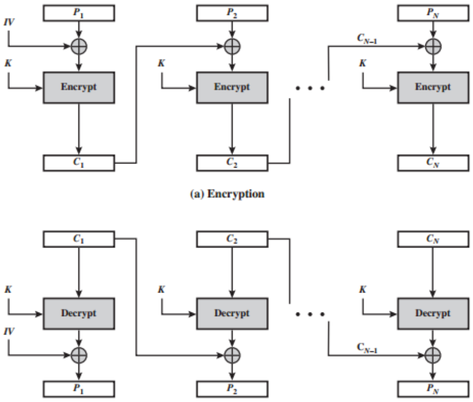
[5] (e) Describe the birthday attack in hash function.

- Answer: Birthday attack is to find two people with the same birthday. For a hash function H, the birthday attack is to find a collision, i.e., find two numbers x and y such that $H(x) = H(y)$.

[5] (f) Describe the homomorphic encryption technique as detailed as possible.

- Answer: Homomorphic encryption is a form of encryption that allows computation on ciphertexts, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the plaintexts. For example, if an encryption technique $E(\cdot)$ satisfies homomorphic property, $E(x) \circ E(y) = E(x \odot y)$, where "$\circ$" is the operation over ciphertext data and "$\odot$" is the operation over plaintext data.

[**6**] 2. Suppose that a message has been encrypted using DES in ciphertext block chaining mode. One bit of ciphertext in block C_i, and another bit of ciphertext in block C_{i+1} are accidentally transformed from 0 to 1 during transmission. How much plaintext will be garbled as a result?



(a) Encryption

(b) Decryption

- Answer: As shown in the above figure, the decryption algorithm in CBC model is $P_i = Dec(C_i) \oplus C_{i-1}$. Thus, if C_i and C_{i+1} are changed during transformation, the plaintexts $P_i = Dec(C_i) \oplus C_{i-1}$,

$P_{i+1} = Dec(C_{i+1}) \oplus C_i$ and $P_{i+2} = Dec(C_{i+2}) \oplus C_{i+1}$ will be affected and other plaintexts will not be affected. Therefore, three plaintexts, i.e., P_i, P_{i+1} and P_{i+2} will be garbled.

[**6**] 3. Using the extended Euclidean algorithm to find the multiplicative inverse of

[3] (a) $12345 \bmod 54321$

• Answer: According to the extended Euclidean algorithm,

$$\begin{aligned} 54321 &= 12345 * 4 + 4941 && (1) \\ 12345 &= 4941 * 2 + 2463 && (2) \\ 4941 &= 2463 * 2 + 15 && (3) \\ 2463 &= 15 * 164 + 3 && (4) \end{aligned}$$

Thus, $gcd(12345, 54321) = 3$, so the inverse of $12345 \bmod 54321$ does not exist.

[3] (b) $350 \bmod 1769$

• Answer: According to the extended Euclidean algorithm,

$$\begin{aligned} 1769 &= 350 * 5 + 19 && (5) \\ 350 &= 19 * 18 + 8 && (6) \\ 19 &= 8 * 2 + 3 && (7) \\ 8 &= 3 * 2 + 2 && (8) \\ 3 &= 2 * 1 + 1 && (9) \end{aligned}$$

Thus,

$$\begin{aligned} 1 &= 1 - 2 * 1 && (10) \\ &= 3 - (8 - 3 * 2) * 1 = 3 * 3 - 8 * 1 && (11) \\ &= (19 - 8 * 2) * 3 - 8 = 19 * 3 - 8 * 7 && (12) \\ &= 19 * 3 - (350 - 19 * 18) * 7 = 19 * 129 - 350 * 7 && (13) \\ &= (1769 - 350 * 5) * 129 - 350 * 7 = 1769 * 129 - 350 * 652 && (14) \end{aligned}$$

Therefore, $350^{-1} \bmod 1769 = -652 \bmod 1769 = 1117 \bmod 1769$.

[**6**] 4. In a public-key system using RSA , you intercept the ciphertext $C = 9$ sent to a user whose public key is $e = 5, n = 35$. What is the plaintext M?

• Answer: According to $n = pq = 35$, $p = 5$ and $q = 7$. So, $\varphi(n) = (p-1)(q-1) = 24$. At the same time, $e = 5$ and $ed \equiv 1 \bmod \varphi(n)$, so $d = 5$.

According to the decryption algorithm $M = C^d \mod n$, so $M = C^d = 9^5 \mod 35 = 4$.

[**6**] 5. Use the Chinese Remainder Theorem (CRT) to solve x, where

$$\begin{cases} x & \equiv & 1 \bmod 3 \\ x & \equiv & 2 \bmod 5 \\ x & \equiv & 3 \bmod 7 \end{cases}$$

• Answer: Let $m_1 = 3, m_2 = 5, m_3 = 7$, $a_1 = 1, a_2 = 2, a_3 = 3$. We have $M = m_1 \cdot m_2 \cdot m_3 = 3 \times 5 \times 7 = 105$.

Then, $M_1 = M/m_1 = 35, M_2 = M/m_2 = 21, M_3 = M/m_3 = 15$.

$\alpha_1 = M_1^{-1} \mod m_1 = 35^{-1} \mod 3 = 2, \alpha_2 = M_2^{-1} \mod m_2 = 21^{-1} \mod 5 = 1, \alpha_3 = M_3^{-1} \mod m_3 = 15^{-1} \mod 7 = 1$.

Therefore, $x = (a_1 \cdot \alpha_1 \cdot M_1 + a_2 \cdot \alpha_2 \cdot M_2 + a_3 \cdot \alpha_3 \cdot M_3) \mod M = (1 \times 2 \times 35 + 2 \times 1 \times 21 + 3 \times 1 \times 15) \mod 105 = 52$

[**6**] 6. Consider an Elgamal encryption scheme with a common prime $q = 11$ and a primitive root $\alpha = 2$. If B has public key $Y_B = 7$ and A choose the random integer $k = 3$, what is the ciphertext of $M = 9$?

• Answer: The ciphertext $C_1 = \alpha^k \mod q = 2^3 \mod 11 = 8$ and $C_2 = M * Y_B^k \mod q = 9 * 7^3 \mod 11 = 7$.

[**10**] 7. Let $F(x)$ be the true answer on input x, and $Geom(\alpha)$ be the noise sampled from Geometric distribution with parameter $\alpha = e^{-\epsilon/S(F)}$. Please prove that the release of $F(x) + Geom(\alpha) = F(x) + Geom(e^{-\epsilon/S(F)})$ can obtain ϵ-DP guarantee.

Proof: Suppose that $A = F(x) + Geom(\alpha)$, and D_1 and D_2 are any two adjacent DBs. Thus, $A(D_1) = F(D_1) + x_1$ and $A(D_2) = F(D_2) + x_2$, where x_1 and x_2 are $Geom(\alpha)$ distributed. Since $\alpha = e^{-\epsilon/S(F)}$, the probability density for x_1 is proportional to $\frac{\alpha-1}{\alpha+1}\alpha^{|x_1|}$. Similarly, the probability density for x_2 is proportional to $\frac{\alpha-1}{\alpha+1}\alpha^{|x_2|}$. Therefore, for any $T \in range(A)$

$$\frac{Pr[A(D_1) = T]}{Pr[A(D_2) = T]} = \frac{Pr[F(D_1) + x_1 = T]}{Pr[F(D_2) + x_2 = T]} = \frac{Pr[x_1 = T - F(D_1)]}{Pr[x_2 = T - F(D_2)]} = \frac{\frac{\alpha-1}{\alpha+1}\alpha^{|x_1|}}{\frac{\alpha-1}{\alpha+1}\alpha^{|x_2|}}$$

$$= \frac{\alpha^{|x_1|}}{\alpha^{|x_2|}} = \alpha^{|x_1|-|x_2|} = \alpha^{|T-F(D_1)|-|T-F(D_2)|} \le \alpha^{|F(D_1)-F(D_2)|} = (e^{-\epsilon/S(F)})^{|F(D_1)-F(D_2)|}$$

where the inequality follows form the triangle inequality. By the definition of sensitivity,

$$S(F) = Max_{D_1, D_2 : |D_1 - D_2| = 1} |F(D_1) - F(D_2)|$$

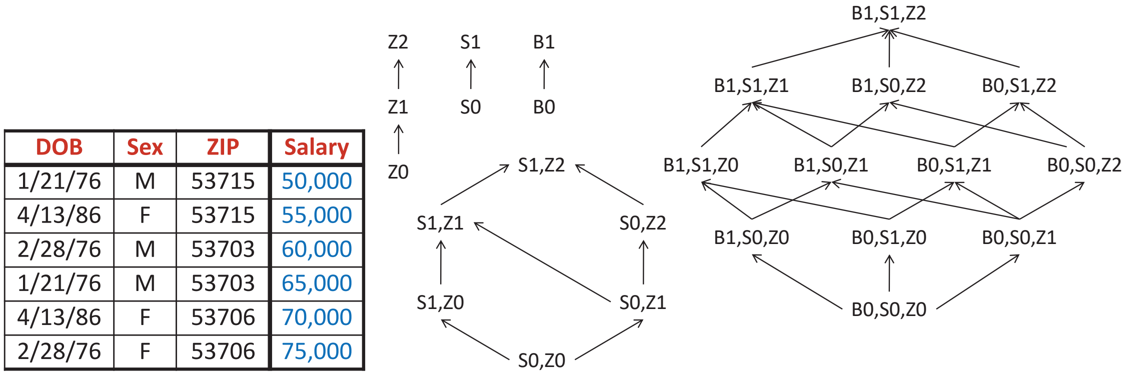
.

Thus, $(e^{-\epsilon/S(F)})^{|F(D_1)-F(D_2)|} \le e^{-\epsilon}$. The ration is bounded by $e^{-\epsilon}$, yields ϵ-Differential Privacy.

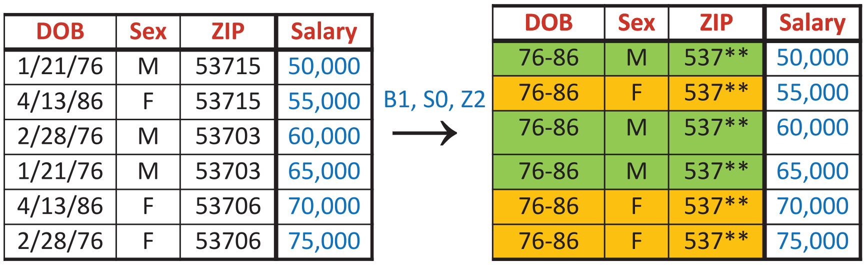
[**10**] 8. Incognito is one of approaches to implement the k-anonymity. Given a table below, the full-domain generalizations described by "domain vectors" are represented as follows.

- $B_0 = \{1/21/76, 2/28/76, 4/13/86\} \rightarrow B_1 = \{76 - 86\}$
- $S_0 = \{M, F\} \rightarrow S_1 = \{*\}$
- $Z_0 = \{53715, 53710, 3706, 53703\} \rightarrow Z_1 = \{5371*, 5370*\} \rightarrow Z_2 = \{537**\}$

Based on the Lattice of Domain Vectors, we can achieve the k-anonymity for the original table. For example, with (B_1, S_0, Z_2) generalization, we can achieve $k = 3$-anonymity.



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

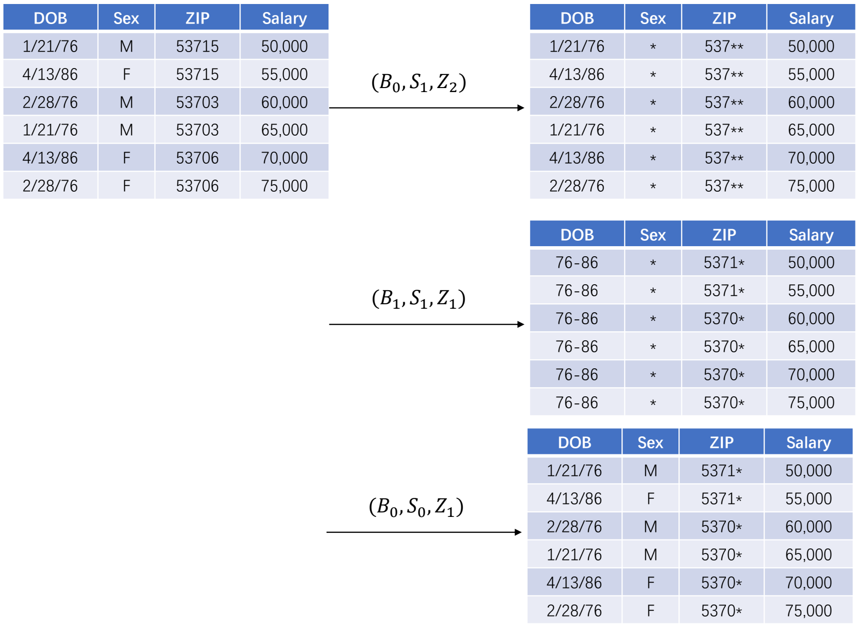


DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

DOB	Sex	ZIP	Salary
76-86	M	537**	50,000
76-86	F	537**	55,000
76-86	M	537**	60,000
76-86	M	537**	65,000
76-86	F	537**	70,000
76-86	F	537**	75,000

Please follow the example to apply (B_0, S_1, Z_2), (B_1, S_1, Z_1), (B_0, S_0, Z_1) to generalize the original table, and discuss what is the value of k in each generalized case?

- Answer: The generalized tables are shown as the following figure.



DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

$\xrightarrow{(B_0, S_1, Z_2)}$

DOB	Sex	ZIP	Salary
1/21/76	*	537**	50,000
4/13/86	*	537**	55,000
2/28/76	*	537**	60,000
1/21/76	*	537**	65,000
4/13/86	*	537**	70,000
2/28/76	*	537**	75,000

$\xrightarrow{(B_1, S_1, Z_1)}$

DOB	Sex	ZIP	Salary
76-86	*	5371*	50,000
76-86	*	5371*	55,000
76-86	*	5370*	60,000
76-86	*	5370*	65,000
76-86	*	5370*	70,000
76-86	*	5370*	75,000

$\xrightarrow{(B_0, S_0, Z_1)}$

DOB	Sex	ZIP	Salary
1/21/76	M	5371*	50,000
4/13/86	F	5371*	55,000
2/28/76	M	5370*	60,000
1/21/76	M	5370*	65,000
4/13/86	F	5370*	70,000
2/28/76	F	5370*	75,000

From the figure, we know that (1) when applying (B_0, S_1, Z_2) to the original table, $k = 2$; (2) when applying (B_1, S_1, Z_1) to the original table, $k = 2$; (3) when applying (B_0, S_0, Z_1) to the original table, $k = 1$.

[**10**] 9. Alice and Bob are good friends, they have shared a secret key sk in advance. Now, Alice wants to send 20 messages $x_1, x_2, \cdots, x_{20}$ to Bob, because there may be errors occurring in communication channel and also possible injection false data attack from external attackers, Alice hopes to use the bloom filter to enhance the security of these messages in term of source authentication and data integrity. Can you help Alice and Bob to design an efficient bloom filter?

[5] (a) For all hash functions $h_1, h_2, \cdots, h_k$ used in the bloom filter, we define each hash function h_i : $\{0,1\}^* \rightarrow \{2^0, 2^1, 2^2, \cdots, 2^{159}\}$, which means we are using an array D with length $n = 160$ as a bloom filter. Then, with the bloom filter D, Alice just needs to use 160 bits overheads for authenticating 20 messages. Can you compute the optimal number of hash functions, k , and the corresponding false positive FP?

• Answer: The optimal number of hash function $k = \ln 2 * \frac{n}{m} = \ln 2 * \frac{160}{20} \approx 6$ and the false positive is $(1 - e^{-\frac{km}{n}})^k = (1 - e^{-\frac{6*20}{160}})^6 = 0.0216$.

[5] (b) For the <u>same values</u> of k, n in Question 9(a), we now design two bloom filters $(D1, D2)$ as the authenticator as follow: the authenticate overhead is still $n = 160$ bits in total, but each hash function is $h_i : 0, 1^* \rightarrow \{2^0, 2^1, 2^2, \cdots, 2^{79}\}$, which means the bloom filters $D1$ and $D2$ are with length $n/2 = 80$ bits. Among k hash functions, half of them, i.e., $k/2$ hash functions, are used for $D1$, the rest $k/2$ hash functions are used for $D2$. With these settings, can you compute the new false positive FP? Compared with the bloom filter D in Question 9(a), which one is better, $(D1, D2)$ or D?

• Answer: For the bloom filter D_1, $n_1 = 80$, $k_1 = \frac{k}{2} = 3$ and $m = 20$.

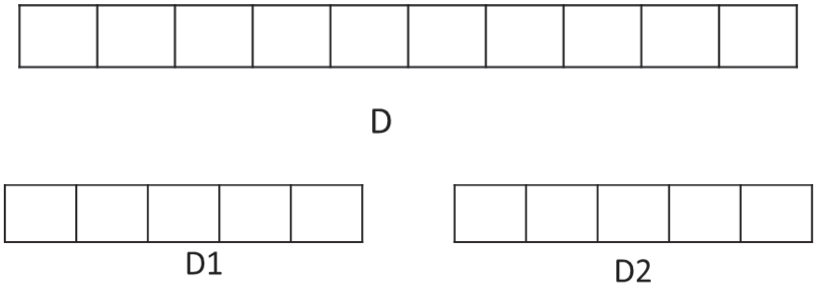
Thus, the false positive in bloom filter D_1 is $p_1 = (1 - e^{-\frac{k_1 m}{n_1}})^{k}_{1} = (1 - e^{-\frac{3*20}{80}})^3 = 0.1489$.

Similarly, for the bloom filter D_2, $n_2 = 80$, $k_2 = \frac{k}{2} = 3$ and $m = 20$.

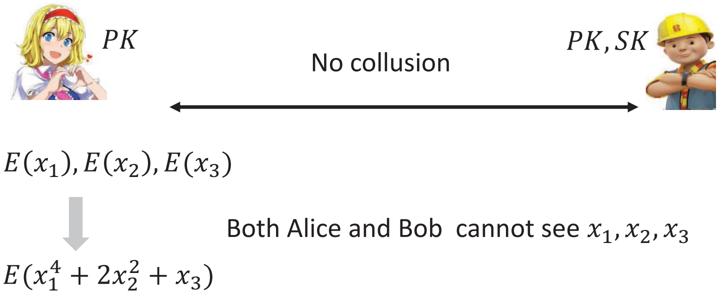
Thus, the false positive in bloom filter D_2 is $p_2 = p_1 = 0.1489$.

So the overall false positive probability $p = p_1 * p_2 = 0.1489 * 0.1489 = 0.0216$.

Since the false positive probability in two bloom filters (D_1, D_2) is the same as that in bloom filter D, the bloom filter D is the same as (D_1, D_2).



[**10**] 10. Let $E()$ be a BGN homomorphic encryption scheme. Assume Bob has the public/private key pair (pk, sk) of $E()$, and Alice only has the public key pk of $E()$. Consider there is no collusion between Alice and Bob. When Alice has three ciphertexts $E(x_1)$, $E(x_2)$ and $E(x_3)$, please design a protocol run between Alice and Bob. With the protocol, Alice can finally obtain the ciphertext $E(x_1^4 + 2x_2^2 + x_3)$ while both Alice and Bob have no idea on the plaintexts x_1, x_2, and x_3.


PK
No collusion
PK, SK
$E(x_1), E(x_2), E(x_3)$
Both Alice and Bob cannot see x_1, x_2, x_3
$E(x_1^4 + 2x_2^2 + x_3)$

• Answer:

Solution 1:

- Compute $E'(x_1^4)$:
  - Alice chooses a random number r_1 and computes $E(x_1 + r_1) = E(x_1) * E(r_1)$. Then, Alice sends $E(x_1 + r_1)$ to Bob.
  - Bob recovers $(x_1 + r_1)$ with the private key sk, and then computes $E((x_1 + r_1)^2) = E(x_1 + r_1)^{x_1 + r_1}$. Then, Bob sends $E((x_1 + r_1)^2)$ to Alice.
  - After receiving $E((x_1 + r_1)^2)$, Alice computes $E(x_1^2) = E((x_1 + r_1)^2 - 2x_1r_1 - r_1^2) = E((x_1 + r_1)^2) * [E(x_1)^{2r_1}]^{-1} * E(r_1^2)^{-1}$.
  - Alice computes $E'(x_1^4) = e(E(x_1^2), E(x_1^2))$.
- Compute $E'(x_2^2)$: Alice first computes $E(2x_2) = E(x_2)^2$ and then $E'(2x_2^2) = e(E(2x_2), E(x_2))$.
- Compute $E'(x_3)$: Alice computes $E'(x_3) = e(E(x_3), E(1))$.
- Compute $E'(x_1^4 + 2x_2^2 + x_3) = E'(x_1^4) * E'(2x_2^2) * E'(x_3)$.

Solution 2:

- Compute $E'(x_1^4)$:
  - Alice chooses a random number r_1 and computes $E(x_1 + r_1) = E(x_1) * E(r_1)$. Then, Alice sends $E(x_1 + r_1)$ to Bob.
  - Bob recovers $(x_1 + r_1)$ with the private key sk, and then computes $E((x_1 + r_1)^2) = E(x_1 + r_1)^{x_1 + r_1}$. Then, Bob sends $E((x_1 + r_1)^2)$ to Alice.
  - After receiving $E((x_1 + r_1)^2)$, Alice computes $E(x_1^2) = E((x_1 + r_1)^2 - 2x_1r_1 - r_1^2) = E((x_1 + r_1)^2) * [E(x_1)^{2r_1}]^{-1} * E(r_1^2)^{-1}$.
  - Similarly, Alice computes $E(x_1^4)$ by $E(x_1^2)$ with the help of Bob.
- Alice computes $E(x_2^2)$ using the same algorithm that computes $E(x_1^2)$.
- Alice computes $E(x_1^4 + 2x_2^2 + x_3) = E(x_1^4) * E(x_2^2)^2 * E(x_3)$.

University of New Brunswick
Faculty of Computer Science
CS4413/6413: Foundations of Privacy
Theory Homework Assignment 2, ***Due Time, Date*** *5:00 PM, March 28, 2019*

Student Name: ____________________________ Matriculation Number: ________________

Instructor: Rongxing Lu
The marking scheme is shown in the left margin and [100] constitutes full marks.

[**20**] 1. Please prove the following results.

[10] (a) Let p be a prime number, and $a^p \equiv b^p \bmod p$, prove $a^p \equiv b^p \bmod p^2$.

[10] (b) Let $\gcd(m,n)=1$, prove $m^{\phi(n)} + n^{\phi(m)} \equiv 1 \bmod mn$.

Answer:

(a) $a^p - b^p = (a-b)(a^{p-1} + a^{p-2}b + \cdots + b^{p-1}) = (a-b)\Sigma_{i=0}^{p-1} a^{p-1-i}b^i$.

Since a and b are prime numbers, according to the Femat's Little Theorem,

$a^{p-1} \equiv 1 \mod p$ and $b^{p-1} \equiv 1 \mod p$.

So $a^p \equiv b^p \mod p \Rightarrow a^{p-1} \cdot a \equiv b^{p-1} \cdot b \mod p \Rightarrow a \equiv b \mod p$.

$a - b \equiv 0 \mod p \Rightarrow p|(a-b)$.

In addition, for any $i \geq 0$, $a^i \equiv b^i \mod p$, so

$\Sigma_{i=0}^{p-1} a^{p-1-i}b^i \equiv \Sigma_{i=0}^{p-1} a^{p-1-i}a^i \equiv \Sigma_{i=0}^{p-1} a^{p-1} \equiv \Sigma_{i=0}^{p-1} 1 \equiv p \equiv 0 \mod p \Rightarrow p|\Sigma_{i=0}^{p-1} a^{p-1-i}b^i$.

Then, $p|(a-b)\Sigma_{i=0}^{p-1} a^{p-1-i}b^i \Rightarrow p|(a^p - b^p) \Rightarrow a^p \equiv b^p \mod p$.

(b) According to the Euler's theorem,

$$gcd(m,n) = 1 \Rightarrow \begin{cases} m^{\phi(n)} + n^{\phi(m)} & \equiv & 1 \bmod n \\ m^{\phi(n)} + n^{\phi(m)} & \equiv & 1 \bmod m \end{cases}$$

According to the Chinese Remainder Theorem,

$a_1 = 1, a_2 = 1, m_1 = n, m_2 = m \Rightarrow M = mn, M_1 = m, M_2 = n$

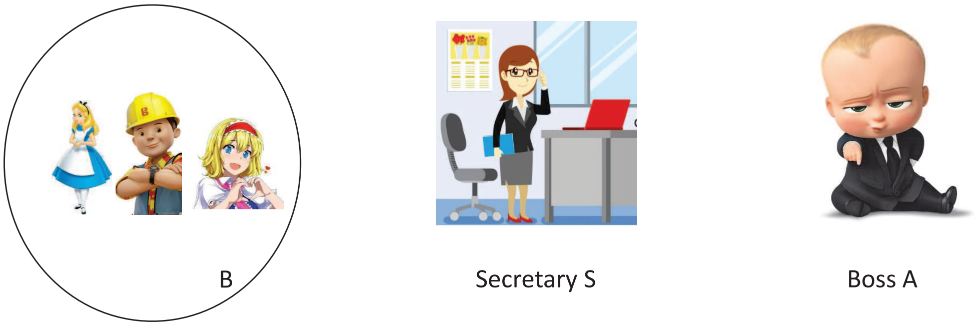
$gcd(m,n) = 1 \Rightarrow \exists a, b$ such that $am + bn = 1$

$M_1^{-1} \bmod m_1 \Rightarrow m^{-1} \bmod n \equiv a$ and $M_2^{-1} \bmod m_2 \Rightarrow n^{-1} \bmod m \equiv b$.

$m^{\phi(n)} + n^{\phi(m)} \equiv (a_1 \cdot (M_1^{-1} \bmod m_1) \cdot M_1 + a_2 \cdot (M_2^{-1} \bmod m_2) \cdot M_2) \bmod M$

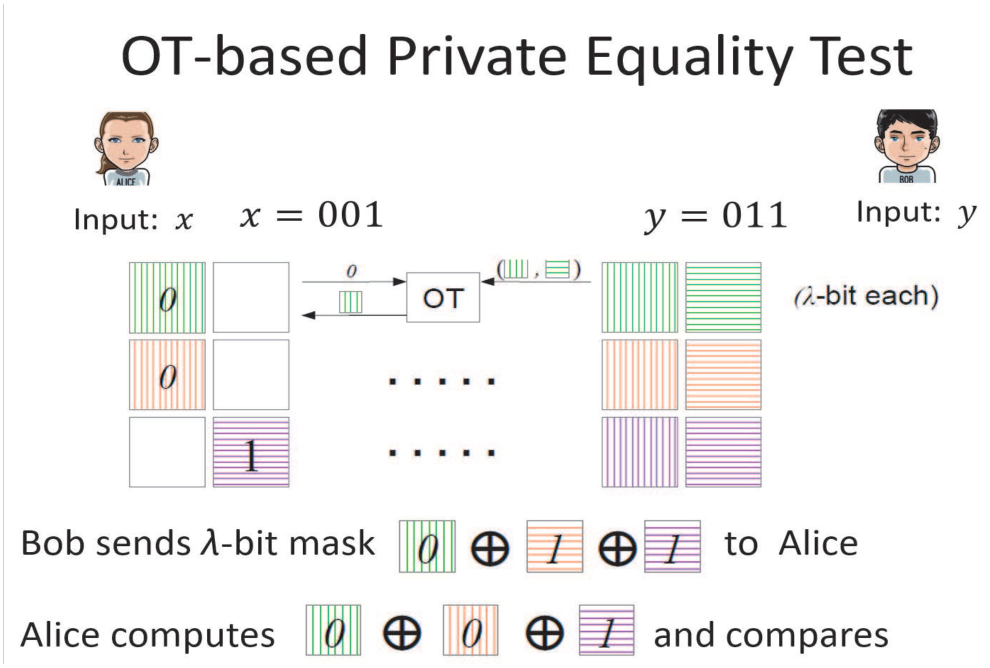
Thus, $m^{\phi(n)} + n^{\phi(m)} \equiv am + bn \equiv 1 \mod mn$.

[**30**] 2. Boss A has a list of keywords $K = \{k_1, k_2, \cdots, k_n\}$ and a set of friends $B = \{B_1, B_2, \cdots, B_n\}$. Boss A asks his secretary S to only forward messages (that include at least one keyword in K and the sender belongs to B) to him. The conditions are i) the secretary S cannot know K; ii) the secretary cannot know B and the message content. (Hint: you can apply the symmetric key encryption and the bloom filter techniques.)



Answer: The set of friends generates and shares a security key SK, and sends SK to Boss A. After receiving SK, Boss A encrypts $\{k_1, k_2, \cdots, k_l\}$ as $\{H(k_1||SK), \cdots, H(k_n||SK)\}$. Then, he/she generates a Bloom Filter array and stores encrypted keywords $\{H(k_1||SK), \cdots, H(k_n||SK)\}$ to the Bloom Filer array. Boss A also sends the Bloom filter array to his secretary S. When B_j sends a message to Boss A, he/she will encrypt each keyword k_j in the message as $H(k_j||SK)$, and then send these encrypted keywords to the secretary S. After receiving the message, S uses the Bloom Filter array to check whether there is a keywords belonging to the keywords set K. If yes, forward to the Boss A. Otherwise, drop out the message.

[**25**] 3. Assume Alice, Bob and Carlo respectively have the data set $A = \{\cdots\}$, $B = \{\cdots\}$, $C = \{\cdots\}$. How to use the OT-protocol to design a Private Set Intersection protocol among Alice, Bob, and Carlo, so that each one can obtain $A \cap B \cap C$. You can design your solution based on the OT-based PET (Private Equality Test) protocol in the figure.



OT-based Private Equality Test

Input: x $x = 001$

$y = 011$ Input: y

(λ-bit each)

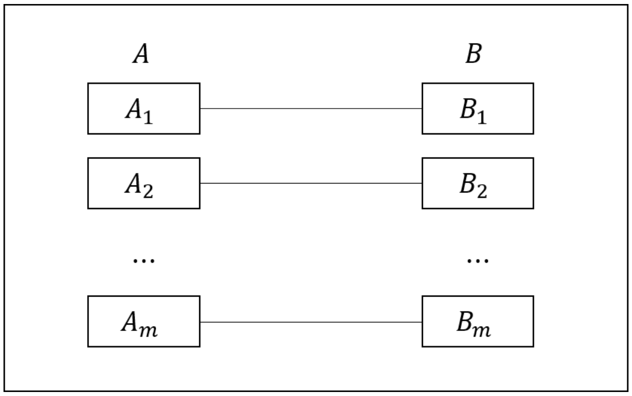
Bob sends λ-bit mask $0 \oplus 1 \oplus 1$ to Alice

Alice computes $0 \oplus 0 \oplus 1$ and compares

Answer: The intersection of $A \cap B$ can be computed as follows.

- As shown in the following figure, put the elements of A and B into m hash tables $\{A_1, A_2, \cdots, A_m\}$ and $\{B_1, B_2, \cdots, B_m\}$, respectively.
- Compute $A_i \cap B_i$ by using the OT-based private equality test protocol to compare each element of A_i with each element of B_i for $i = 1, 2, \cdots, m$.
- Then, $A \cap B = (A_1 \cap B_1) \cup (A_2 \cap B_2) \cup (A_m \cap B_m)$.

Finally, $A \cap B \cap C$ is the intersection of $A \cap B$ and C, and can be computed with the same method.

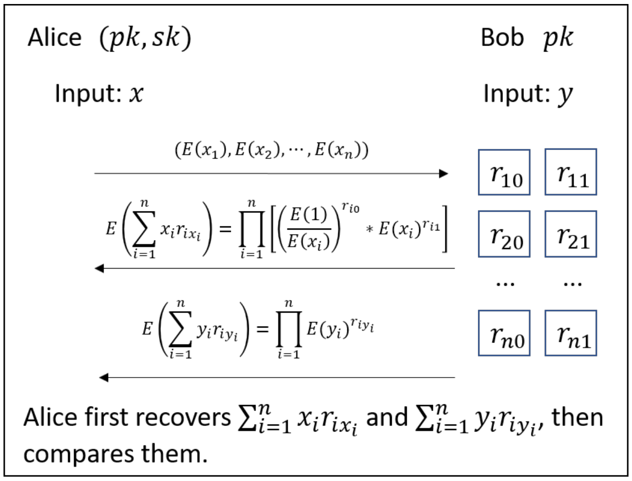


[25] 4. How to use the OT-protocol to design a privacy-preserving integer comparison protocol between two parties, e.g., two integers x, y, both of them are n bits, where $x > y$, $x < y$, or $x = y$. You can design your solution based on the above OT-based PET protocol. (Hint: you may disclose two bits information in your solution!)

Answer: Suppose that Alice has x and Bob has y, and x and y can be denoted as $x = x_1 x_2 \cdots x_n$ and $y = y_1 y_2 \cdots y_n$, respectively. As shown in the following figure, Alice and Bob can use the Paillier homomorphic encryption technique and OT protocol to compare x and y. In specific, Alice has public key pk and private key sk, while Bob only has public key pk. They can compare x and y as follows.

- Alice encrypts x as $E(x) = (E(x_1), E(x_2), \cdots, E(x_n))$. Then, he/she sends $E(x)$ to Bob.
- On receiving $E(x)$, Bob selects $2n$ random positive numbers $\{r_{i0}, r_{i1} | i = 1, 2, \cdots, n\}$ such that $r_{i1} > r_{i0} + \sum_{j=i+1}^{n} r_{j1}$. Then, Bob computes $\prod_{i=1}^{n}[(\frac{E(1)}{E(x_i)})^{r_{i0}} * E(x_i)^{r_{i1}}]$, i.e., $E(\sum_{i=1}^{n} x_i r_{ix_i})$, and returns it to Alice. At the same time, Bob computes $\prod_{i=1}^{n} E(y_i)^{r_{iy_i}}$, i.e., $E(\sum_{i=1}^{n} y_i r_{iy_i})$, and sends it to Alice.
- Alice recovers $\sum_{i=1}^{n} x_i r_{ix_i}$ and $\sum_{i=1}^{n} y_i r_{iy_i}$ from $E(\sum_{i=1}^{n} x_i r_{ix_i})$ and $E(\sum_{i=1}^{n} y_i r_{iy_i})$, respectively. Then, he/she compares them to obtain the comparison result of x and y.

Since the positive numbers satisfy that $r_{i1} > r_{i0} + \sum_{j=i+1}^{n} r_{j1}$ for $i = 1, 2, \cdots, n$, the comparison result of $\sum_{i=1}^{n} x_i r_{ix_i}$ and $\sum_{i=1}^{n} y_i r_{iy_i}$ is equal to that of x and y.


Alice (pk, sk)
Bob pk
Input: x
Input: y
$(E(x_1), E(x_2), \cdots, E(x_n))$
r_{10}
r_{11}
r_{20}
r_{21}
$\cdots$
$\cdots$
r_{n0}
r_{n1}
$E\left(\sum_{i=1}^{n} x_i r_{ix_i}\right) = \prod_{i=1}^{n}\left[\left(\frac{E(1)}{E(x_i)}\right)^{r_{i0}} * E(x_i)^{r_{i1}}\right]$
$E\left(\sum_{i=1}^{n} y_i r_{iy_i}\right) = \prod_{i=1}^{n} E(y_i)^{r_{iy_i}}$
Alice first recovers $\sum_{i=1}^{n} x_i r_{ix_i}$ and $\sum_{i=1}^{n} y_i r_{iy_i}$, then compares them.