

CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 0: Course Overview

Lecturer: Rongxing LU

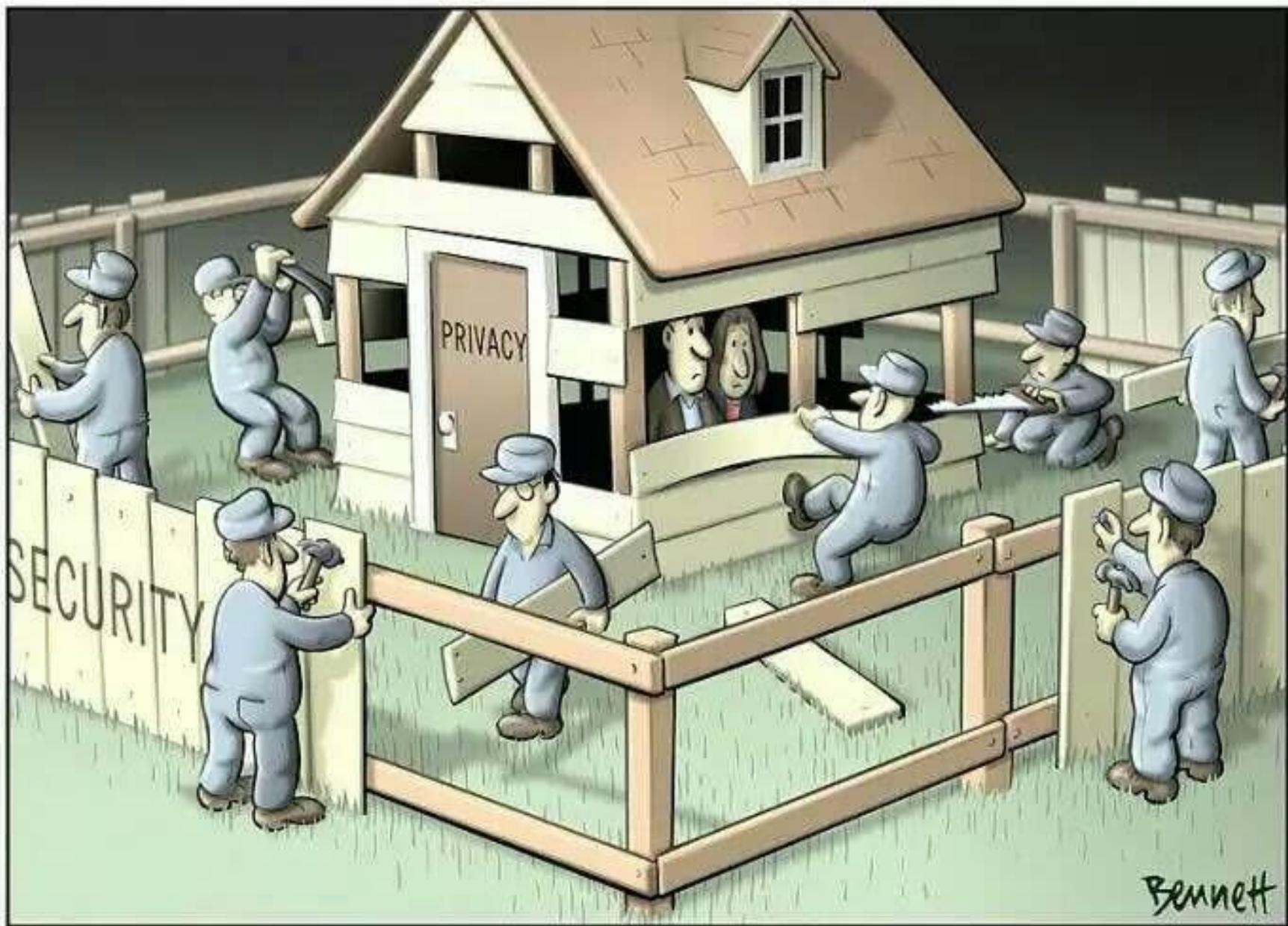
Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

Personal Information is Everywhere





Questions will be Discussed in This Course

- What is Privacy?
- How can we protect our privacy?



Course Instructor

- Instructor: Rongxing Lu
 - Office: GE 114
 - Email: rlu1@unb.ca
 - Office Hours:
 - T 2:30PM-3:30PM
 - Extra office hours on demand
 - Website: <http://www.cs.unb.ca/~rlu1/>

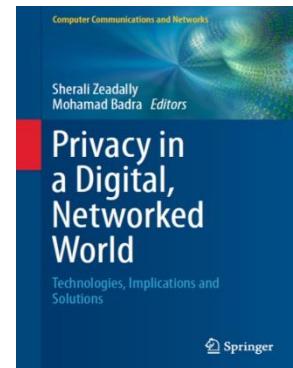
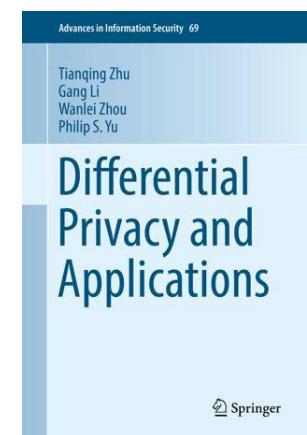


Logistics

- Lectures:
- Tutorials:
- Course work and grading:
 - Homework (30%) – 2 x 15%
 - Two theory assignments – 10% + 5%
 - Two programming assignments – 10% + 5%
 - Due of the First Theory & Programming Assignments
 - **February 28, 2019 (D2L Dropbox)**
 - Due of the Second Theory & Programming Assignments
 - **March 28, 2019 (D2L Dropbox)**
 - Course project (20%)
 - Midterm Exam (20%)
 - The test will be held during the middle of the term based on materials covered till that time.
 - Final Exam (30%)
 - Based on most of tutorial questions discussed in this course.

Logistics: Textbook

- Lecture notes prepared by the Instructor
- Textbook:
 - Tianqing Zhu, Gang Li, Wanlei Zhou, Philip S. Yu, **Differential Privacy and Applications**, Springer; 1st ed. 2017 edition, ISBN-10: 9783319620022, ISBN-13: 978-3319620022
- Recommended References:
 - S. Zeadally and M. Badra, **Privacy in a Digital, Networked World: Technologies, Implications and Solutions**, Springer Publishing Company, 2015, ISBN:3319084690 9783319084695
 - W. Mao, Modern Cryptography: Theory and Practice. Prentice Hall PTR, 2003, ISBN: 0130669431
 - D. Stinson, Cryptography: Theory and Practice (Third Edition). CRC Press, 2005, 978-1584885085
 - Papers of interest from selected conferences and journals by the Instructor



Logistics: Course Projects

- Project Proposal
 - Group based (no more than 3 each group), Based on the paper selected
- Deliverable Part I: in-class presentation
 - 25-minute presentation 5%
 - **Presentation Arrangement: March 18, 2019 – April 5, 2019**
- Deliverable Part II: written report
 - CS4413: 3-page reports 15%
 - CS6413: 6-page reports 15%
 - **Due: April 8, 2019 (D2L Dropbox)**

Logistics: Collaboration Policy

- You are allowed to discuss homework problems and approaches for their solution with other students in the class, but are required to figure out and write out detailed solutions independently and to acknowledge any collaboration or other source.

Outline of Topics in the Course

Topic #	Topics
Topic 0	<ul style="list-style-type: none">• Course Overview
Topic 1	<ul style="list-style-type: none">• Database privacy
Topic 2	<ul style="list-style-type: none">• Big data privacy
Topic 3	<ul style="list-style-type: none">• Review some basic cryptographic techniques for privacy
Topic 4	<ul style="list-style-type: none">• Homomorphic cryptographic techniques for privacy
Topic 5	<ul style="list-style-type: none">• Anonymous communication network techniques
Topic 6	<ul style="list-style-type: none">• Private information retrieval techniques
Topic 7	<ul style="list-style-type: none">• Oblivious transfer protocols
Topic 8	<ul style="list-style-type: none">• Zero knowledge proof techniques
Topic 9	<ul style="list-style-type: none">• Private matching protocols in mobile social networks
Topic 10	<ul style="list-style-type: none">• Secure data sharing in cloud computing
Topic 11	<ul style="list-style-type: none">• Privacy-preserving data aggregation in smart grid

Student Introductions

- Who are you?
- Why are you here?



What is personal data?



- 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 eg. health data, religious beliefs

Why data privacy matters to us?

- 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 entity's 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

Overview of General Data Protection Regulation (GDPR)

- It has come in to force on 25 May 2018
- **What?**
 - a European law that will replace the current Data Protection Act.
 - The UK government will still implement the rules after Brexit.
- **Why?**
 - The aim is to strengthen and unify personal data protection for all individuals living in the European Union.
- **Who?**
 - The Information Commissioner's Office (ICO) will lead on GDPR in the UK and will hand out penalties for organisations who are in breach of the new law.



What's changing In GDPR?

- Many GDPR principles are similar to those in current the Data Protection Act.
- There are also new and strengthened requirements for how we protect people's data.
- Changes include:
 - new rights (e.g. 'right to be forgotten')
 - greater emphasis on transparency and record-keeping
 - mandatory data breach reporting
 - much larger fines for when organisations get things wrong



What is data privacy all about?

- Being open with people about how we use their information
- Not keeping their information longer than necessary
- Making sure it is accurate
- Making sure that it is safe
- Knowing what information we've got and what we can do with it (eg. sharing)
- Recognising a breach and knowing what to do



Who does this affect?

- All of us - we all have a responsibility to keep people's information safe.
- Particularly those involved in:
 - Research involving personal data and/or human participants
 - Finance
 - IT



Privacy vs Security



- Isn't it the same thing?
 - Not really. But they are kissing cousins.
- **Data privacy** is focused on the use and governance of personal data—things like putting policies in place to ensure that consumers' personal information is being collected, shared and used in appropriate ways.
- **Security** focuses more on protecting data from malicious attacks and the exploitation of stolen data for profit.
- While security is necessary for protecting data, it's not sufficient for addressing privacy.

Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 1: Database Privacy

Lecturer: Rongxing LU

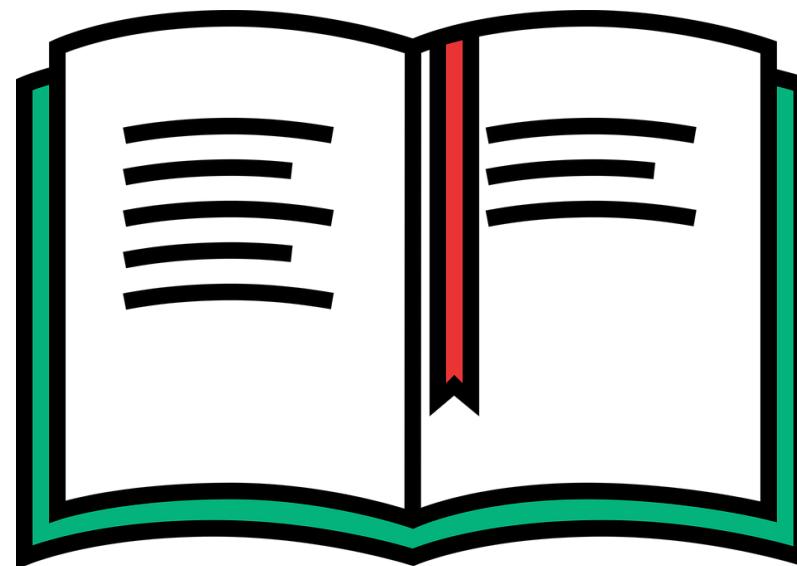
Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

Outline

- Database Security and Privacy
- Database Anonymity Techniques
- Differential Privacy Techniques



Database Security and Privacy



Database Overview

- Every university (e.g., UNB), hospital, company need places to store institutional knowledge and data.
- Frequently that data contains proprietary information
 - Personally Identifiable Data
 - Employee HR Data
 - Financial Data
- The security and privacy of this data is of critical importance.



Security Issues in Database

- There are four key security issues that we need to consider in database
 - Availability
 - Authenticity
 - Integrity
 - Confidentiality



Availability

- Data should be available at all necessary times
- Data should be available to only the appropriate users
- Data should be able to track who has access to and who has accessed what data



Authenticity

- To ensure that the data has been edited by an authorized source
- To confirm that users accessing the database system are who they say they are
- To verify that all report requests are for authorized users
- To verify that any outbound data is going to the expected receiver



Integrity

- To verify that any external data has the correct formatting and other metadata
- To verify that all input data is accurate and verifiable
- To ensure that data is following the correct work flow rules for the institution/company
- To be able to report on all data changes and who authored them to ensure compliance with corporate rules and privacy laws.



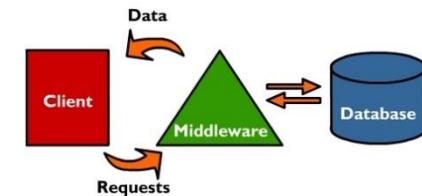
Confidentiality

- To ensure that confidential data is only available to correct people
- To ensure that entire database is secure from external and internal system breaches
- To 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



More Security Concerns

- How to keep your data confidential?
 - Internal data loss
 - External hacking
 - Securing data if hardware stolen
 - Unapproved administrator access
- Middleware security concerns
 - From middleware that sits between the user and the data
 - Single sign on authentication (the theft of one password endangers all systems)



Security Solutions for Database

- 3rd party security options
 - Most companies have several types of databases so to ensure total security across databases, they hire 3rd party database security vendors
 - Those companies have solutions for Database Activity Monitoring (DAM)
- Built in database protection
 - Companies create built-in solutions such as
 - Password controls
 - Data access based on roles and profiles
 - IP restrictions
 - Auditing capabilities of who have run what reports
 - Security logging



Pros and Cons of 3rd Party Solutions

Solution Description	Pros	Cons
Data Obfuscation (Masking, Scrambling)	Fake or Scrambled data set for use by design and implementation teams	Can be very expensive – good fake data can range in cost from \$200,000 to \$1 Million
Encryption of Data	Allows personally identifiable data to be scrambled if intrusion takes place.	Adds overhead and possible performance issues.
Database Intrusion/Extrusion Prevention	Looks for SQL Injections, Bad access commands and odd outbound data	Can eat into over head and cause performance issues – also expensive. Needs very specific criteria to set up.
Data Leak Prevention	Catches any data that is being sent out of the system	Does not protect data in the actual data warehouse.

Pros and Cons of Built In Solutions

Solution Description	Pros	Cons
Complex Passwords (require numbers and symbols) as well as frequent password changes	Makes passwords harder to guess and harder to crack	Users write them down and keep them next to computer or forget and need multiple resets
Keep Internal and External facing databases separate	Makes it very hard to hack one and then get through to the other	Reduces functionality of databases and restricts flow of internal data
Restrict Downloading	Keeps data in the database and not loose in excel, etc	Restricts reporting capabilities and off line functionality
Restrict Unwanted Connections	Again makes it harder to worm from one system to another	Makes integration more difficult and can reduce user acceptance
SAML (Security Assertion Markup Language)	SAML is the standard that is used for Single Sign On functionality	If not in use blocks the usage of single sign on

Database Privacy

“Database privacy concerns the protection of information about individuals... it is based on a balance of confidentiality, integrity and availability.”

- What is the intended use of stored data?
- Who carries the risk if data is disclosed to an unauthorized party?
- Why does someone need to know a specific piece of information?



Privacy Threats to Database Privacy

- Knowledge Discovery Data Mining (KDDM)
 - Extract information from database, and suggest a pattern regarding the data stored in the database
 - Discover patterns that classify individuals into categories, revealing in the way confidential personal information with certain probability
- Future Threat
 - United State Department of Homeland Security is planning to create a massive database to track all citizens.



Strategies for Privacy Protection

- What do we need to consider when choosing a privacy protection system:
 - Knowledge of Technology
 - Understanding of Technology
 - Implementation of Technology



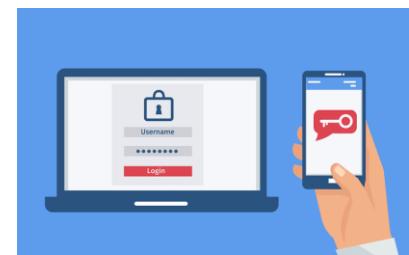
Historical Methods

- Lock and Key
 - One of the most trusted forms of protection
 - Still in use today, even with advanced technology
 - Examples
 - Trunks, Doors, Automobiles



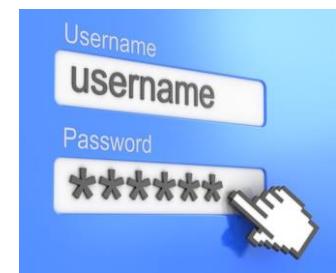
Authentication Methods

- Three main types of authentication techniques
 - What you know
 - What you have
 - What you are
- A computer or device will authenticate or validate the information the user provides



What You Know

- Most well known form of authentication
- User/Password Verification
 - User ID and Password used in order to gain access
- Problems and Security Issues
 - Potential for password to be written down on paper
 - Password can be guessed or broken easily



What You Have

- Physical object is needed in order to gain access
 - Key Card, ID Card, Token
- Most common uses include
 - Gaining access to buildings or restricted areas
- Problems and Security Issues
 - Loss of object
 - Replication

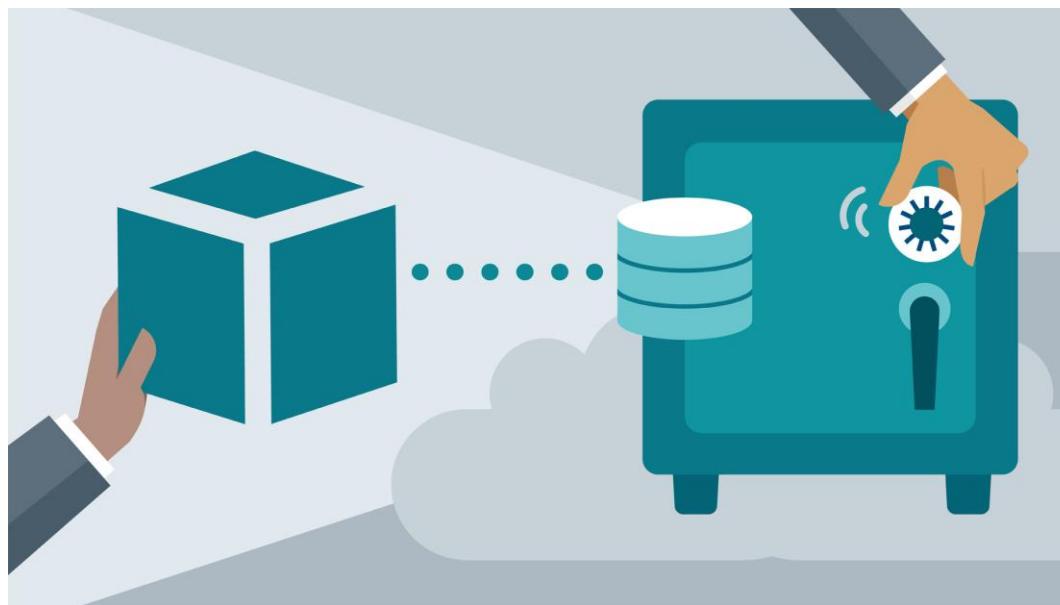


What You Are

- Biometrics
 - Unique biological characteristics used for identification
 - Technology still being developed
 - Examples
 - Fingerprint Scanners, Voice Print Recognition, Hand and/or Palm Geometry, Retinal Scan, Iris Scan, Facial Scan
- Problems and Security Issues
 - Biological or Environmental Conditions
 - Aging, Injury, Scarring, Corrective Lenses, Background Noise
 - False-Acceptance Rate and False-Rejection Rate
 - Some users might gain access without permission or vice versa



Database Anonymity Techniques



Why Anonymize in Database?

- Privacy is a complex concept !
 - Data subjects have inherent right and expectation of privacy
 - Significant legal framework relating to privacy
 - UN Declaration of Human Rights, US Constitution
 - HIPAA, Video Privacy Protection, Data Protection Acts
- For Data Sharing
 - Give real(istic) data to others to study without compromising privacy of individuals in the data



Case Study: US Census

- **Raw data:** information about every US household
 - Who, where; age, gender, racial, income and educational data
- **Why released:** determine representation, planning
- **How anonymized:** aggregated to geographic areas (Zip code)
 - Broken down by various combinations of dimensions
 - Released in full after 72 years
- **Attacks:** no reports of successful deanonymization
 - Recent attempts by FBI to access raw data rebuffed
- **Consequences:** greater understanding of US population
 - Affects representation, funding of civil projects
 - Rich source of data for future historians and genealogists

Case Study: AOL Search Data

- **Raw data:** 20M search queries for 650K users from 2006
- **Why released:** allow researchers to understand search patterns
- **How anonymized:** user identifiers removed
 - All searches from same user linked by an arbitrary identifier
- **Attacks:** many successful attacks identified individual users
 - Ego-surfers: people typed in their own names
 - Zip codes and town names identify an area
- **Consequences:** CTO resigned, two researchers fired
 - Well-intentioned effort failed due to inadequate anonymization



Two Abstract Examples

- “Census” data recording incomes and demographics
 - Schema: (**SSN, DOB, Sex, ZIP, Salary**)
 - Tabular data—best represented as a table
- “Search” data recording web searches
 - Schema: (**Uid, Kw1, Kw2, ...**)
 - Set data—each user has different set of keywords
- Each example has different anonymization needs



M	E	M
a	a	b
b	b	b

Models of Anonymization

- **Interactive Model** (akin to statistical databases)
 - Data owner acts as “gatekeeper” to data
 - Users submit queries in some agreed language
 - Gatekeeper gives an (anonymized) answer, or refuses to answer
- **“Send me your code” model**
 - Data owner executes code on their system and reports result
 - Cannot be sure that the code is not malicious
- • **Offline**, aka “publish and be damned” model
 - Data owner somehow anonymizes data set
 - Publishes the results to the world, and retires

Objectives for Anonymization

- Prevent (high confidence) inference of **associations**
 - Prevent inference of salary for an individual in “census”
 - Prevent inference of individual’s search history in “search”
 - All aim to prevent **linking** sensitive information to an individual
- Prevent inference of **presence** of an individual in the data set
 - Satisfying “presence” also satisfies “association” (not vice-versa)
 - Presence in a data set can violate privacy (eg STD clinic patients)
- Have to model what knowledge might be known to attacker
 - **Background knowledge**: facts about the data set (X has salary Y)
 - **Domain knowledge**: broad properties of data (illness Z rare in men)

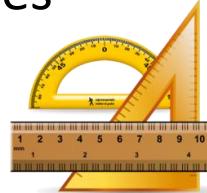


Anonymity & Utility

- Anonymization is meaningless if **utility** of data not considered
 - The empty data set has perfect privacy, but no utility
 - The original data has full utility, but no privacy
- What is “**utility**”? Depends what the application is...
 - For fixed query set, can look at max, average distortion
 - Problem for publishing: want to support unknown applications!
 - Need some way to **quantify** utility of alternate anonymizations

Measures of Utility

- Define a **surrogate measure** and try to optimize
 - Often based on the “**information loss**” of the anonymization
 - Simple example: number of rows suppressed in a table
- Give a guarantee for all queries in some **fixed class**
 - Hope the class is representative, so other uses have low distortion
 - Costly: some methods enumerate all queries, or all anonymizations
- **Empirical Evaluation**
 - Perform experiments with a reasonable workload on the result
 - Compare to results on original data (e.g. Netflix prize problems)
- Combinations of multiple methods
 - Optimize for some surrogate, but also evaluate on real queries



Definitions of Technical Terms

- **Identifiers**—uniquely identify, e.g. Social Security Number (SSN)
 - Step 0: remove all identifiers
 - Was not enough for AOL search data
- **Quasi-Identifiers (QI)**—such as DOB, Sex, ZIP Code
 - Enough to partially identify an individual in a dataset
 - DOB+Sex+ZIP unique for 87% of US Residents
- **Sensitive attributes (SA)**—the associations we want to hide
 - Salary in the “census” example is considered sensitive
 - Not always well-defined: only some “search” queries sensitive
 - SA can be identified: bonus may identify salary...



Anonymization as Uncertainty

- Anonymization -- adding uncertainty to certain data
 - To ensure an attacker can't be sure about associations, presence
- It is important to use the tools and models of uncertainty
 - To quantify the uncertainty of an attacker
 - To understand the impact of background knowledge
 - To allow efficient, accurate querying of anonymized data



Uncertainty & Possible Worlds

- Uncertain Data typically represents multiple **possible worlds**
 - Each possible world corresponds to a database (or...)
 - The uncertainty model may attach a probability to each world
 - Queries conceptually range over all possible worlds
- **Possibilistic** interpretations
 - Is a given fact possible (\exists a world W where it is true) ?
 - Is a given fact certain (\forall worlds W it is true) ?
- **Probabilistic** interpretations
 - What is the probability of a fact being true?
 - What is the distribution of answers to an aggregate query?
 - What is the (min, max, mean) answer to an aggregate query?

Tabular Data Example

- Census data recording incomes and demographics

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

- Releasing SSN → Salary association **violates** individual's privacy
 - SSN is an identifier, Salary is a sensitive attribute (SA)



Tabular Data Example: De-Identification

- **Census data:** remove SSN to create 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

- Does the de-identified table preserve an individual's privacy?
 - Depends on what other information an attacker knows



Tabular Data Example: Linking Attack

- De-identified private data + publicly available data

DOB	Sex	ZIP	Salary	SSN	DOB
1/21/76	M	53715	50,000	11-1-111	1/21/76
4/13/86	F	53715	55,000	33-3-333	2/28/76
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		

- Cannot uniquely identify either individual's salary
 - DOB is a **quasi-identifier** (QI)



Tabular Data Example: Linking Attack

- De-identified private data + publicly available data

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
11-1-111	1/21/76	M
33-3-333	2/28/76	M

- Uniquely identified one individual's salary, but not the other's
 - DOB, Sex are **quasi-identifiers** (QI)



Tabular Data Example: Linking Attack

- De-identified private data + publicly available data

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

- Uniquely identified both individuals' salaries
 - [DOB, Sex, ZIP] is unique for lots of US residents



Tabular Data Example: Anonymization

- Anonymization through **tuple suppression**

DOB	Sex	ZIP	Salary
*	*	*	*
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
*	*	*	*
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

- Cannot link to private table even with knowledge of QI values
 - Missing tuples could take any value from the space of all tuples
 - Introduces a lot of uncertainty



Tabular Data Example: Anonymization

- Anonymization through QI attribute generalization

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

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

- Cannot uniquely identify tuple with knowledge of QI values
 - More precise form of uncertainty than tuple suppression
 - E.g., $\text{ZIP} = 537** \rightarrow \text{ZIP} \in \{53700, \dots, 53799\}$



Tabular Data Example: Anonymization

- Anonymization through sensitive attribute (SA) permutation

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

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

- Can uniquely identify tuple, but uncertainty about SA value
 - Much more precise form of uncertainty than generalization



Tabular Data Example: Anonymization

- Anonymization through sensitive attribute (SA) perturbation

DOB	Sex	ZIP	Salary
1/21/76	M	53715	60,000
4/13/86	F	53715	45,000
2/28/76	M	53703	60,000
1/21/76	M	53703	55,000
4/13/86	F	53706	80,000
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

- Can uniquely identify tuple, but get “noisy” SA value



k-Anonymization

- k-anonymity: Table T satisfies k-anonymity wrt quasi-identifier QI iff each tuple in (the multiset) $T[QI]$ appears at least k times
 - Protects against “linking attack”
- k-anonymization: Table T' is a k-anonymization of T if T' is a generalization/suppression of T , and T' satisfies k-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	537**	50,000
4/13/86	F	537**	55,000
2/28/76	*	537**	60,000
1/21/76	M	537**	65,000
4/13/86	F	537**	70,000
2/28/76	*	537**	75,000



k-Anonymization and Uncertainty

- **Intuition:** A k-anonymized table T' represents the set of all “possible world” tables T_i s.t. T' is a k-anonymization of T_i
- The table T from which T' was originally derived is one of the possible worlds

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



k-Anonymization and Uncertainty

- **Intuition:** A k-anonymized table T' represents the set of all “possible world” tables T_i s.t. T' is a k-anonymization of T_i
- (Many) other tables are also possible

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



DOB	Sex	ZIP	Salary
1/21/76	M	53710	50,000
4/13/86	F	53715	55,000
2/28/76	F	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	M	53715	75,000



k-Anonymization and Uncertainty

- **Intuition:** A k-anonymized table T' represents the set of all “possible world” tables T_i s.t. T' is a k-anonymization of T_i
 - If no background knowledge, all possible worlds are equally likely
 - Easily representable in systems for uncertain data (see later)
- **Query Answering**
 - Queries should (implicitly) range over all possible worlds
 - Example query: what is the salary of individual (1/21/76, M, 53715)? Best guess is 57,500 (weighted average of 50,000 and 65,000)
 - Example query: what is the maximum salary of males in 53706? Could be as small as 50,000, or as big as 75,000



Computing k-Anonymizations

- Variations depend on search space and algorithm
 - Generalization vs (tuple) suppression
 - Global (e.g., full-domain) vs local (e.g., multidimensional) recoding
 - Hierarchy-based vs partition-based



Incognito

- Computes all “minimal” full-domain generalizations
 - Uses ideas from data cube computation, association rule mining
- Key intuitions for efficient computation:
 - **Subset Property:** If table T is k -anonymous wrt a set of attributes Q , then T is k -anonymous wrt any set of attributes that is a subset of Q
 - **Generalization Property:** If table T_2 is a generalization of table T_1 , and T_1 is k -anonymous, then T_2 is k -anonymous
- Properties useful for stronger notions of privacy too!
 - ℓ -diversity, t -closeness



Incognito - Example

- Every full-domain generalization described by a “domain vector”
 - $B0=\{1/21/76, 2/28/76, 4/13/86\} \rightarrow B1=\{76-86\}$
 - $S0=\{M, F\} \rightarrow S1=\{*\}$
 - $Z0=\{53715, 53710, 53706, 53703\} \rightarrow Z1=\{5371*, 5370*\} \rightarrow Z2=\{537**\}$

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

$B0, S1, Z2$ →

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



Incognito - Example

- Every full-domain generalization described by a “domain vector”
 - $B0=\{1/21/76, 2/28/76, 4/13/86\} \rightarrow B1=\{76-86\}$
 - $S0=\{M, F\} \rightarrow S1=\{*\}$
 - $Z0=\{53715, 53710, 53706, 53703\} \rightarrow Z1=\{5371*, 5370*\} \rightarrow Z2=\{537**\}$

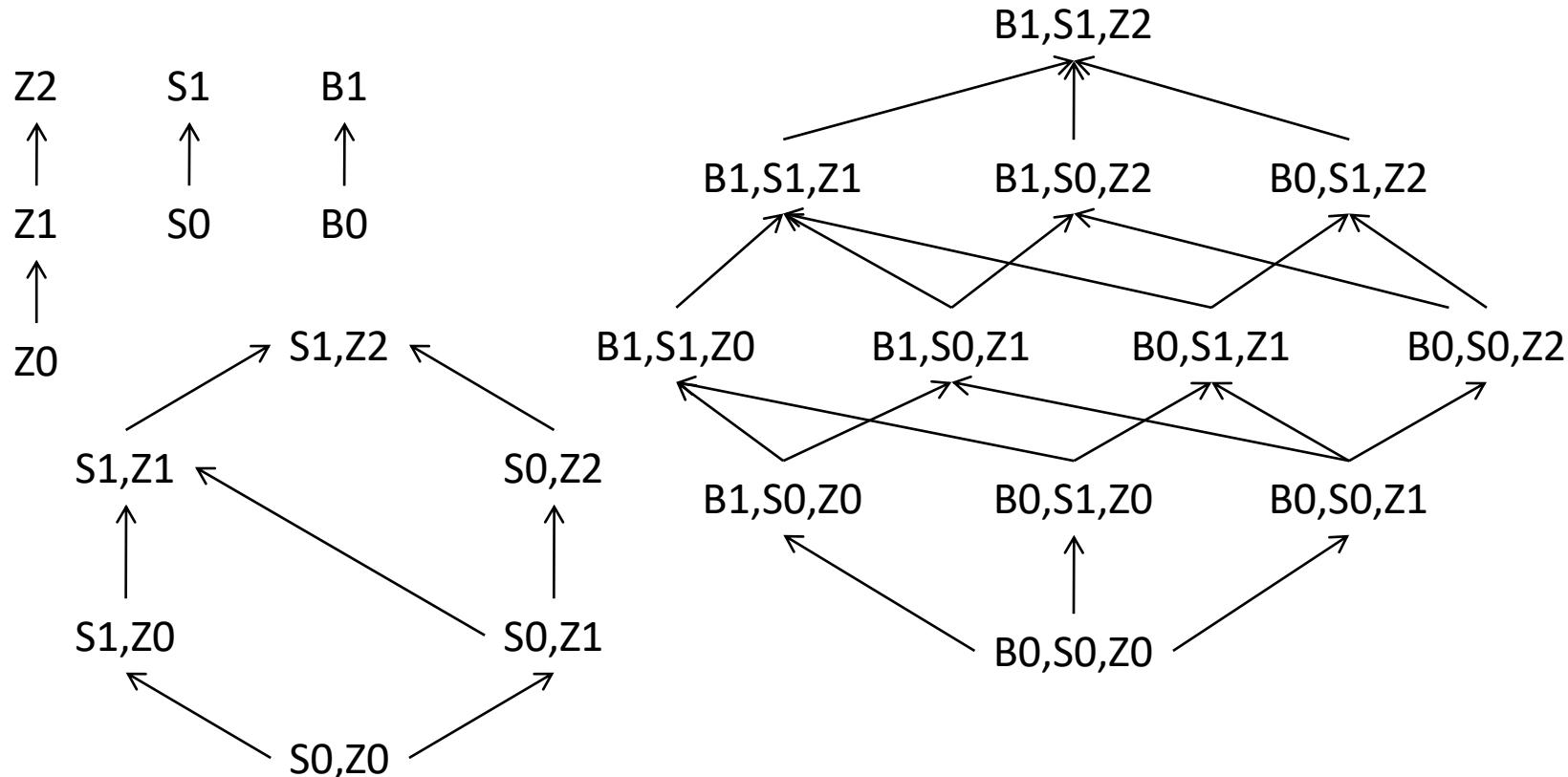
DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

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



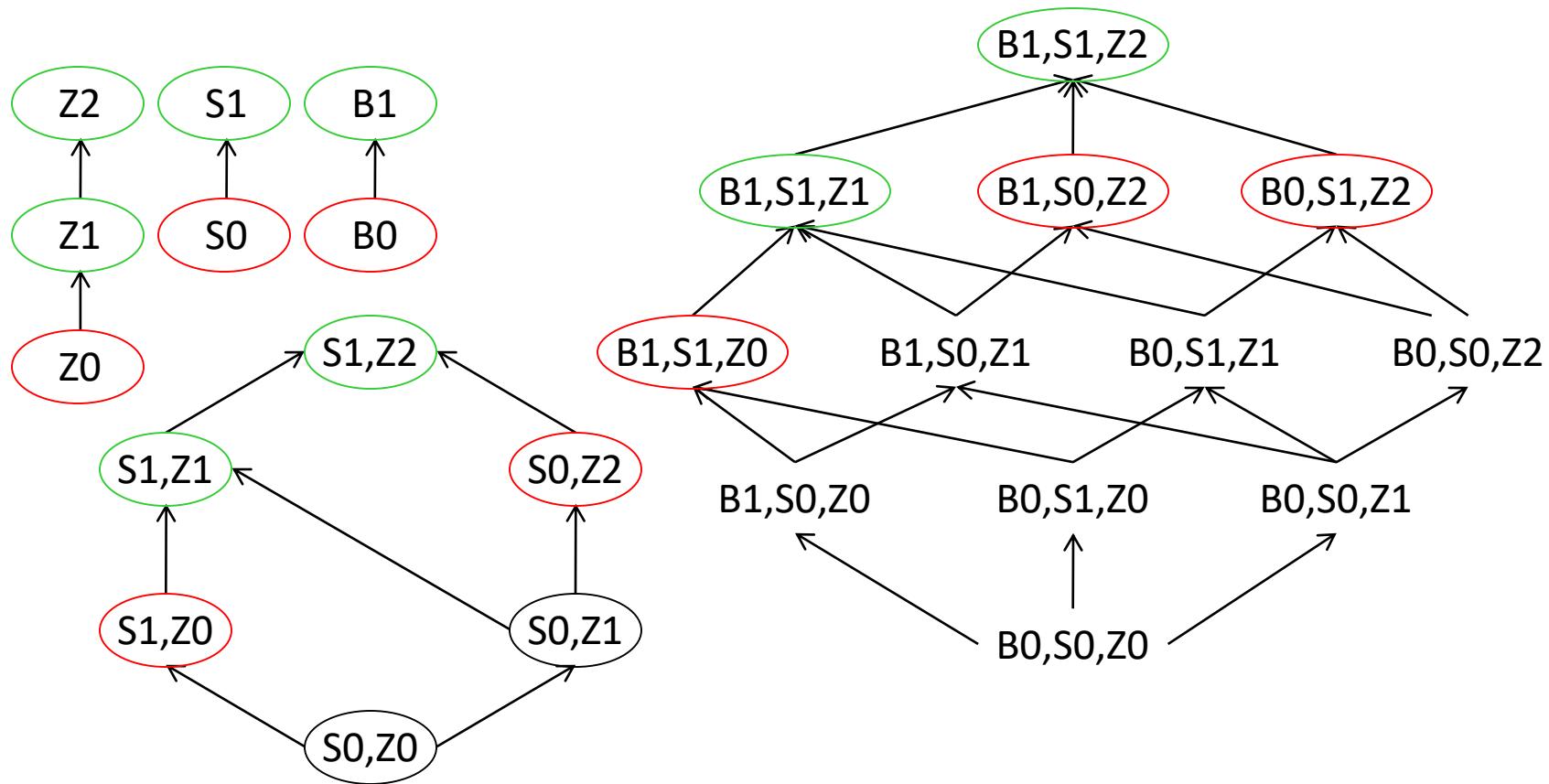
Incognito: Lattice of Domain Vectors



- Computes all “**minimal**” full-domain generalizations
 - Set of minimal full-domain generalizations forms an anti-chain
 - Can use any reasonable utility metric to choose “**optimal**” solution



Incognito: Lattice of Domain Vectors ...

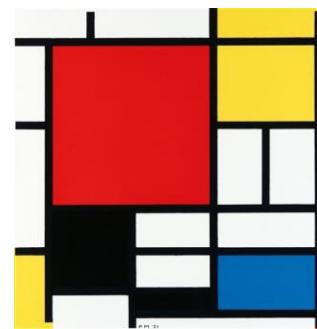


- Computes all “**minimal**” full-domain generalizations
 - Set of minimal full-domain generalizations forms an anti-chain
 - Can use any reasonable utility metric to choose “**optimal**” solution



Mondrian

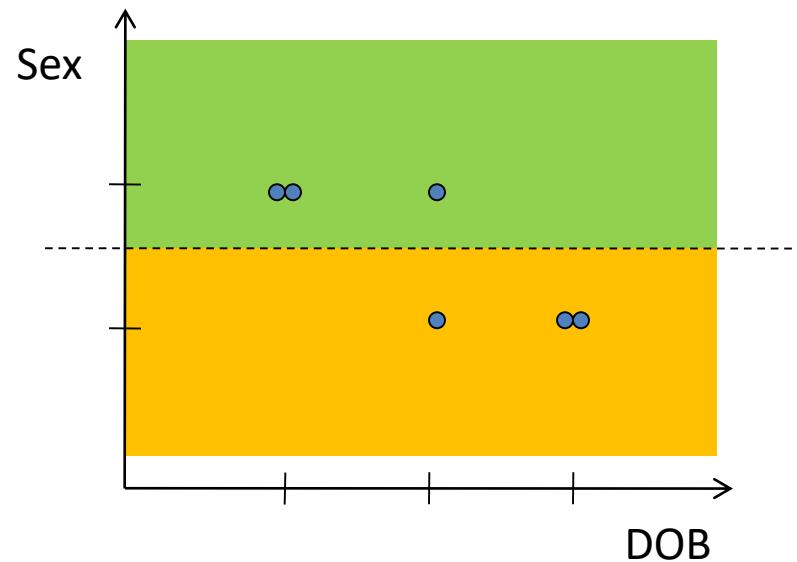
- Computes one “good” multi-dimensional generalization
 - Uses local recoding to explore a larger search space
 - Treats all attributes as ordered, chooses partition boundaries
- Utility metrics
 - **Discernability**: sum of squares of group sizes
 - **Normalized average group size** = $(\text{total tuples} / \text{total groups}) / k$
- **Efficient**: greedy $O(n \log n)$ heuristic for NP-hard problem
- **Quality guarantee**: solution is a constant-factor approximation



Mondrian - Example

- Uses ideas from spatial kd-tree construction
 - QI tuples = points in a multi-dimensional space
 - Hyper-rectangles with $\geq k$ points = k-anonymous groups
 - Choose axis-parallel line to partition point-multiset at median

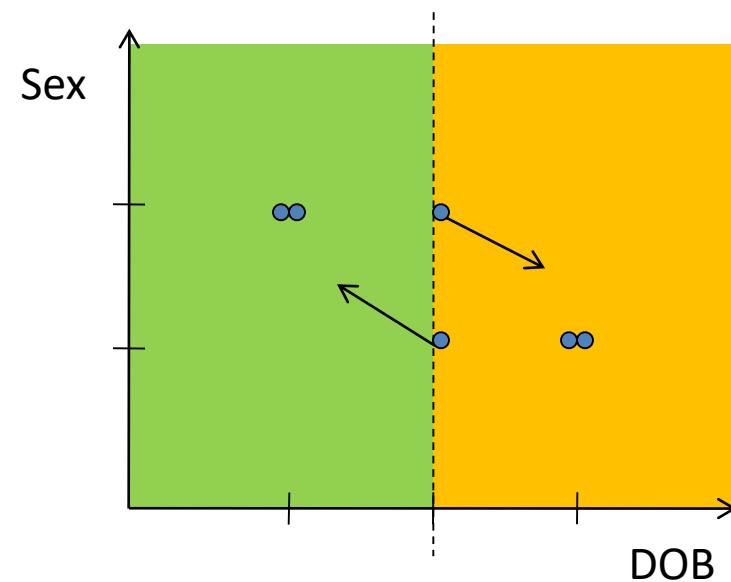
DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000



Mondrian – Example ...

- Uses ideas from spatial kd-tree construction
 - QI tuples = points in a multi-dimensional space
 - Hyper-rectangles with $\geq k$ points = k-anonymous groups
 - Choose axis-parallel line to partition point-multiset at median

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



Homogeneity Attack

- **Issue:** k-anonymity requires each tuple in (the multiset) $T[QI]$ to appear $\geq k$ times, but does not say anything about the SA values
 - If (almost) all SA values in a QI group are equal, loss of privacy!
 - The problem is with the choice of grouping, not the data

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	50,000
4/13/86	F	53706	55,000
2/28/76	F	53706	60,000

Not Ok! →

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**	50,000
4/13/86	*	537**	55,000
2/28/76	*	537**	60,000



Homogeneity Attack ...

- **Issue:** k-anonymity requires each tuple in (the multiset) $T[QI]$ to appear $\geq k$ times, but does not say anything about the SA values
 - If (almost) all SA values in a QI group are equal, loss of privacy!
 - The problem is with the choice of grouping, not the data

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	50,000
4/13/86	F	53706	55,000
2/28/76	F	53706	60,000

Ok! →

DOB	Sex	ZIP	Salary
76-86	*	53715	50,000
76-86	*	53715	55,000
76-86	*	53703	60,000
76-86	*	53703	50,000
76-86	*	53706	55,000
76-86	*	53706	60,000



Homogeneity and Uncertainty

- **Intuition:** A k-anonymized table T' represents the set of all “possible world” tables T_i s.t. T' is a k-anonymization of T_i
- Lack of diversity of SA values implies that in a large fraction of possible worlds, some fact is true, which can violate privacy

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**	50,000
4/13/86	*	537**	55,000
2/28/76	*	537**	60,000

SSN	DOB	Sex	ZIP
11-1-111	1/21/76	M	53715



ℓ -Diversity

- ℓ -Diversity Principle: a table is ℓ -diverse if each of its QI groups contains at least ℓ “well-represented” values for the SA
 - Statement about possible worlds
- Different definitions of ℓ -diversity based on formalizing the intuition of a “well-represented” value
 - Entropy ℓ -diversity: for each QI group g , $\text{entropy}(g) \geq \log(\ell)$
 - Recursive (c, ℓ) -diversity: for each QI group g with m SA values, and r_i the i 'th highest frequency, $r_1 < c (r_1 + r_{\ell+1} + \dots + r_m)$
 - Folk ℓ -diversity: for each QI group g , no SA value should occur more than $1/\ell$ fraction of the time = Recursive($1/\ell$, 1)-diversity



ℓ -Diversity - Example

- Intuition: Most frequent value does not appear too often compared to the less frequent values in a QI group
- Entropy ℓ -diversity: for each QI group g , $\text{entropy}(g) \geq \log(\ell)$
 - ℓ -diversity((1/21/76, *, 537**)) = ?? = 1

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**	50,000
4/13/86	*	537**	55,000
2/28/76	*	537**	60,000



Computing ℓ -Diversity

- **Key Observation:** entropy ℓ -diversity and recursive(c, ℓ)-diversity possess the Subset Property and the Generalization Property
- **Algorithm Template:**
 - Take **any** algorithm for k-anonymity and replace the k-anonymity test for a generalized table by the ℓ -diversity test
 - Easy to check based on counts of SA values in QI groups



t-Closeness

- Limitations of ℓ -diversity
 - Similarity attack: SA values are distinct, but semantically similar

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**	50,001
4/13/86	*	537**	55,001
2/28/76	*	537**	60,001

SSN	DOB	Sex	ZIP
11-1-111	1/21/76	M	53715

- **t-Closeness Principle:** a table has t-closeness if in each of its QI groups, the distance between the distribution of SA values in the group and in the whole table is no more than threshold t



Answering Queries on Generalized Tables

- **Observation:** Generalization loses a lot of information, resulting in inaccurate aggregate analyses
- How many people were born in 1976?
 - Bounds = [1,5], selectivity estimate = 1, actual value = 4

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



Answering Queries on Generalized Tables

- **Observation:** Generalization loses a lot of information, resulting in inaccurate aggregate analyses
- What is the average salary of people born in 1976?
 - Bounds = [50K,75K], actual value = 62.5K

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

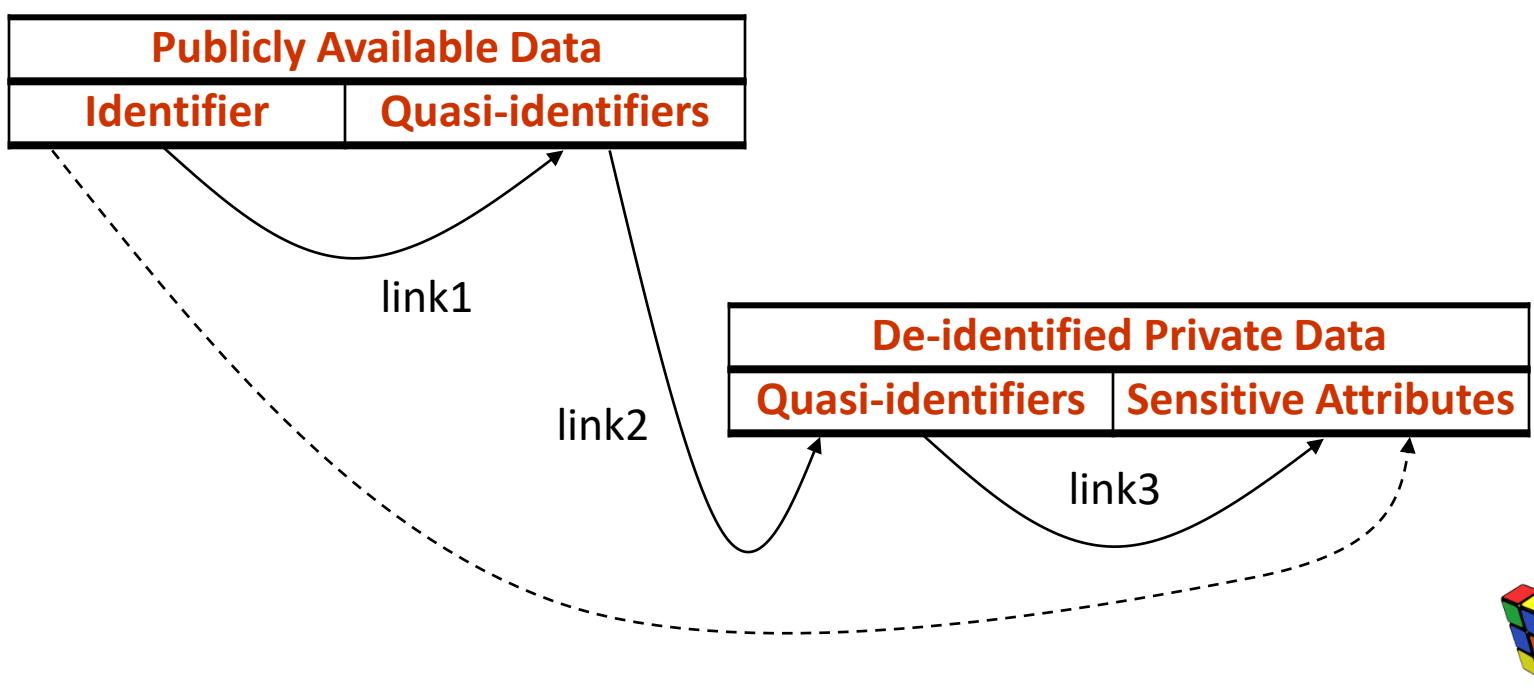


DOB	Sex	ZIP	Salary
76-86	M	537**	50,000
76-86	F	537**	55,000
76-86	M	537**	60,000
76-86	M	537**	65,000
76-86	F	537**	70,000
76-86	F	537**	75,000



Permutation: A Viable Alternative

- **Observation:** Identifier → SA is a composition of link1, link2, link3
 - Generalization-based techniques weaken link2
- **Alternative:** Weaken link 3 (QI → SA association in private data)



Permutation: Basics

- Partition private data into groups of tuples, permute SA values wrt QI values in each group
- For individuals known to be in private data, **same privacy guarantee** as generalization

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



Permutation: Aggregate Analyses

- **Key observation:** Exact QI and SA values are available
- How many people were born in 1976?
 - Estimate = 4, actual value = 4

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



Permutation: Aggregate Analyses ...

- **Key observation:** Exact QI and SA values are available
- What is the average salary of people born in 1976?
 - Estimated bounds = [57.5K, 62.5K], actual value = 62.5K

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

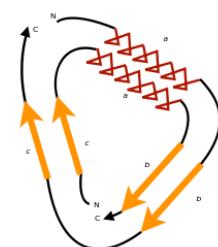


DOB	Sex	ZIP	Salary
1/21/76	M	53715	60,000
4/13/86	F	53715	75,000
2/28/76	M	53703	65,000
1/21/76	M	53703	50,000
4/13/86	F	53706	70,000
2/28/76	F	53706	55,000



Computing Permutation Groups

- Can use grouping obtained by any previously discussed approach
 - Instead of generalization, use permutation
 - For same groups, permutation **always** has lower information loss
- Anatomy: form l -diverse groups
 - Hash SA values into buckets
 - Iteratively pick 1 value from each of the l most populated buckets
- Permutation: use numeric diversity
 - Sort (ordered) SA values
 - Pick k adjacent values subject to numeric diversity condition



Permutation and Uncertainty

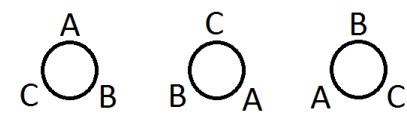
- **Intuition:** A permuted (QI, SA) table T' represents the set of all “possible world” tables T_i s.t. T' is a (QI, SA) permutation of T_i
- **Issue:** The SA values taken by different tuples in the same QI group are not independent of each other

The diagram illustrates the concept of permutation and uncertainty. On the left, there are two separate tables: one for QI (DOB, Sex, ZIP) and one for Salary. The QI table has 6 rows, and the Salary table has 6 rows. An arrow labeled "No!" points from these two tables to a third table on the right, which represents a permuted version where the salary values are no longer independent across the QI groups.

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

→

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

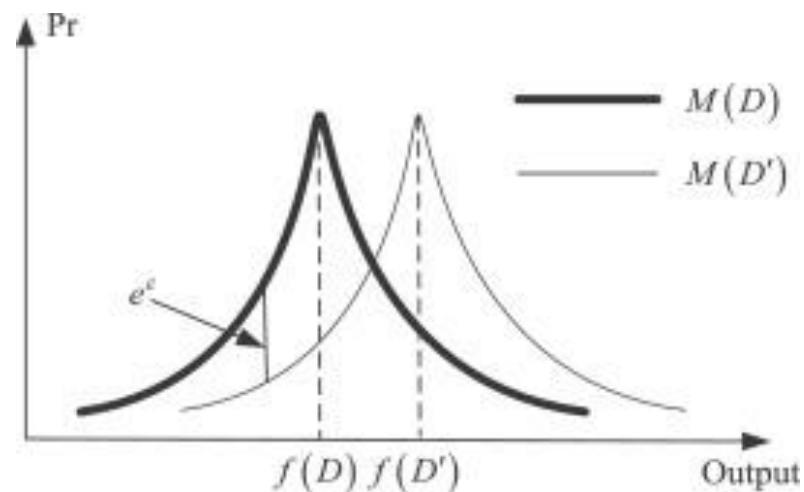


Summary of Tabular Anonymization

- Anonymization Techniques
 - Generalization + Suppression: natural representation and efficient reasoning using Uncertain Database models
 - Permutation:
- Recent Attacks
 - Minimality Attack:
 - Uses knowledge of anonymization algorithm to argue some possible worlds are not consistent with output
 - deFinetti Attack:
 - Uses knowledge from anonymized data to argue some associations are more likely than others
 - New research: analyze, understand their practical impact
 - Best understood via probability and uncertainty



Differential Privacy Techniques



Cynthia Dwork. 2006. Differential privacy. In Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP'06), Springer-Verlag, Berlin, Heidelberg, 1-12. DOI=http://dx.doi.org/10.1007/11787006_1

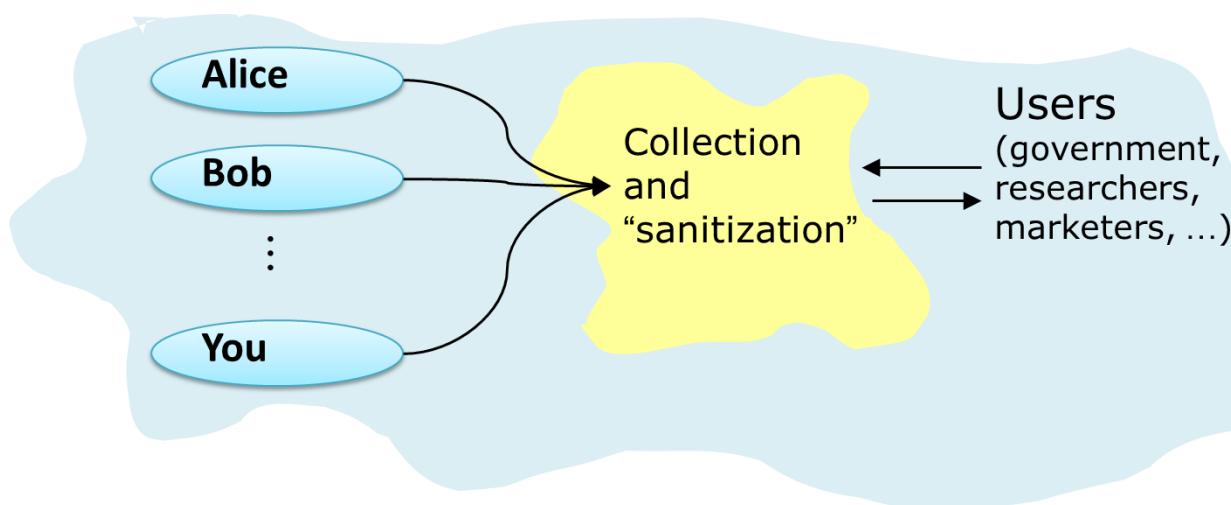
Recall of Database Privacy

- Blending/hiding into a crowd
 - K-anonymity, l-diversity, etc. approaches
 - Adversary may have various background knowledge to breach privacy
 - Privacy models often assume “**the adversary’s background knowledge is given**”, which is **impractical**



Interactive Database Query (IDQ)

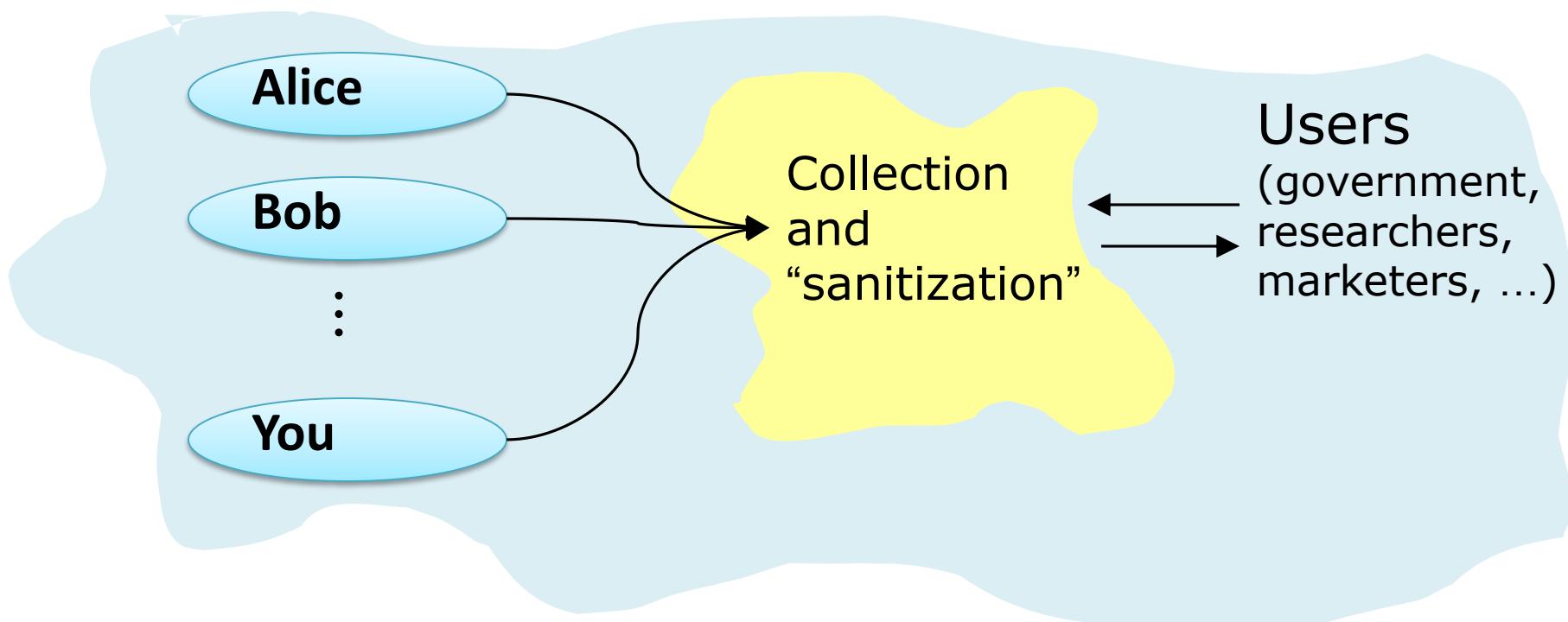
- A classical research problem for statistical databases
- Prevent query inferences – malicious users submit multiple queries to infer private information about some person



Privacy & Utility

Two conflicting goals

- **Utility**: Users can extract “global” statistics
- **Privacy**: Individual information stays hidden
- How can these be formalized?



Privacy & Utility

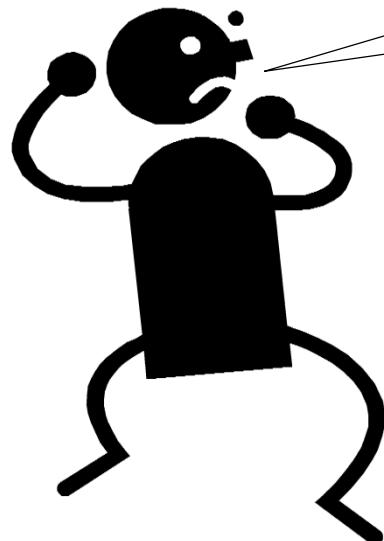
- **Challenge:** How can you allow meaningful usage of such datasets while preserving individual privacy?
- Blatant Non-Privacy
 - Leak individual records
 - Can link with public databases to re-identify individuals
 - Allow adversary to reconstruct database with significant probability
 - **Privacy models often assume “the adversary’s background knowledge is given”, which is impractical**



Attempts

- Attempt 1: Crypto-ish Definitions

I am releasing some useful statistic $f(D)$,
and nothing more will be revealed.



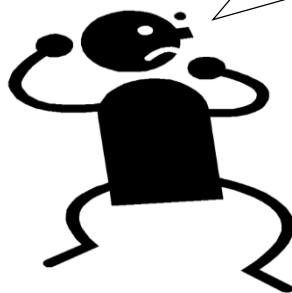
What kind of statistics are
safe to publish?

Example:

If I know Alice's height is 2 inches higher than the average American's height, by looking at the census database, I can find the average and then calculate Alice's exact height. Therefore, Alice's privacy is breached.

Attempt 2: Absolute Disclosure Prevention

I am releasing research findings showing that people who smoke are very likely to get cancer.



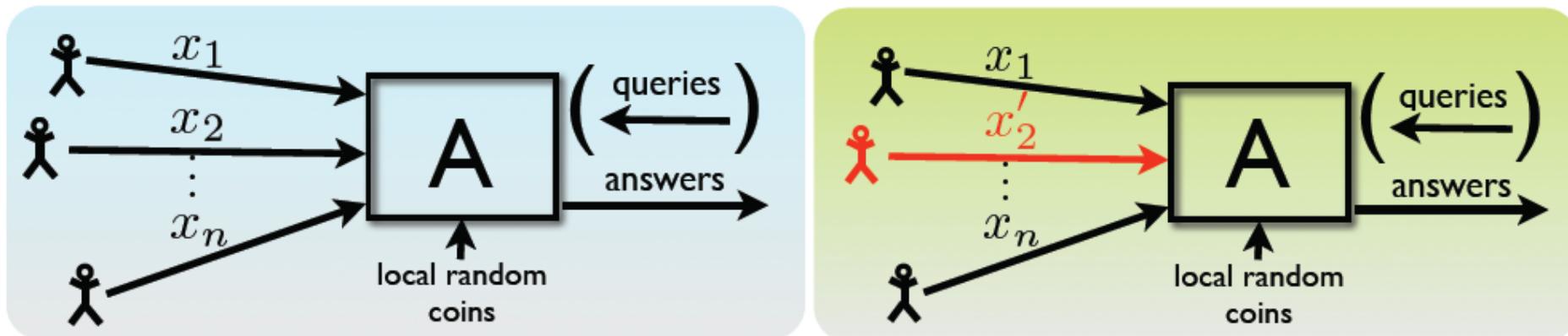
You cannot do that, since it will break my privacy. My insurance company happens to know that I am a smoker...



- “If the release of statistics S makes it possible to determine the value [of private information] more accurately than is possible without access to S , a disclosure has taken place.” [Dalenius]
- It is not possible to design any non-trivial mechanism that satisfies such strong notion of privacy. [Dalenius]

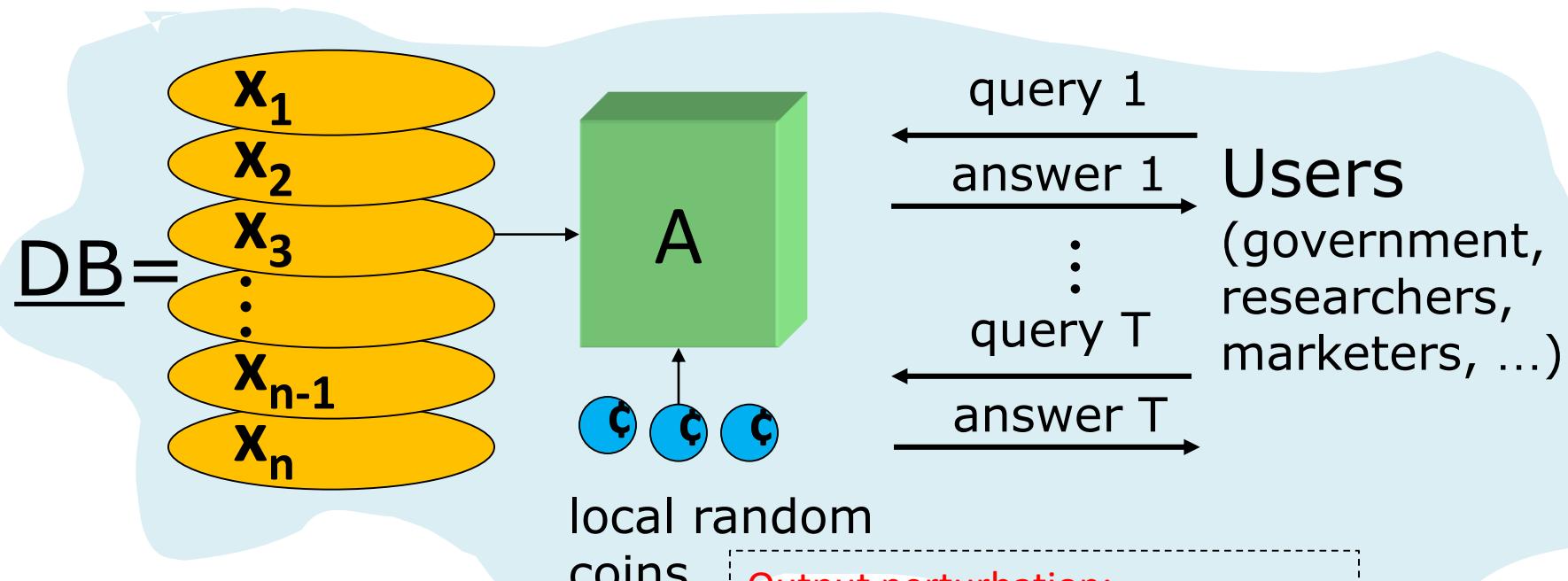
Attempt 3: Differential Privacy (DP)

- The risk to my privacy should not substantially increase as a result of participating in a statistical database
- From the released statistics, it is hard to tell which case it is.



X and X' are two datasets, X is a neighbor of X' if they differ in one row

Can DP Make IDQ Privacy-Preserving?



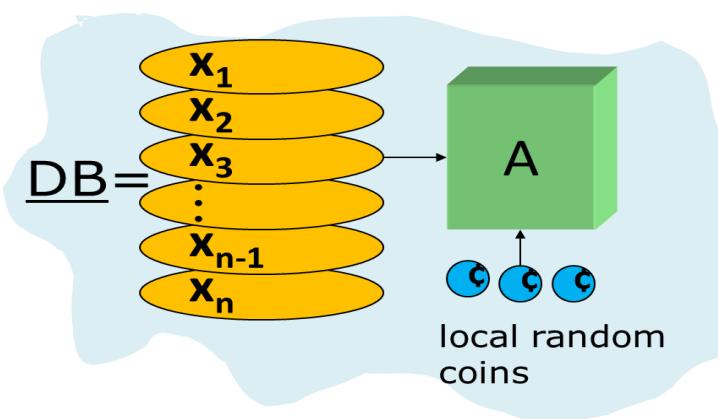
Notable Properties of DP

- Adversary knows arbitrary auxiliary information
 - No linkage attacks
- Sanitizer need not know the adversary's prior distribution on the DB



Differential Privacy

- Goal: **Protect individual participants**



D: All Databases

A: Sanitizer: must be a random algorithm, why?

T: All Output Events

Query: The number of people who smoke, i.e.,
of 1's in D ?

$$\text{? } A(D) = \sum x_i \in T \rightarrow \checkmark A(D) = \sum x_i + \text{noise} \in T$$

DB

No.	Name	Smoke? $x_i \in \{0,1\}$
1	Alice	1
2	Bob	0
3	Carlo	0
...
i	Lu	1
...
n	Tom	1

DB

No.	Name	Smoke? $x_i \in \{0,1\}$
1	Alice	1
2	Bob	0
3	Carlo	0
...
i	Lu	1
...
n	Tom	1

D

No.	Smoke? $x_i \in \{0,1\}$
1	1
2	0
3	0
...	..
i	1
...	..
n	1

DP: Protect individual participants

- **I**: all individuals in **D**
- **D+I**: Database includes the record **I**
- **D-I**: Database does not include the record **I**
- **(D+I, D-I)**: called adjacent databases
 - Differing on **at most one** element
- Example

No.	Name	Smoke? $x_i \in \{0,1\}$
1	Alice	1
2	Bob	0
3	Carlo	0
...
<i>i</i>	Lu	1
...
<i>n</i>	Tom	1

DB with Bob

No.	Name	Smoke? $x_i \in \{0,1\}$
1	Alice	1
2	Bob	0
3	Carlo	0
...
<i>i</i>	Lu	1
...
<i>n</i>	Tom	1

DB without Bob

No.	Smoke? $x_i \in \{0,1\}$
1	1
2	0
3	0
...	..
<i>i</i>	1
...	..
<i>n</i>	1

 $D + I$ 

No.	Smoke? $x_i \in \{0,1\}$
1	1
2	0
3	0
...	..
<i>i</i>	1
...	..
<i>n</i>	1

 $D - I$

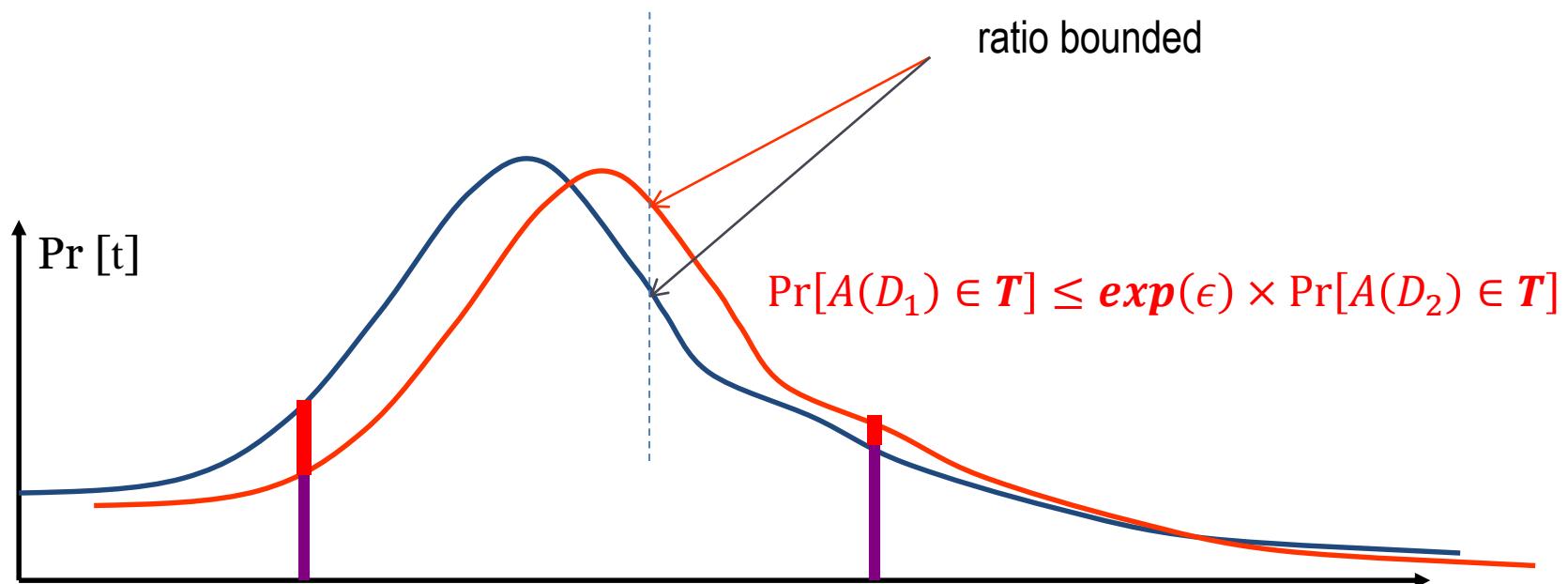
ϵ -Differential Privacy



- Goal:
 - The risk to my privacy should not substantially increase as a result of participating in a statistical database.
 - With or without including me in the database, my privacy risk should not change much
- Probability of every **bad** event - or any event - **increases only by small multiplicative factor (SMF) when I enter the DB.**
 - For example $\Pr[A(D + I) \in T] \approx \text{SMF} * \Pr[A(D - I) \in T]$
- Formally,
 - A randomized sanitizer function A gives the ϵ -Differential Privacy if for all adjacent DBs $D_1 = D + I$ and $D_2 = D - I$, and all possible output $T \in \text{Range}(A)$
$$\Pr[A(D_1) \in T] \leq \exp(\epsilon) \times \Pr[A(D_2) \in T]$$
The probability is taken over the coin tosses of A .

ϵ -Differential Privacy ...

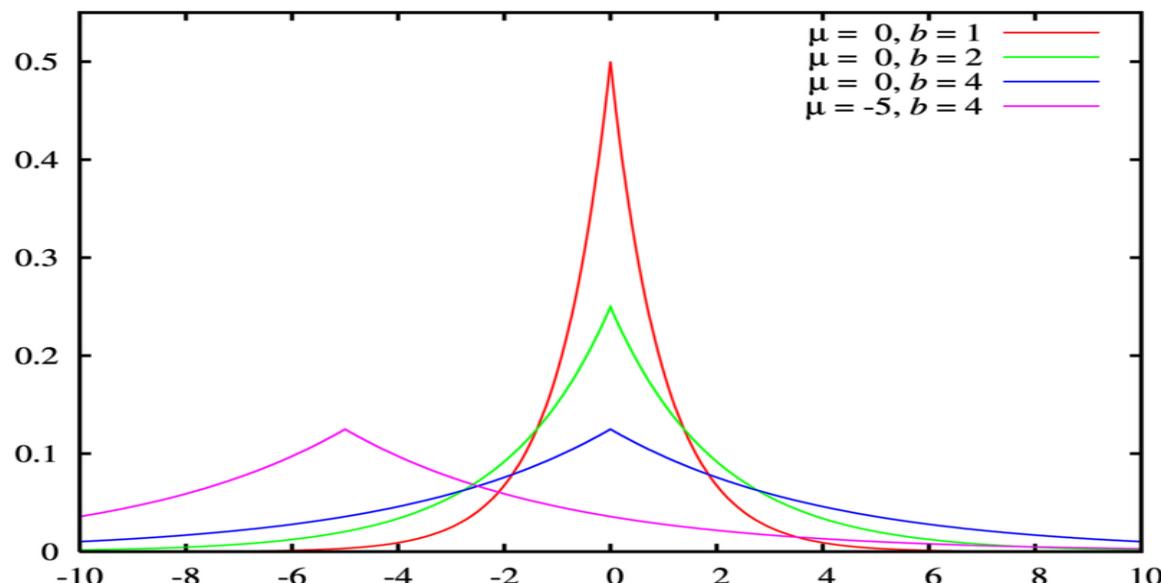
- No perceptible risk is incurred by joining DB.
 - Participation in the DB poses no **additional** risk. Any info adversary can obtain, it could obtain without **me** (my data).



What is the noise?

- Using laplacian distribution (a symmetric exponential distribution) to generate noise.
 - Add noise from a symmetric continuous distribution to true answer

$$x \in [-\infty, +\infty]$$



- Mean: μ
- Variance: $2b^2$
- probability density function (PDF) is

$$\frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

Laplace distribution

- We generate noise using the Laplace distribution.
- The Laplace distribution, denoted $\text{Lap}(b)$, is defined with parameter b and has density function:

$$\frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \quad \text{when considering } \mu = 0, \text{ we have } \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$$

- Taking $b = 1/\epsilon$ we have immediately that the density is proportional to $\exp(-\epsilon|x|) = e^{-\epsilon|x|}$.
- This distribution has its highest density at 0.
- **For any x, x' such that $|x - x'| \leq 1$, the density at x is at most e^ϵ times the density at x' .**
- The distribution is symmetric about 0.
- The distribution flattens as ϵ decreases. More likely to deviate from the true value.



How to add the noise?

- Sensitivity of Number Functions
 - we need to calibrate the noise to the influence an individual can have on the output -> **Privacy & Utility**
- The (**global**) sensitivity of a function F is the maximum (absolute) change over all possible adjacent inputs
$$S(F) = \text{Max}_{D_1, D_2 : |D_1 - D_2| = 1} |F(D_1) - F(D_2)|$$
- **Intuition:** $S(F)$ characterizes the scale of the influence of one individual, and hence how much noise we must add.



Sensitivity of Number Functions

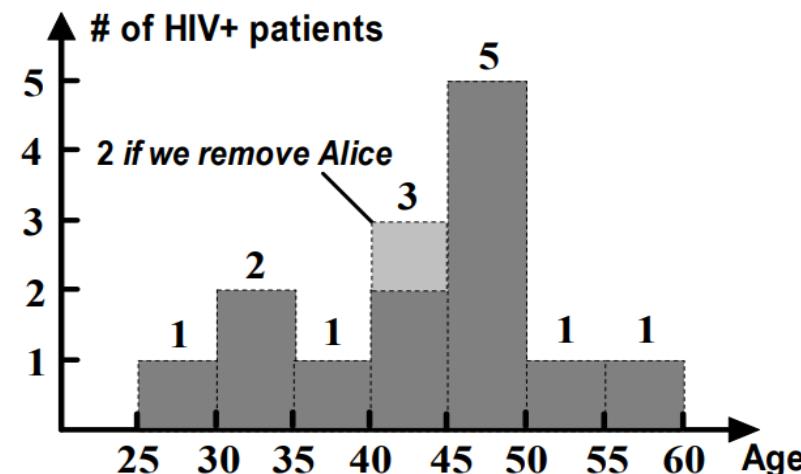
- **$S(F)$** is small for a lot of common functions
 - COUNT: $S(F) = 1$
 - $n = |D|$ Distance $|F(D_1) - F(D_2)|$
 - HISTOGRAM: $S(F) = 2$
 - $n = \|\mathbf{D}\|_1 = \sum_{i=1}^{|X|} |D[i]|$ Distance $\|F(D_1) - F(D_2)\|_1$
 - Bounded for other functions (MEAN, covariance matrix...)
 - Many data-mining algorithms can be implemented through a sequence of low-sensitivity queries
- **Definition** l_1 (Manhattan) Distance
 - For a vector $v = (v_1, v_2, \dots, v_d) \in R^d$, $\|\mathbf{v}\|_1 = \sum_{i=1}^d |v_i|$
- High Global Sensitivity
 - Order statistics (smallest order statistic, largest order statistic))
 - Clustering $x_{(1)} = \min(x_1, x_2, \dots, x_n)$ $x_{(n)} = \max(x_1, x_2, \dots, x_n)$



HISTOGRAM: $S(F) = 2$ Why?

Name	Age	HIV+
Alice	42	Yes
Bob	31	Yes
Carol	32	Yes
Dave	36	No
Ellen	43	Yes
Frank	41	Yes
Grace	26	Yes
...

(a) Example sensitive data



(b) Unperturbed histogram

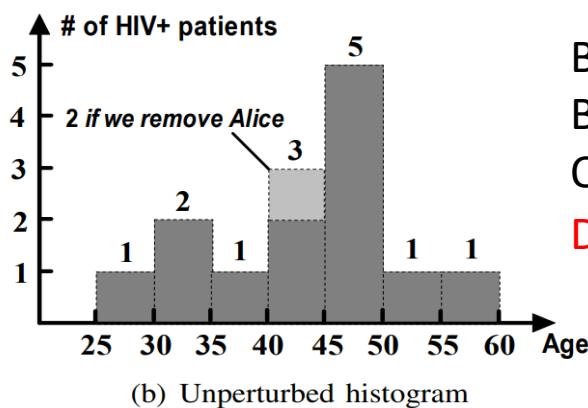
- Such histograms are commonly found, e.g., in the published statistics by Singapore's Ministry of Health².
- The application of DP to such histograms guarantees that **changing or removing** any record from the database has negligible impact on the output histogram.
- This means that the adversary cannot infer whether a specific patient (say, Alice) is infected by HIV, even if s/he knows the HIV status of all the remaining patients in the database.

HISTOGRAM: $S(F) = 2$ Why?

- Removing any record affects one bin
- Changing any record affects two bin, **in the worst case**, e.g., changing the age. Therefore, $S(F)=2$

Name	Age	HIV+
Alice	42	Yes
Bob	31	Yes
Carol	32	Yes
Dave	36	No
Ellen	43	Yes
Frank	41	Yes
Grace	26	Yes
...

(a) Example sensitive data



(b) Unperturbed histogram

Bin(35-40): Difference $|1-2|=1$ Bin(40-45): Difference $|3-2|=1$

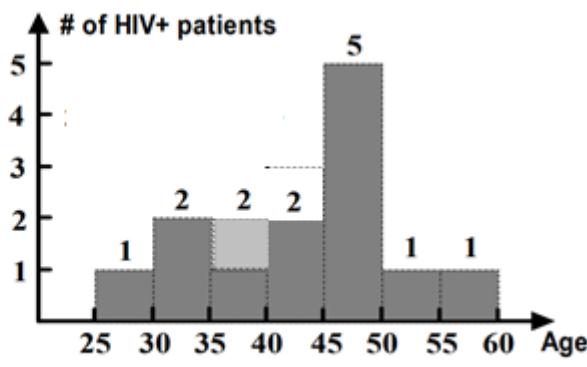
Other Bins: No Difference

$$\text{Distance } \left\| F(D_1) - F(D_2) \right\|_1 = 1 + 1 = 2$$

37

Name	Age	HIV+
Alice	42	Yes
Bob	31	Yes
Carol	32	Yes
Dave	36	No
Ellen	43	Yes
Frank	41	Yes
Grace	26	Yes
...

(a) Example sensitive data



(b) Unperturbed histogram

Add the Noise After Knowing $S(F)$

- Laplace Mechanism with Sensitivity
 - Release $F(x) + \text{Lap}(S(F)/\epsilon)$ to obtain ϵ -DP guarantee
 - $F(x)$ = true answer on input x
 - $\text{Lap}(\lambda)$ = noise sampled from Laplace distribution with parameter $\lambda = S(F)/\epsilon$
 - Intuition on impact of parameters of differential privacy (DP):
 - Larger $S(F)$, more noise (need more noise to mask an individual)
 - Smaller ϵ , more noise (more noise increases privacy)
 - Expected magnitude of $|\text{Lap}(\lambda)|$ is (approx) $1/\lambda$?



Proof of ϵ -Differential Privacy

- A randomized sanitizer function A gives the ϵ -Differential Privacy if for all adjacent DBs $D_1 = D + I$ and $D_2 = D - I$, and all possible output $T \in Range(A)$

$$\Pr[A(D_1) \in T] \leq \exp(\epsilon) \times \Pr[A(D_2) \in T]$$

The probability is taken over the coin tosses of A .

Proof.

For D_1 , the probability density at any $T \in Range(A)$ is proportional to $e^{-\|A(D_1)-T\|_1 \left(\frac{\epsilon}{S(F)} \right)}$. Similarly, for D_2 , the probability density at any $T \in Range(A)$ is proportional to $e^{-\|A(D_2)-T\|_1 \left(\frac{\epsilon}{S(F)} \right)}$. Therefore,

$$\begin{aligned} \frac{\Pr[A(D_1) \in T]}{\Pr[A(D_2) \in T]} &= \frac{e^{-\|A(D_1)-T\|_1 \left(\frac{\epsilon}{S(F)} \right)}}{e^{-\|A(D_2)-T\|_1 \left(\frac{\epsilon}{S(F)} \right)}} = \frac{e^{\|A(D_2)-T\|_1 \left(\frac{\epsilon}{S(F)} \right)}}{e^{\|A(D_1)-T\|_1 \left(\frac{\epsilon}{S(F)} \right)}} \\ &= e^{(\|A(D_2)-T\|_1 - \|A(D_1)-T\|_1) \left(\frac{\epsilon}{S(F)} \right)} \leq e^{\|A(D_2)-A(D_1)\|_1 \left(\frac{\epsilon}{S(F)} \right)} \end{aligned}$$

where the inequality follows from the triangle inequality. By the definition of sensitivity, $S(F) = \text{Max}_{D_1, D_2: |D_1 - D_2|=1} |A(D_1) - A(D_2)|$. So the ratio is bounded by $e^{-\epsilon}$, yields ϵ -Differential Privacy

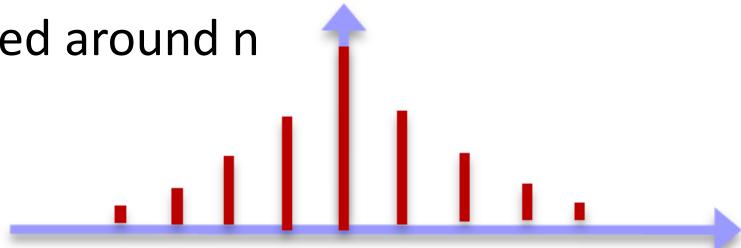


Noise from Geometric Mechanism

- If a function F has only integer values, we can add the discrete counterpart of Laplace noise, geometric noises, to achieve differential privacy, i.e., $F(n) + x = m$, where x is a noise.
- Geometric distribution has the similar properties to those of Laplace distribution
- **Definition (Geometric Distribution):** Let $\alpha > 1$. we denote by $\text{Geom}(\alpha)$ the symmetric geometric distribution that takes integer values such that the probability mass function (PMF) at k is $\frac{\alpha-1}{\alpha+1} \cdot \alpha^{-|k|}$.
- What does this mean?
 - For input n , output distribution is

$$\Pr[F(n) + x = m] = \frac{\alpha - 1}{\alpha + 1} \cdot \alpha^{-|m - F(n)|}$$

- Symmetric geometric distribution, centered around n



Add the Noise After Knowing S(F) ...

- Geometric Mechanism with Sensitivity
 - Release $F(x) + \text{Geom}(e^{-\frac{\epsilon}{S(F)}})$ to obtain ϵ -DP guarantee
 - $F(x)$ = true answer on input x
 - $\text{Geom}(\alpha)$ = noise sampled from Geometric distribution with parameter $\alpha = e^{-\frac{\epsilon}{S(F)}}$
 - Intuition on impact of parameters of differential privacy (DP):
 - Larger $S(F)$, more noise (need more noise to mask an individual)
 - Smaller ϵ , more noise (more noise increases privacy)



Proof of ϵ -Differential Privacy ...

Proof.

For D_1 , the probability density at any $T \in \text{Range}(A)$ is proportional to $\frac{\alpha-1}{\alpha+1} \cdot \alpha^{-|A(D_1)-T|}$ with parameter $\alpha = S(F)/\epsilon$. Similarly, for D_2 , the probability density at any $T \in \text{Range}(A)$ is proportional to $\frac{\alpha-1}{\alpha+1} \cdot \alpha^{-|A(D_2)-T|}$. Therefore,

$$\begin{aligned}\frac{\Pr[A(D_1) \in T]}{\Pr[A(D_2) \in T]} &= \frac{\frac{\alpha-1}{\alpha+1} \cdot \alpha^{-|A(D_1)-T|}}{\frac{\alpha-1}{\alpha+1} \cdot \alpha^{-|A(D_2)-T|}} = \alpha^{|A(D_2)-T|-|A(D_1)-T|} \\ &\leq \alpha^{|A(D_2)-A(D_1)|} = e^{-\frac{\epsilon}{S(F)}|A(D_2)-A(D_1)|} = e^{-\epsilon}\end{aligned}$$

where the inequality follows from the triangle inequality. By the definition of sensitivity, $S(F) = \text{Max}_{D_1, D_2 : |D_1 - D_2|=1} |A(D_1) - A(D_2)|$. So the ratio is bounded by $e^{-\epsilon}$, yields ϵ -Differential Privacy





Sequential Composition

- What happens if we ask multiple questions about same data?
 - We reveal more, so the bound on ϵ differential privacy weakens
- Suppose we output via A_1 and A_2 with ϵ_1 and ϵ_2 differential privacy:
 - $\Pr[A_1(D) = S_1] \leq \exp(\epsilon_1) \Pr[A_1(D') = S_1]$ and $\Pr[A_2(D) =$

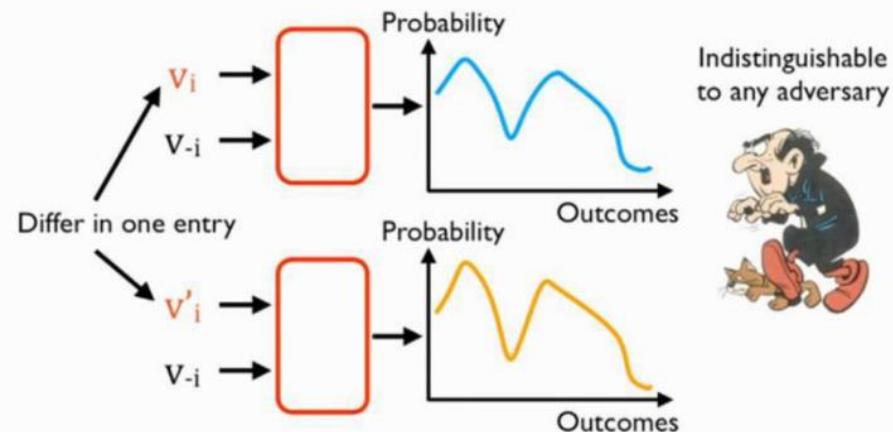
Parallel Composition

- If the inputs are disjoint, then result is $\max(\epsilon_1, \epsilon_2)$ private
- **Example:** Ask for count of people broken down by handedness, hair color
 - Each cell is a disjoint set of individuals
 - So can release each cell with ϵ -differential privacy (parallel composition) instead of 6ϵ -DP (sequential composition)

	Redhead	Blond	Brunette
Left-handed	23	35	56
Right-handed	215	360	493

DP Pros, Cons, and Challenges?

- Utility v.s. privacy
- Privacy budget management and depletion
- Allow non-experts to use?
- Many non-trivial DP algorithms require really large datasets to be *practically* useful
- Advance Mechanisms
 - Sparse Data Processing
 - Multiplicative Weights



Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 2: Big Data Privacy

Lecturer: Rongxing LU

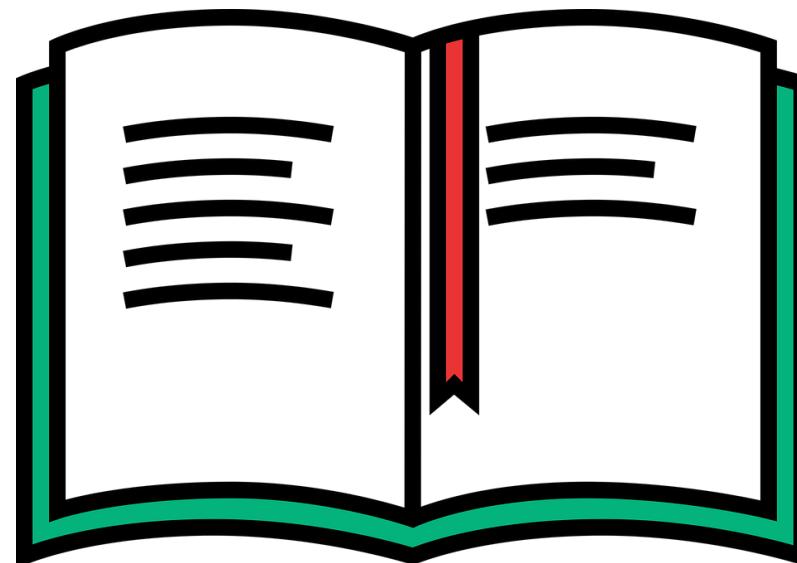
Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

Outline

- Background of Big Data
- Big Data Privacy Techniques



Background of Big Data



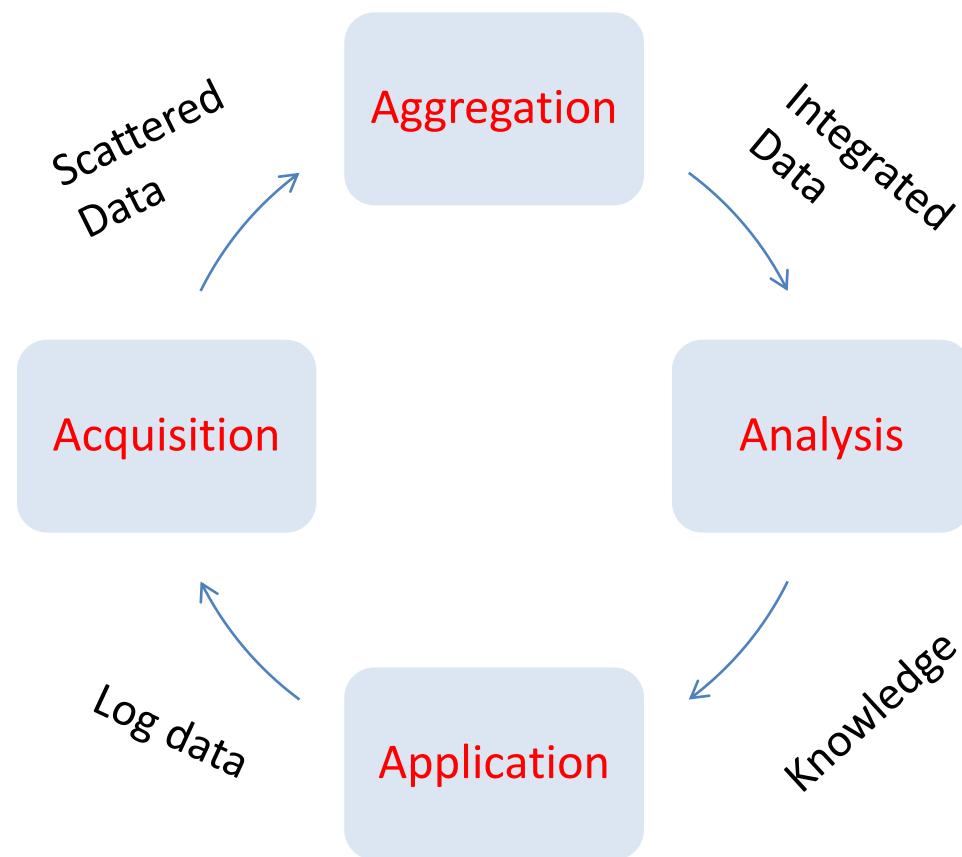
What is “big data”?

- Informally
 - "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)
 - **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)

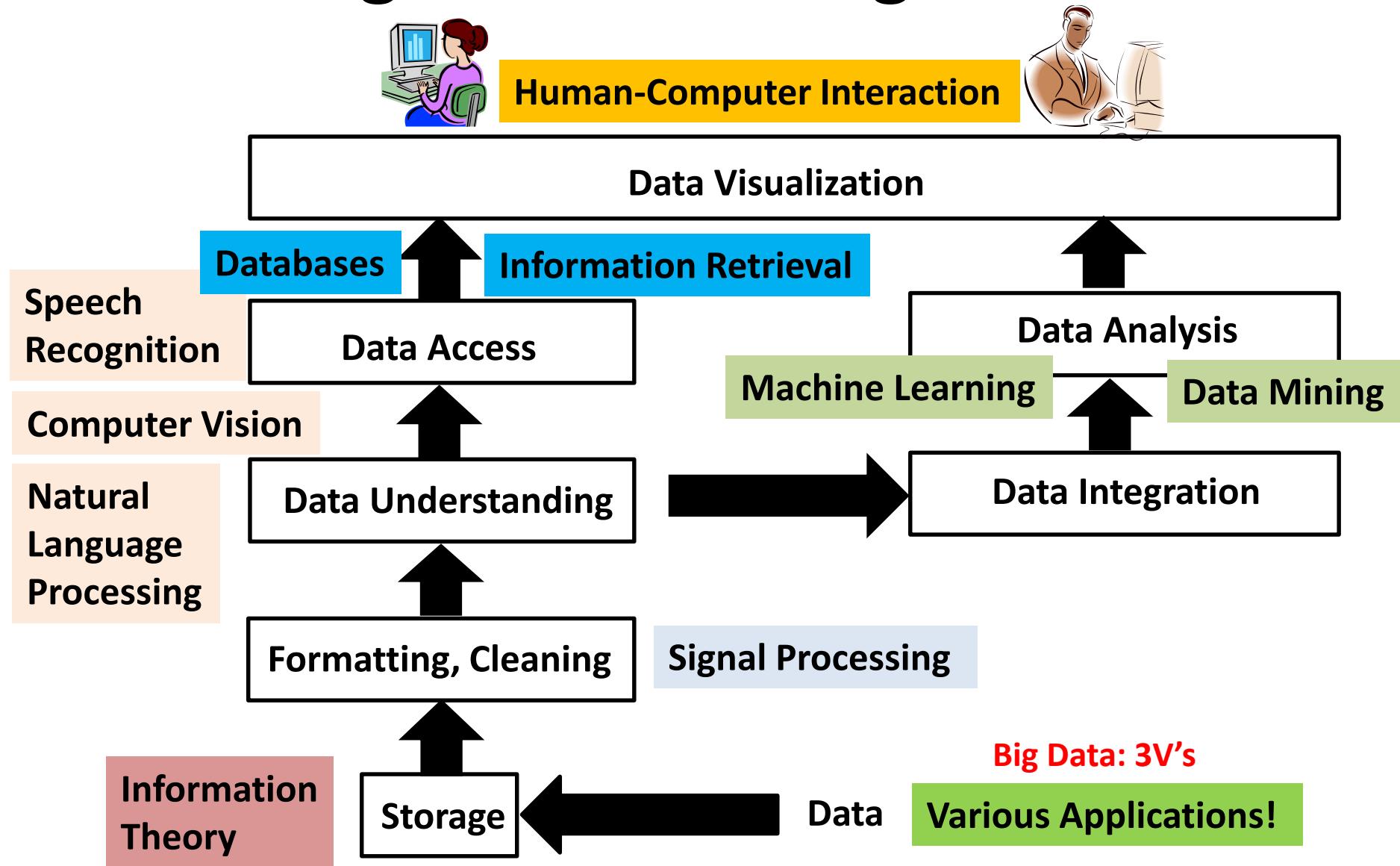
Big Data Challenges & Opportunities

- 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, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

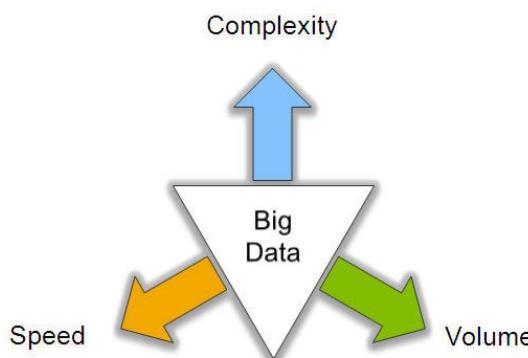
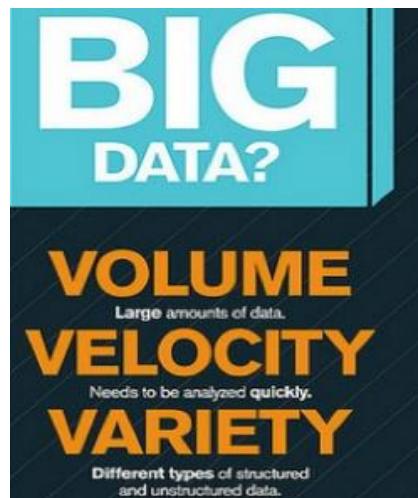
Lifecycle of Big Data: 4 “A”s



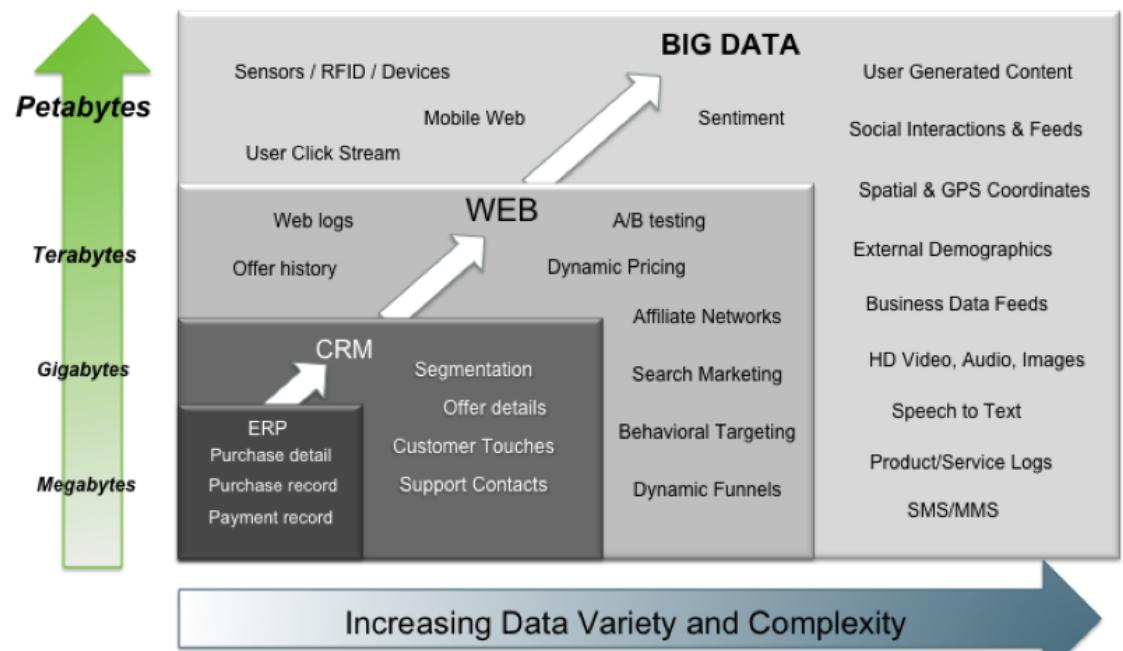
Big Picture of Big Data



Big Data: 3V's



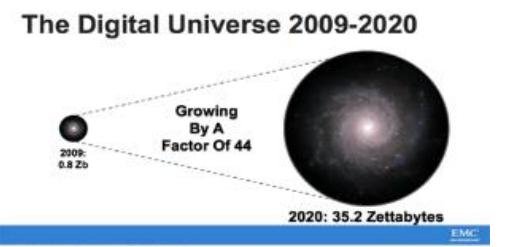
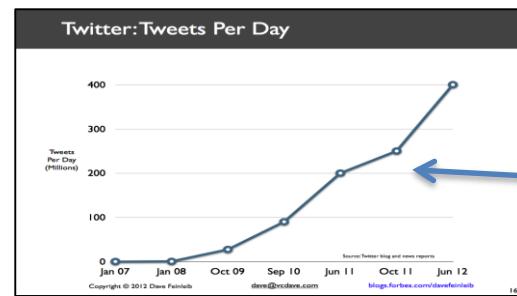
Big Data = Transactions + Interactions + Observations



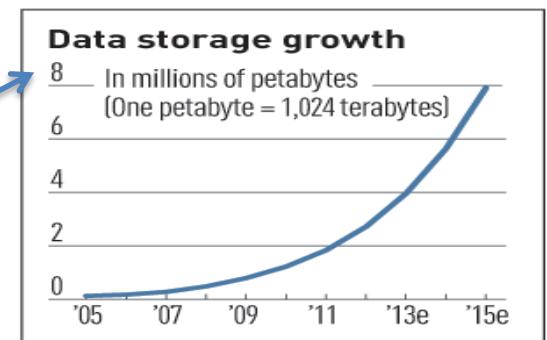
Source: Contents of above graphic created in partnership with Teradata, Inc.

Volume (Scale)

- **Data Volume**
 - 44x increase from 2009 to 2020
 - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially



Exponential increase in collected/generated data



Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data



Volume (Scale) ...

12+ TBs
of tweet data
every day

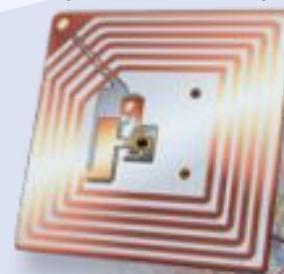


? TBs of
data every day



25+ TBs of
log data
every day

30 billion RFID
tags today
(1.3B in 2005)



76 million smart meters
in 2009...
200M by 2014



4.6
billion
camera
phones
world wide

100s of
millions
of GPS
enabled
devices sold
annually

http://

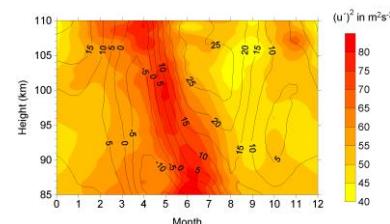
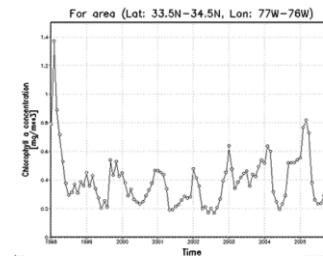
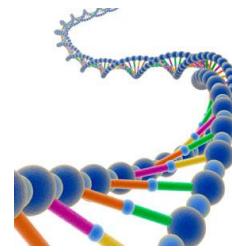
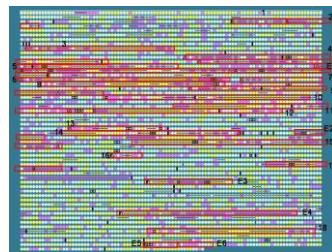
2+
billion
people on
the Web
by end
2011



Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
 - Social Network, Semantic Web (RDF), ...
- Streaming Data
 - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)

To extract knowledge → all these types of data need to linked together



Velocity (Speed)

- Data is generated fast and need to be processed fast
- Online Data Analytics
- Late decisions → missing opportunities
- **Examples**
 - **E-Promotions:** Based on your current location, your purchase history, what you like → send promotions right now for store next to you
 - **Healthcare monitoring:** sensors monitoring your activities and body → any abnormal measurements require immediate reaction



Real-time/Fast Data

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion



Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)

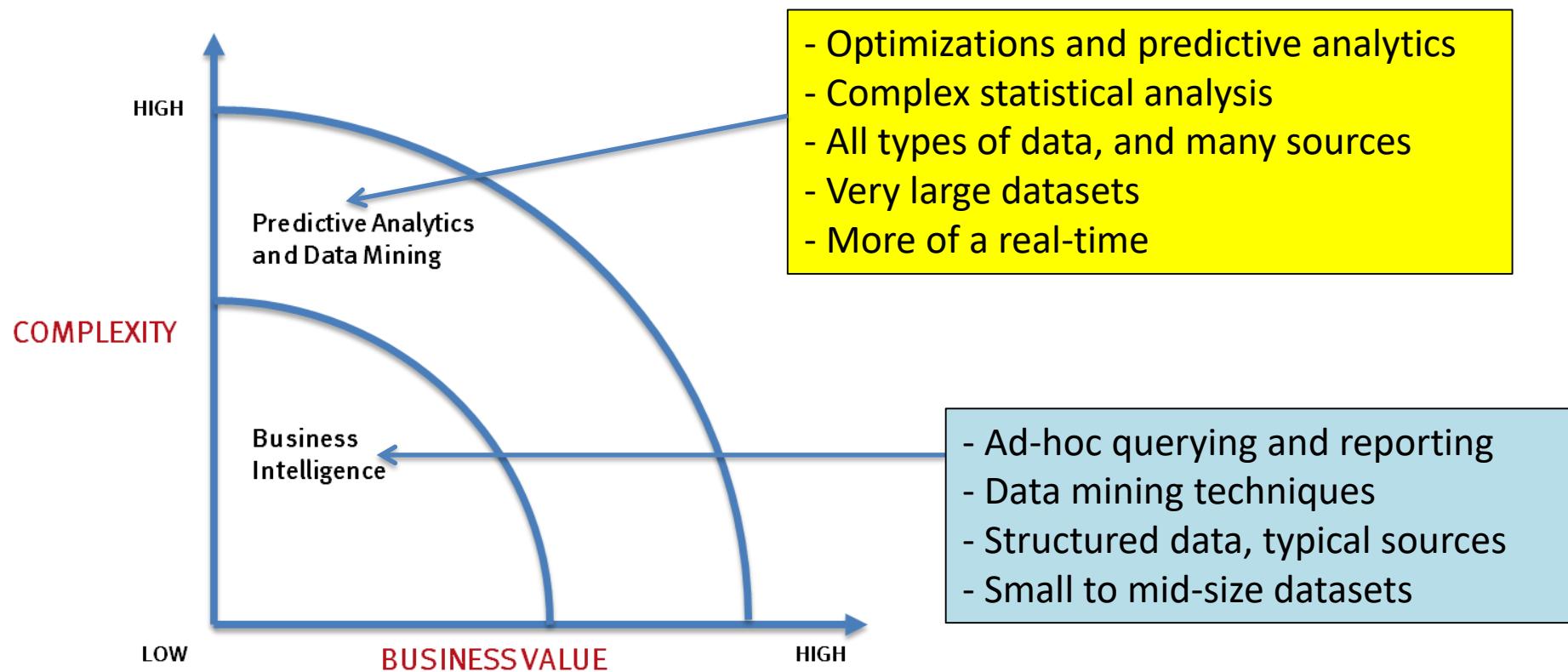


Mobile devices
(tracking all objects all the time)



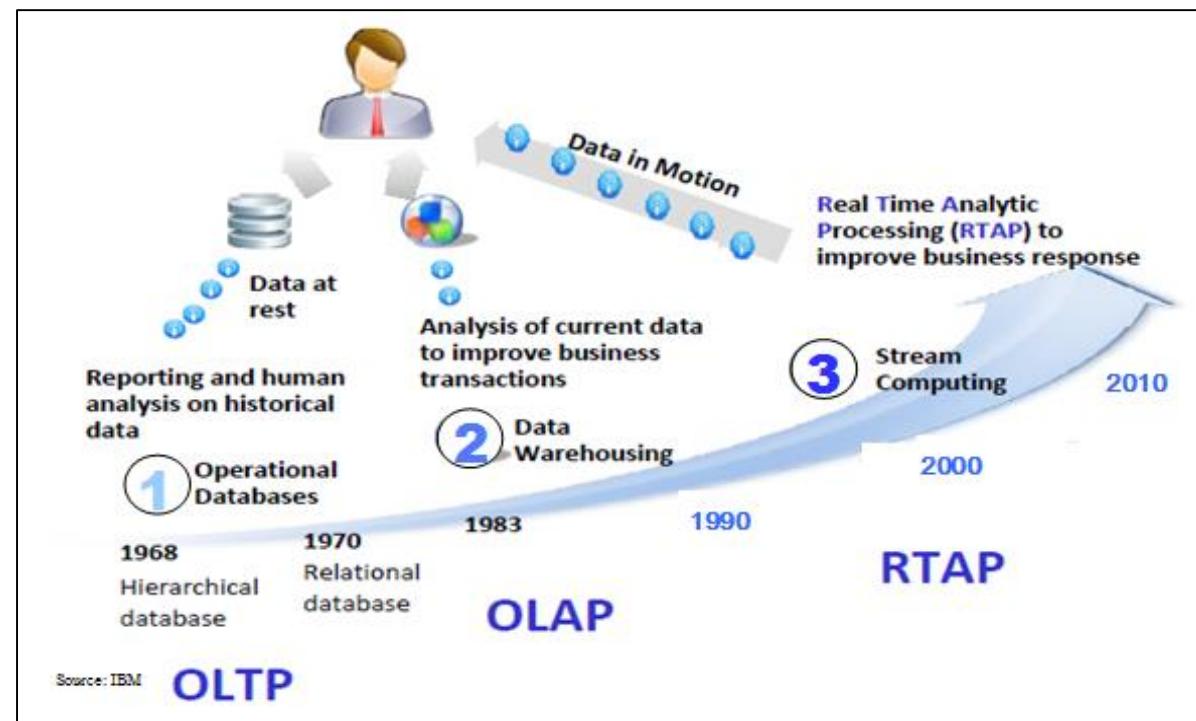
Sensor technology and networks
(measuring all kinds of data)

What's driving Big Data

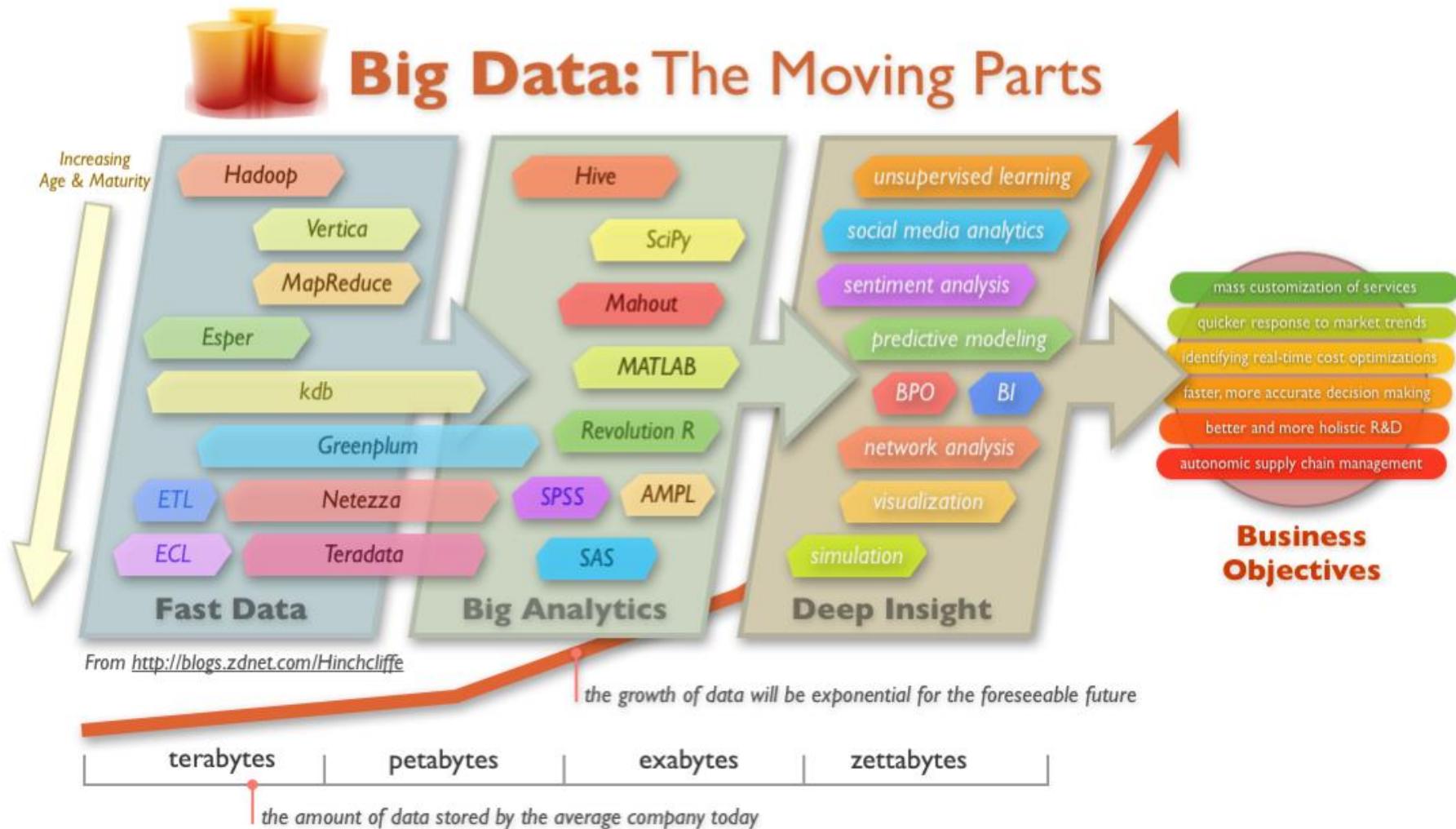


Harnessing Big Data

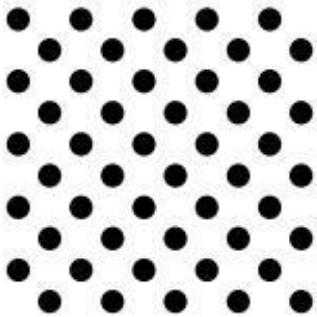
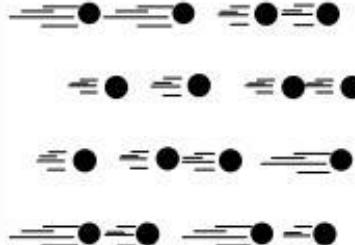
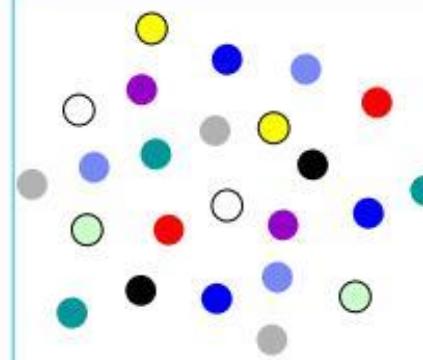
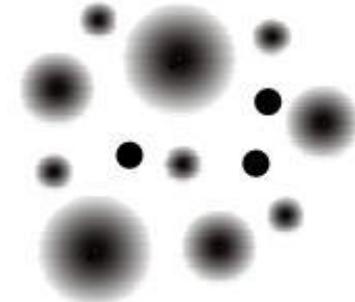
- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)



Big Data Technology

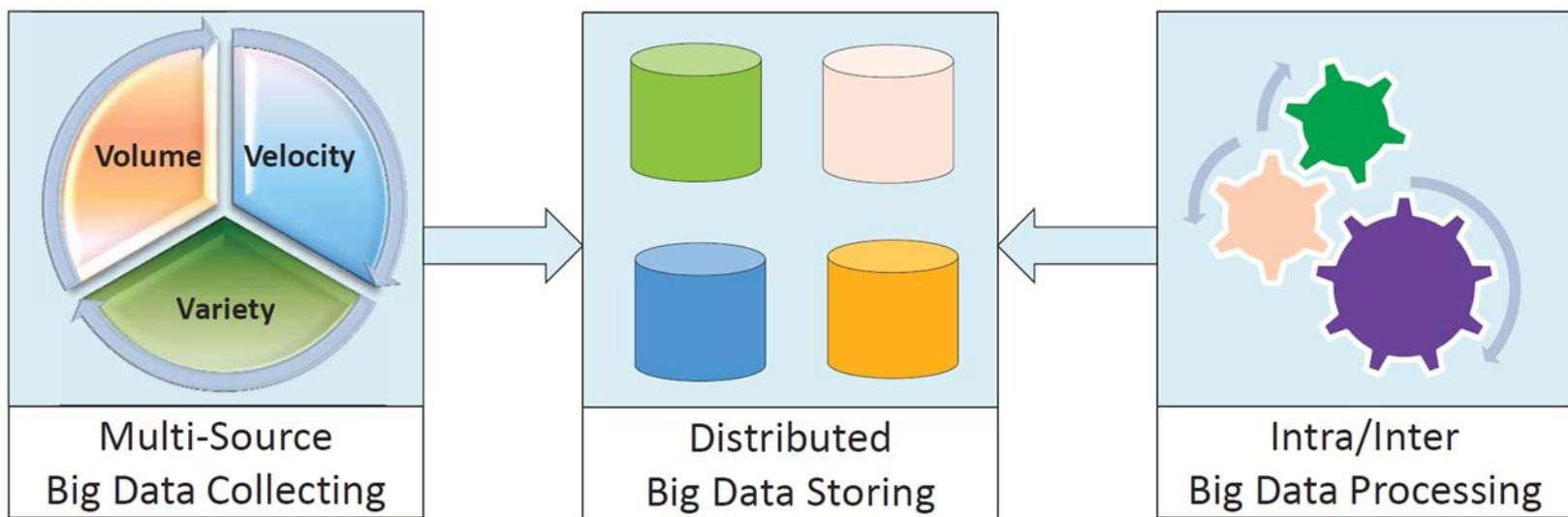


Big Data: 3V's to 4V's

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

Security Issues in Big Data

- Big Data Security is not only referred to that existing in database
 - But includes security issues in whole data cycle



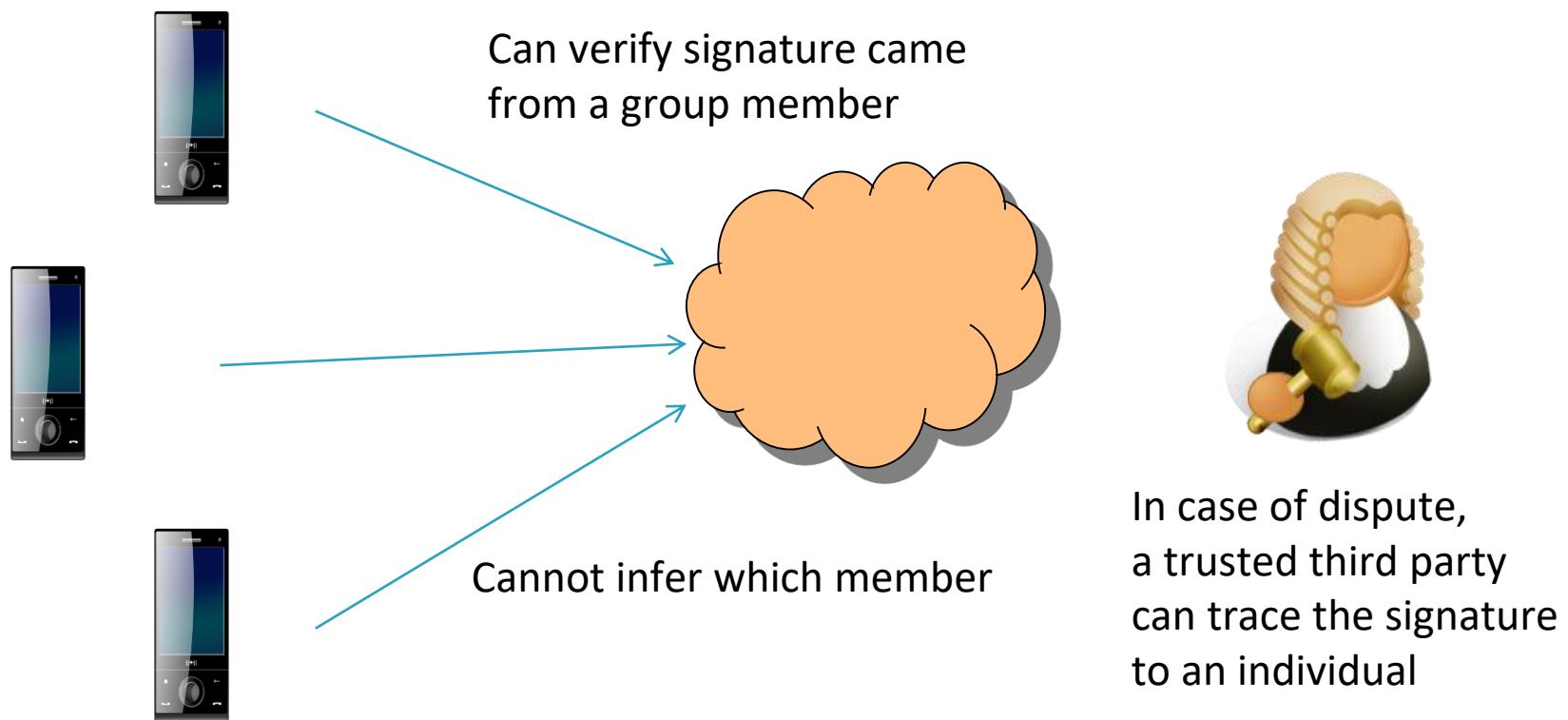
Typical Security Issues Discussed in Big Data

- Secure data collection
- Secure data filtration
- Data Integrity and Poisoning Concerns
- Proof of Data Storage
- Secure Outsourcing of Computation



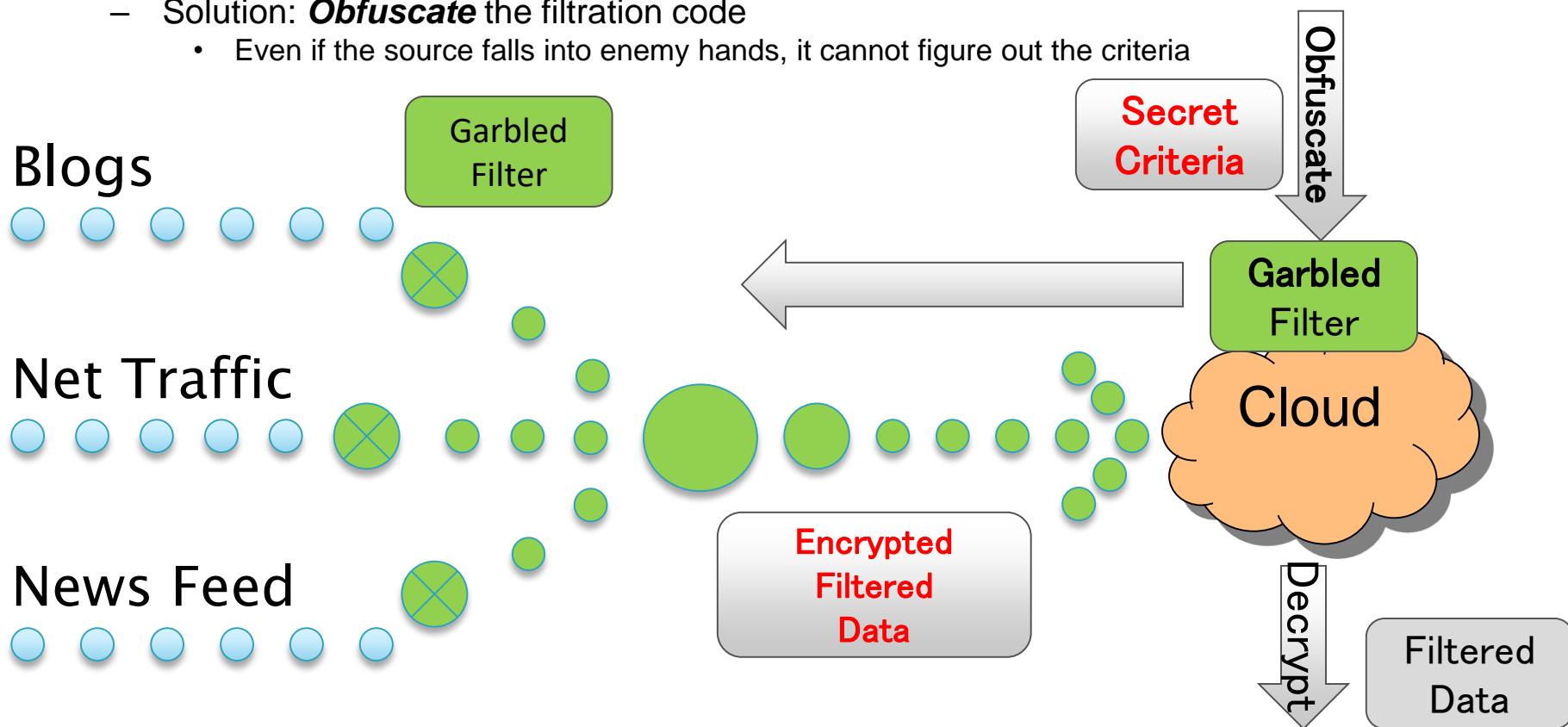
Secure data collection

- How to make collection of data *private* as well as *authenticated*?



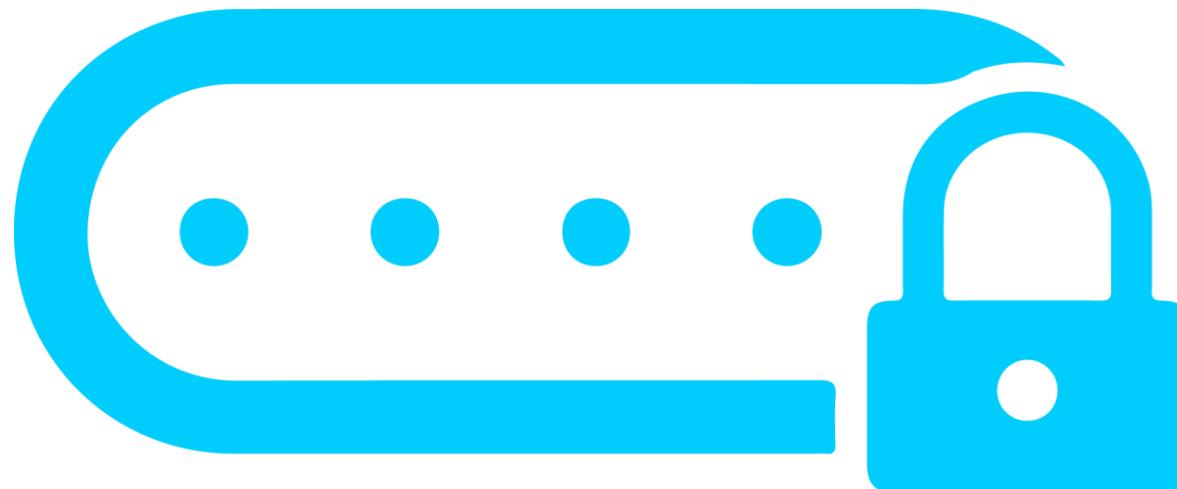
Secure Data Filtration

- Problem Scenario:
 - The intelligence gathering community needs to collect a useful subset of huge streaming sources of data
 - The criteria for being useful may be classified – **private criteria**
 - Most of the streaming data is useless and storing it all may be impractical – **filter at source**
 - How do we keep the filtering criteria secret even if it is executing at the source?
 - Solution: **Obfuscate** the filtration code
 - Even if the source falls into enemy hands, it cannot figure out the criteria



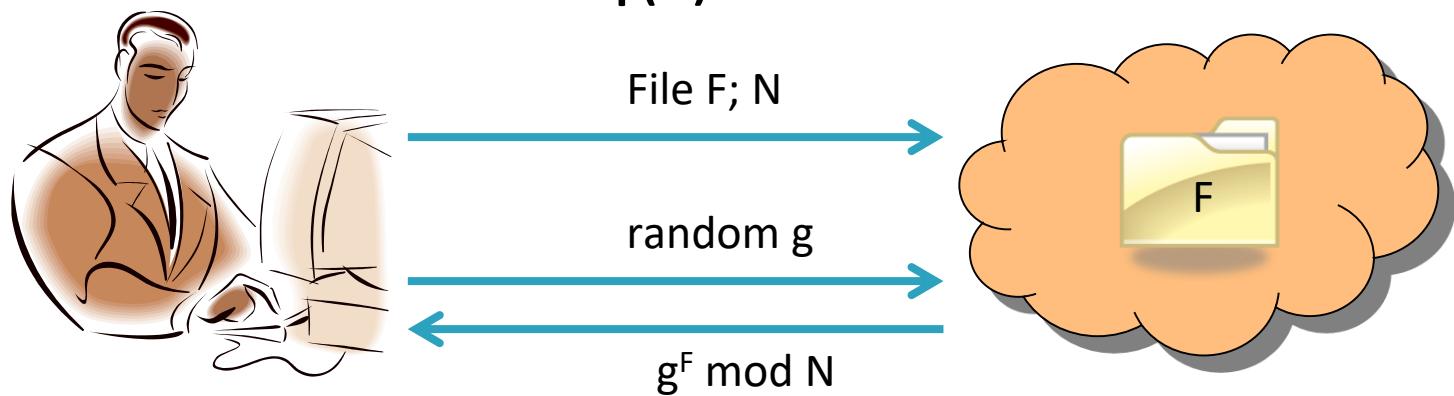
Data Integrity and Poisoning Concerns

- Computing on Authenticated Data



Proof of Data Storage

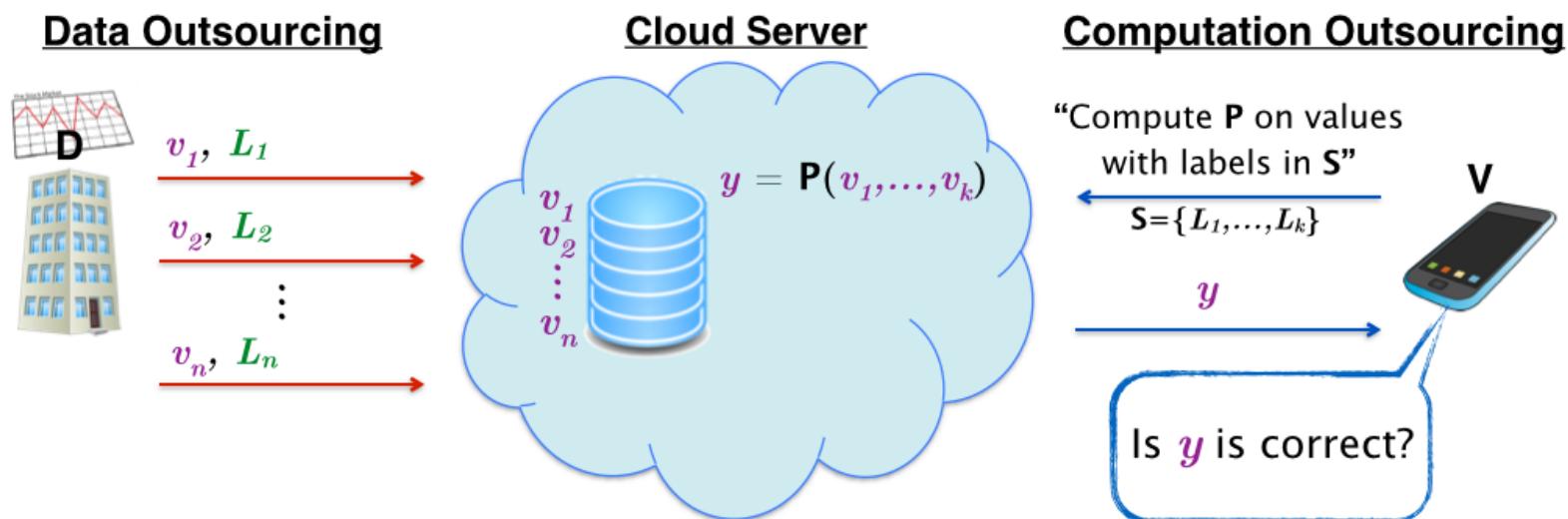
$$N = pq$$
$$f = F \bmod \phi(N)$$



**Check if
 $g^f = g^F \bmod N$**

Secure Outsourcing of Computation

- Problem Scenario:
 - A “weak client” wants to outsource a computation
 - The provider returns the result along with a “proof” that the computation was carried out correctly
 - Catch: verification of the proof should require substantially less computational effort than computing the result from scratch



Big Data Privacy Techniques



Why are Privacy Issues in Big Data?

- Users have an inadequate understanding of how privacy violations **of big data** impact individuals as well as social behavior.
- There is a lack of transparency regarding **privacy policies** or predictive analytics applied to users, and there is a lack of data due process of **law**.
- There is an **economic incentive** to disclose users' data.
- There are **technical limitations** to some of the most advanced techniques devised to allow analytics on private data.
- ...

Privacy Protection and Big Data

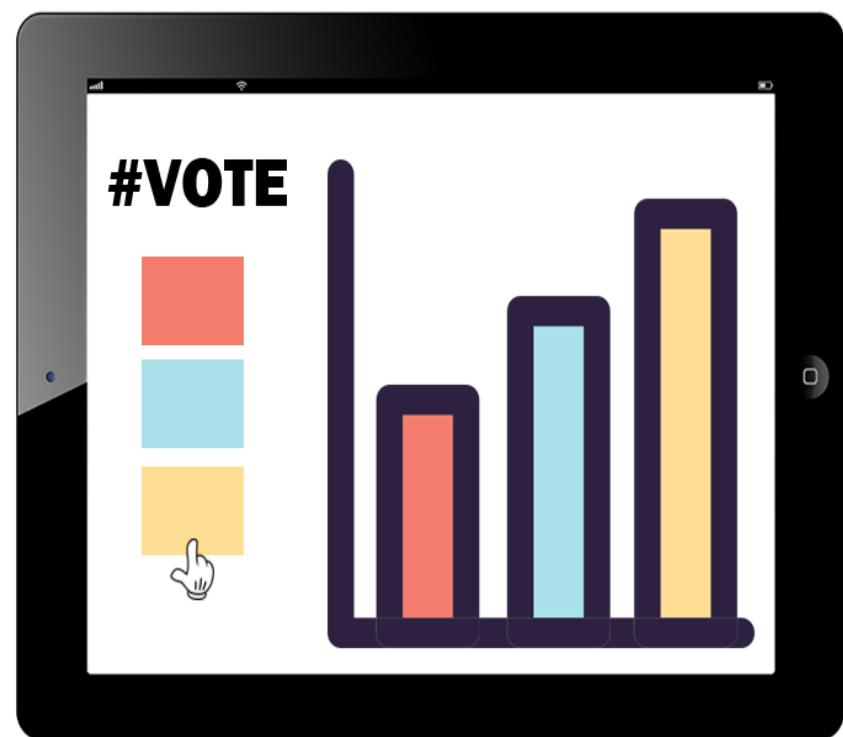
- Big Data analytics applies statistical learning on private data that “prosumers” produce, and uses this learning for purposes of prediction. Natural tensions between learning and privacy.
- There exist a variety of technical measures to protect privacy but not all have the same level of effectiveness or address the same set of problems.

Categories of Privacy-Preserving Techniques

(1) Intentional limitations	(2) Perturbations or transformations	(3) Process guarantees	(4) Ownership guarantees
<ul style="list-style-type: none">• Analytics intentionally limited• Polls• aggregation (as used in census)• etc.	<ul style="list-style-type: none">• Transformations selected to preserve as much of the statistics as possible• suppression,• swapping,• randomization,• synthesis,• k-anonymity• Differential privacy,• etc.	<ul style="list-style-type: none">• Reliance on voluntary or regulatory safeguards honored by service providers and data “custodians”• audit logs,• Accountability systems,• etc.	<ul style="list-style-type: none">• Strict ownership and control by prosumers of data• multi-party Secure computation,• etc.

Intentional limitations: Polls

- A special case of Aggregation



Intentional limitations:

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.

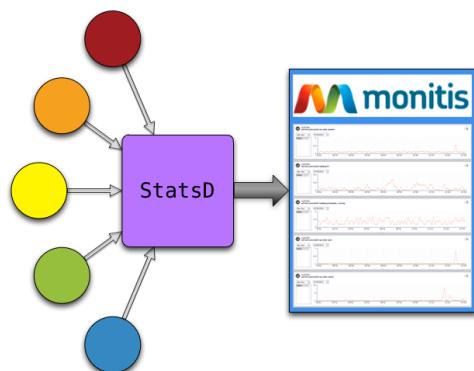


Table 1.1: Definition of Aggregates

Aggregate	Definition
Count	$N = \{v_i\} $, where $\{v_i\}$ is the set of sensed values.
Sum	$S = \sum_{i=1}^N v_i$
Uniform Sample	$U = \{v_{i_1}, v_{i_2}, \dots, v_{i_k}\}$, where these k values are randomly selected from the population set $\{v_i\}$.
Median	$M = \begin{cases} v_i & \text{s.t. } rank(v_i) = \frac{N+1}{2}, \\ \frac{v_i+v_j}{2} & \text{s.t. } rank(v_i) = \frac{N}{2} \text{ and } rank(v_j) = \frac{N}{2} + 1, \end{cases} \quad \begin{matrix} N \text{ is odd} \\ \text{otherwise} \end{matrix}$

Perturbations or transformations: 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.

People by Race and Age Group						
Race	People Reported					
	Age Group					
Race	<18	19-64	65-99	100+	Total	
Martian	15	*	*	0	34	
Asian	*	11	0	*	35	
Black	17	*	*	*	47	
Hispanic	*	*	24	19	56	
White	18	13	19	20	70	
Total	73	48	56	65	242	

* The data is suppressed to protect privacy

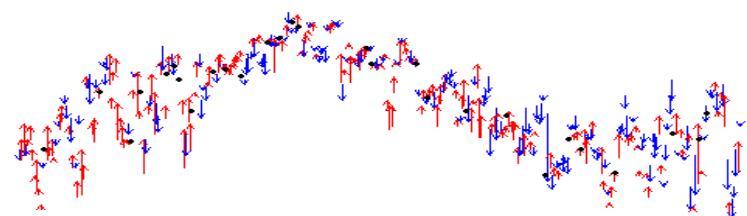
Perturbations or transformations: swapping

- 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.

RecID	Age	State	Diagnosis	Income	Billing
1	44	MI	AIDS	48,000	1,200
2	44	MI	Asthma	37,900	2,500
3	55	MI	AIDS	67,000	3,000
4	44	MI	Asthma	21,000	1,000
5	55	MI	Asthma	90,000	900
6	45	MI	Diabetes	45,500	750
7	25	IN	Diabetes	49,000	1,200
8	35	MI	AIDS	66,000	2,200
9	55	MI	AIDS	69,000	4,200
10	45	MI	Tuberculosis	34,000	3,100

Perturbations or transformations: randomization

- This technique involves adding noise of randomly generated numerical values to data variables to distort the values of sensitive variables and make it difficult to deduce accurate matching.
- The level of privacy protection depends on the nature of the noise distribution. Greater protection is achieved when using a noise distribution with large variance. However, large-variance noise distribution may introduce measurement errors and inaccurate regression coefficients.
- Differential Privacy Techniques



Perturbations or transformations: synthesis

- With synthetic data, the values of sensitive variables are replaced with synthetic values generated by simulation.
- In a way, the synthetic values are basically a random value generated by a probability distribution function simulator.
- These distributions are selected to reproduce as many of the relationships in the original data as possible.



The Simulation

Process guarantees:

Audit logs

- An audit log is a document that records an event in an information (IT) technology system. In addition to documenting what resources were accessed, audit log entries usually include destination and source addresses, a timestamp and user login information.

Data Audit Log

Date (UTC)	Type	Name	Chan...	User IP Address	User	User Type	Chan...
2014-09-14 02:...	Preferences	(name: po_file_hi...	Insert	207.66.184.66	Demo Admin (ID:...	Employee	2
2014-09-12 19:...	Advertisers	Another Avertise...	Update	207.66.184.66	Demo Admin (ID:...	Employee	2
2014-09-09 18:...	Offers	Example Offer (i...	Update	207.66.184.66	Demo Admin (ID:...	Employee	2
2014-09-08 23:...	Preferences	(name: affiliate_...	Delete	54.244.19.51	Demo Admin (ID:...	Employee	2
2014-09-08 23:...	Preferences	(name: affiliate_...	Delete	54.244.19.51	Demo Admin (ID:...	Employee	2
2014-09-08 23:...	Preferences	(name: affiliate_...	Delete	54.244.19.51	Demo Admin (ID:...	Employee	2
2014-09-08 23:...	Preferences	(name: affiliate_...	Delete	54.244.19.51	Demo Admin (ID:...	Employee	2
2014-09-08 23:...	Preferences	(name: affiliate_...	Insert	54.244.19.51	Demo Admin (ID:...	Employee	2
2014-09-08 23:...	Preferences	(name: affiliate_...	Insert	54.244.19.51	Demo Admin (ID:...	Employee	2

Total Items: 654

Page Size: 10

Navigation icons: back, forward, first, last, search, etc.

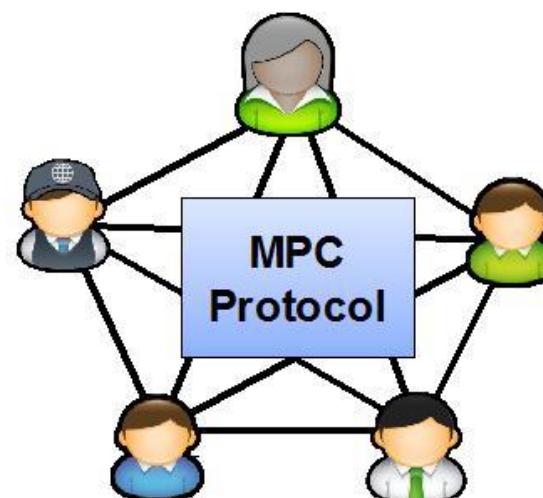
Process guarantees: Accountability systems

- An accountability system establishes the processes for monitoring, analyzing, and improving the performance of individuals and institutions, and as such, it is a key mechanism for achieving good governance outcomes.



Ownership guarantees: Multi-party Secure Computation (MPSC)

- **Secure function evaluation** allows a set $P = \{p_1, \dots, p_n\}$ of n players to compute an arbitrary agreed function of their private inputs (x_1, \dots, x_n) , respectively, even if an adversary may corrupt and control some of the players in various ways.
- More 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.



Example: An Efficient and Privacy-Preserving Cosine Similarity Computing Protocol

 P_A

$$\vec{a} = (a_1, a_2, \dots, a_n) \in F_q^n$$

 P_B

$$\vec{b} = (b_1, b_2, \dots, b_n) \in F_q^n$$

Keep Privacy Step1: (performed by P_A) Given security parameters k_1, k_2, k_3, k_4 , choose two large primes α, p such that $|p| = k_1$, $|\alpha| = k_2$, set $a_{n+1} = a_{n+2} = 0$

Choose a large random number $s \in Z_p$, and $n + 2$ random numbers $c_i, i = 1, 2, \dots, n + 2$, with $|c_i| = k_3$ for each $a_i, i = 1, 2, \dots, n + 2$

$$C_i = \begin{cases} s(a_i \cdot a + c_i) \bmod p, & a_i \neq 0 \\ s \cdot c_i \bmod p, & a_i = 0 \end{cases}$$

end for

compute $A = \sum_{i=1}^n a_i^2$, keep $s^{-1} \bmod p$ secret, and send $(\alpha, p, C_1, \dots, C_{n+2})$ to P_B

$$\alpha, p, C_1, \dots, C_n$$

Step2: (performed by P_B) set $b_{n+1} = b_{n+2} = 0$

for each $b_i, i = 1, 2, \dots, n+2$

$$D_i = \begin{cases} b_i \cdot \alpha \cdot C_i \bmod p, & b_i \neq 0 \\ r_i \cdot C_i \bmod p, & b_i = 0 \end{cases}$$

where r_i is a random number, with $|r_i| = k_4$

end for

 $B = \sum_{i=1}^n b_i^2$ and $D = \sum_{i=1}^{n+2} D_i \bmod p$, send (B, D) back to P_A B, D Step3: (performed by P_A) compute $E = s^{-1} \cdot D \bmod p$

$$\text{compute } \vec{a} \cdot \vec{b} = \sum_{i=1}^n a_i \cdot b_i = \frac{E - (E \bmod \alpha^2)}{\alpha^2}, \cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\sqrt{A} \cdot \sqrt{B}}$$

Output correctly

Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 3: Review of some basic cryptographic
techniques

Lecturer: Rongxing LU

Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

Review of Some Basic Cryptographic Techniques

Symmetric Encryption

Block Cipher

Stream Cipher

Hash Function

Hash Function Properties

Birthday Attack

Bloom Filter

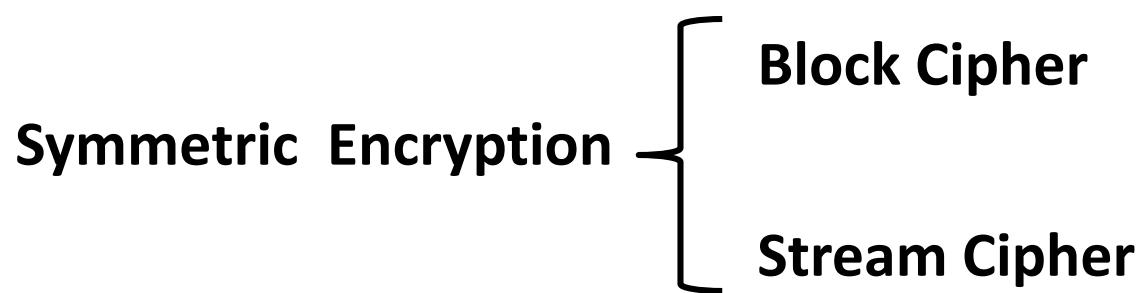
Public Key Encryption

Number Theory

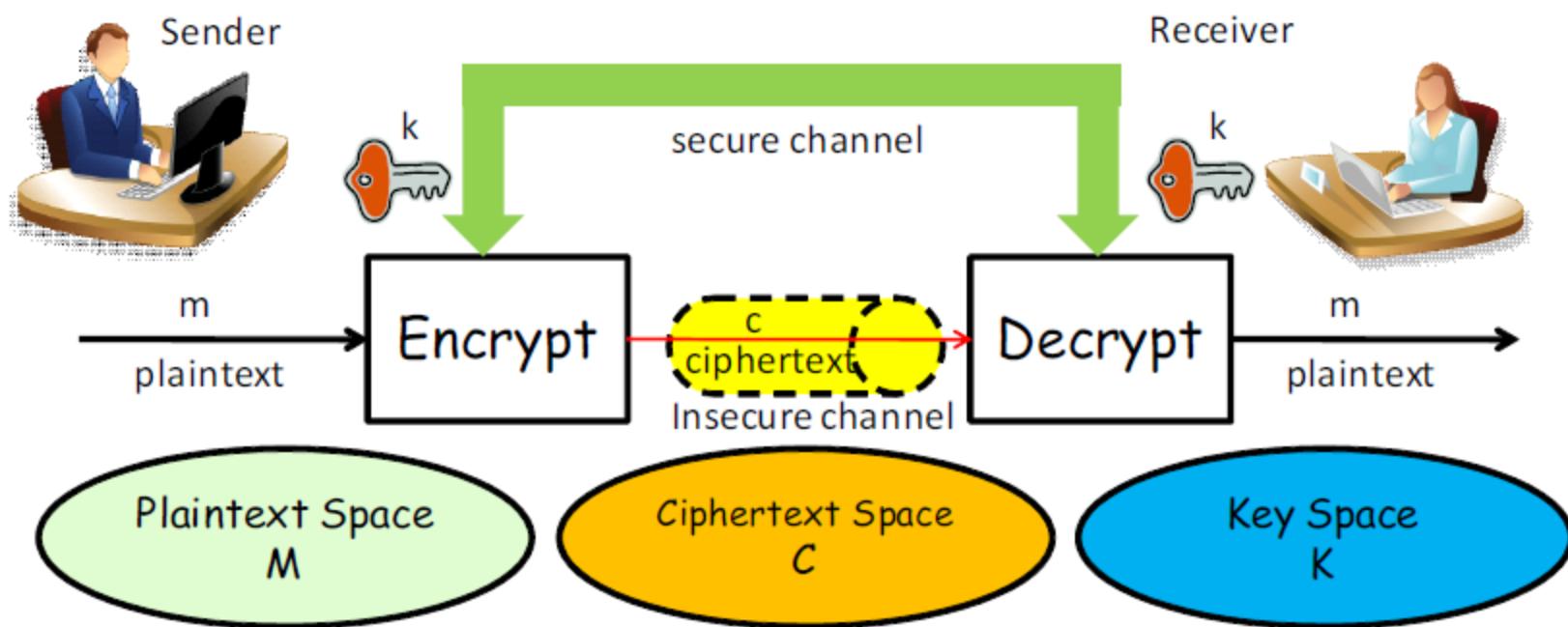
RSA Encryption, Signature

ElGamal Encryption

Diffie-Hellman Key Exchange

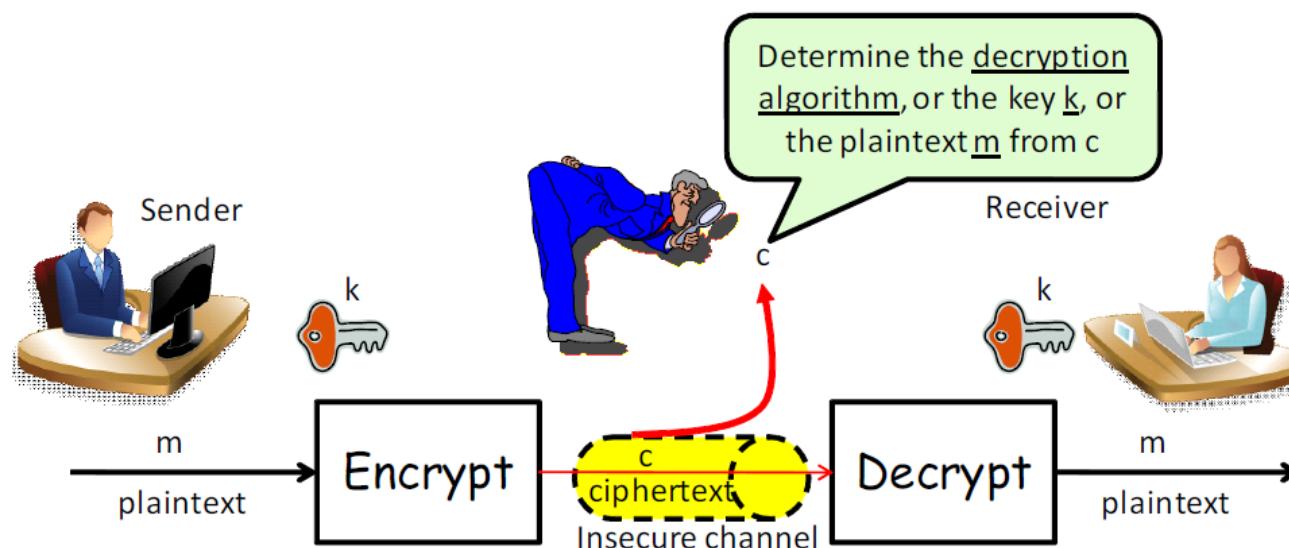


Symmetric Encryption



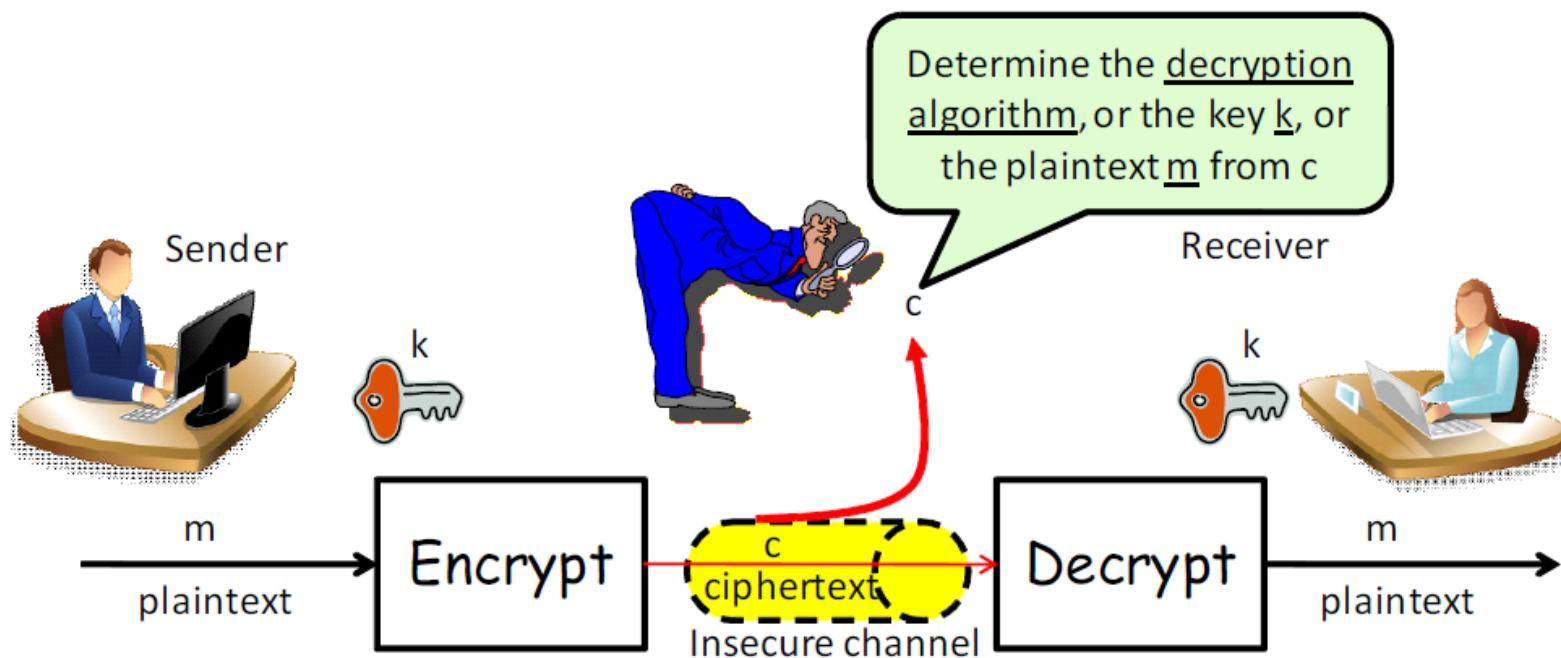
Attacks on Symmetric Encryption

- **Ciphertext-only attack:** An adversary determines the decryption algorithm Dec_k or key k , or the plaintext from intercepted ciphertext c .



Attacks on Symmetric Encryption (2)

- **Known-plaintext attack:** An adversary determines the decryption algorithm Dec_k or key k , from a ciphertext-plaintext (c, m) .



Security Requirements

- According to **Kerckhoffs** Principles, the security should depend on the confidentiality of the key, so it is usually assumed that the algorithms Enc_k and Dec_k are known to an adversary.
- It should be computationally infeasible for an adversary to determine the plaintext $m \in M$, given a ciphertext $c \in C$.
- It should be computationally infeasible for an adversary to systematically determine the decryption algorithm Dec_k or key k from intercepted ciphertext c , even if the corresponding plaintext m is known.

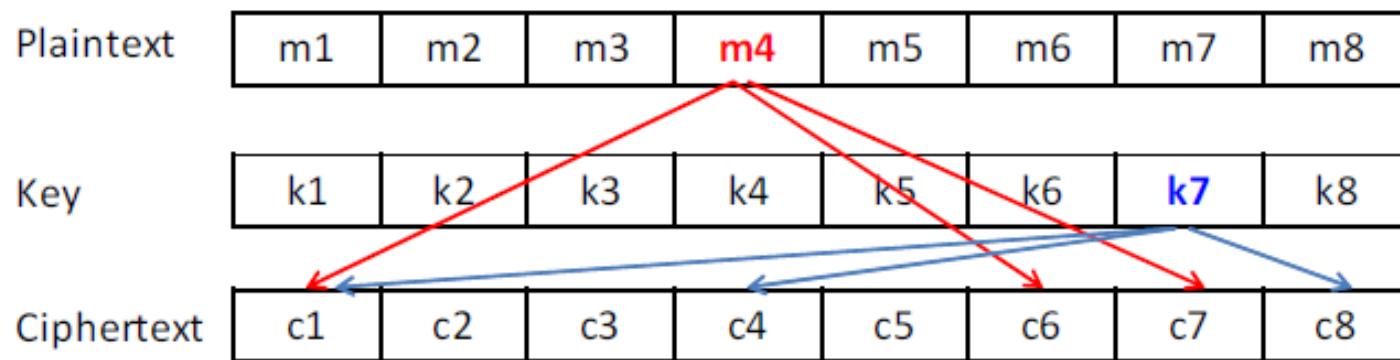
How to design a symmetric encryption to meet the above requirement?

- Two properties are desirable
 - Confusion
 - Diffusion
- **Confusion:** Process of substituting characters or symbols to make the relationship between ciphertext and key as complex as possible.
Attackers's uncertainly as to the contents of a message or the key used for encryption and decryption.

How to design a symmetric encryption to meet the above requirement ? (2)

Plaintext	1	1	0	1	1	0	1	1	0	1
Key	0	1	0	1	1	1	1	0	1	1
Ciphertext = Plaintext \oplus Key	1	0	0	0	0	1	0	1	1	0

- **Diffusion:** Process of spreading effect of plaintext or key as widely as possible over ciphertext.
Dispersion of the effect of individual key or message bits over the plaintext



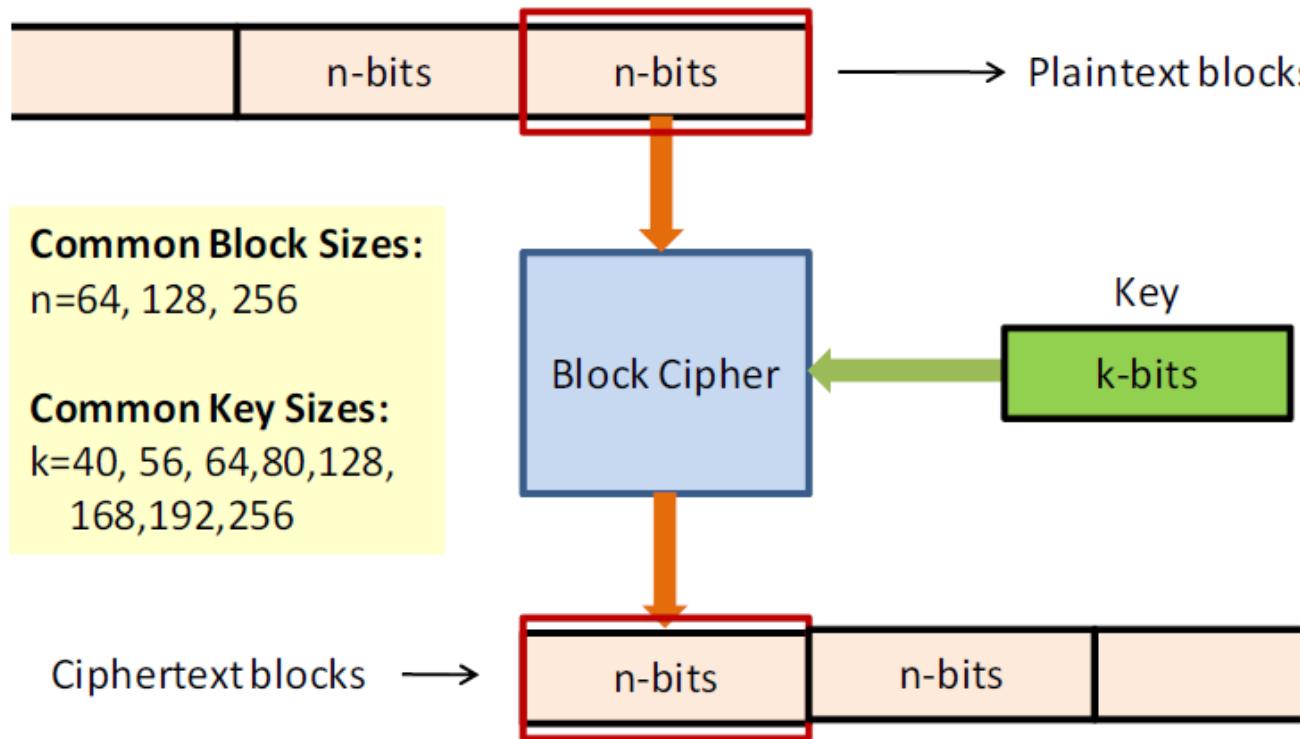
Type of Ciphers



- Block ciphers
 - Block ciphers break messages into fixed length blocks, and encrypt each block using the same key.
 - The Data Encryption Standard (DES) is an example of a block cipher, where blocks of 64 bits are encrypted using a 56-bit key.
- Stream ciphers
 - Stream ciphers, like block ciphers, break message into fixed length blocks, but use a sequence of keys to encrypt the blocks.
 - The Vigenere cipher is an example of a stream cipher.
 - Key = $k_1 k_2 k_3 k_4$ (random, used one-time only)
 - Plaintext = $m_1 m_2 m_3 m_4$; Ciphertext = $c_1 c_2 c_3 c_4$, where $c_i = m_i \oplus k_i$

Block Cipher

- Message is divided into fixed size blocks (block size) using padding if necessary
- Ciphertext is block of (usually) the same size



Block Cipher

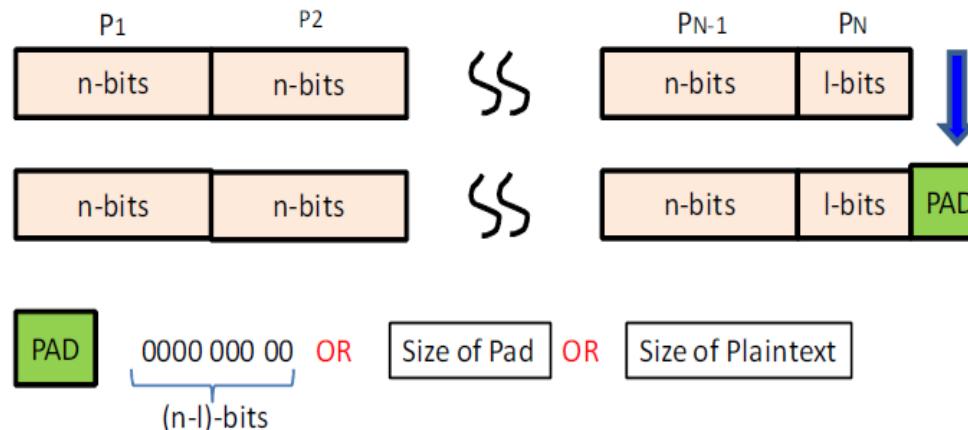
- Formal definition of Block Cipher
 - Let E be an encryption algorithm, and let $Ek(b)$ be the ciphertext of the message b with key k .
 - a message $m = b_1 b_2 \dots$ where each b_i is of a fixed length.
 - A block cipher is a cipher for which $Ek(m) = Ek(b_1)Ek(b_2)\dots$
- Properties of Block Ciphers
 - Adds Confusion about message and key
 - Should have good Diffusion Properties
 - Single bit change in input plaintext should produce changes in approx. 50% of output bits (at random)
 - Single bit change in key should produce changes in approx. 50% of output bits (at random)

How to use a block cipher

- Block ciphers encrypt fixed-size blocks
 - e.g. DES encrypts 64-bit blocks
- We need some way to encrypt a message of arbitrary length
 - e.g. a message of 1000 bytes
- NIST (National Institute of Standards and Technology) defines several ways to deal with it, called modes of operation
 - ECB (Electronic Code Book)
 - CBC (Cipher Block Chaining)

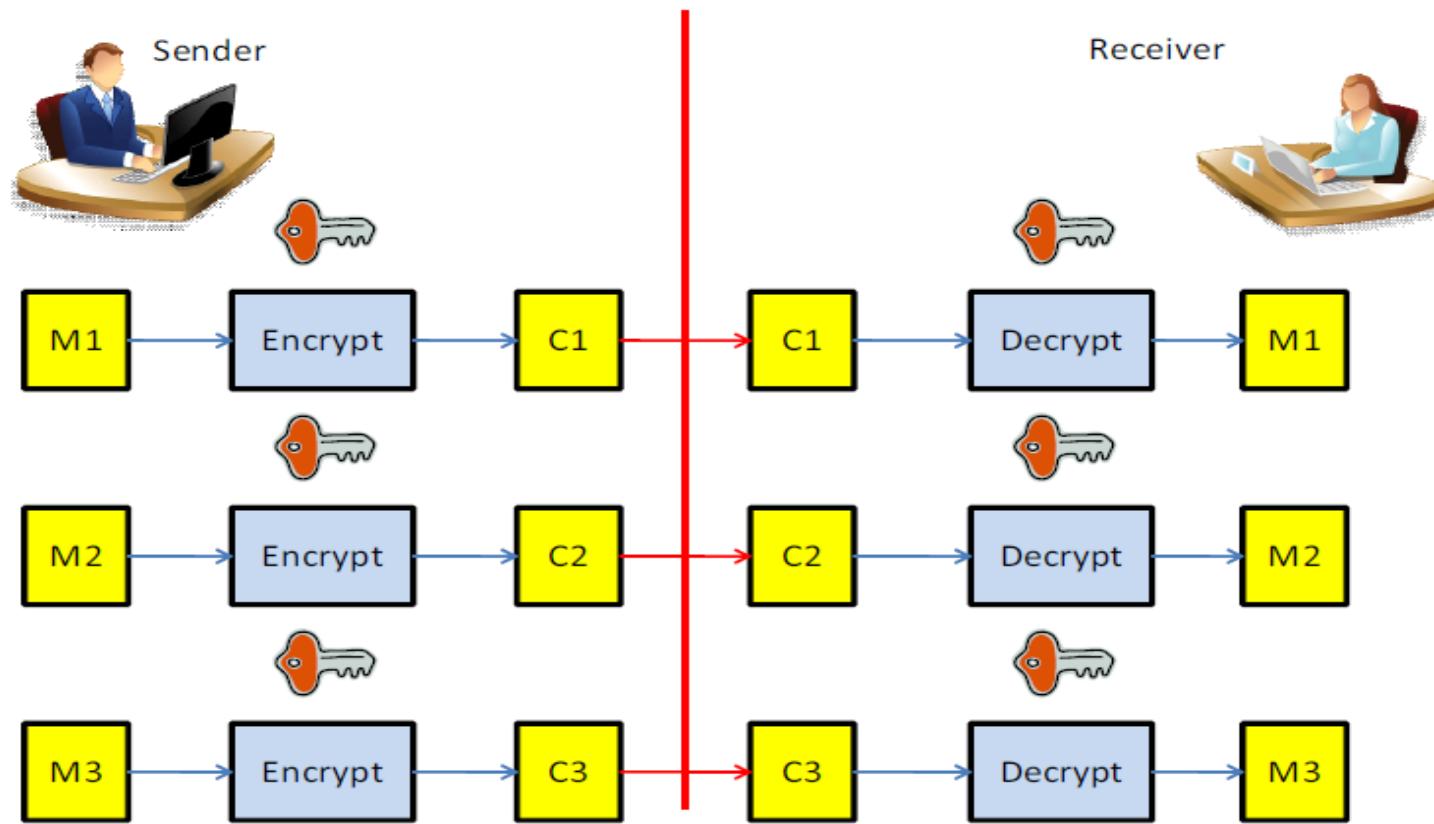
Message Padding

- The plaintext message is broken into blocks, $P_1; P_2; P_3;$
- The last block may be short of a whole block and needs padding.
- Possible padding:
 - Known non-data values (e.g. nulls)
 - Or a number indicating the size of the pad
 - Or a number indicating the size of the plaintext
 - The last two may require an extra block.



Electronic Code Book (ECB) Mode

- Cipher acts as simple block substitution determined by key

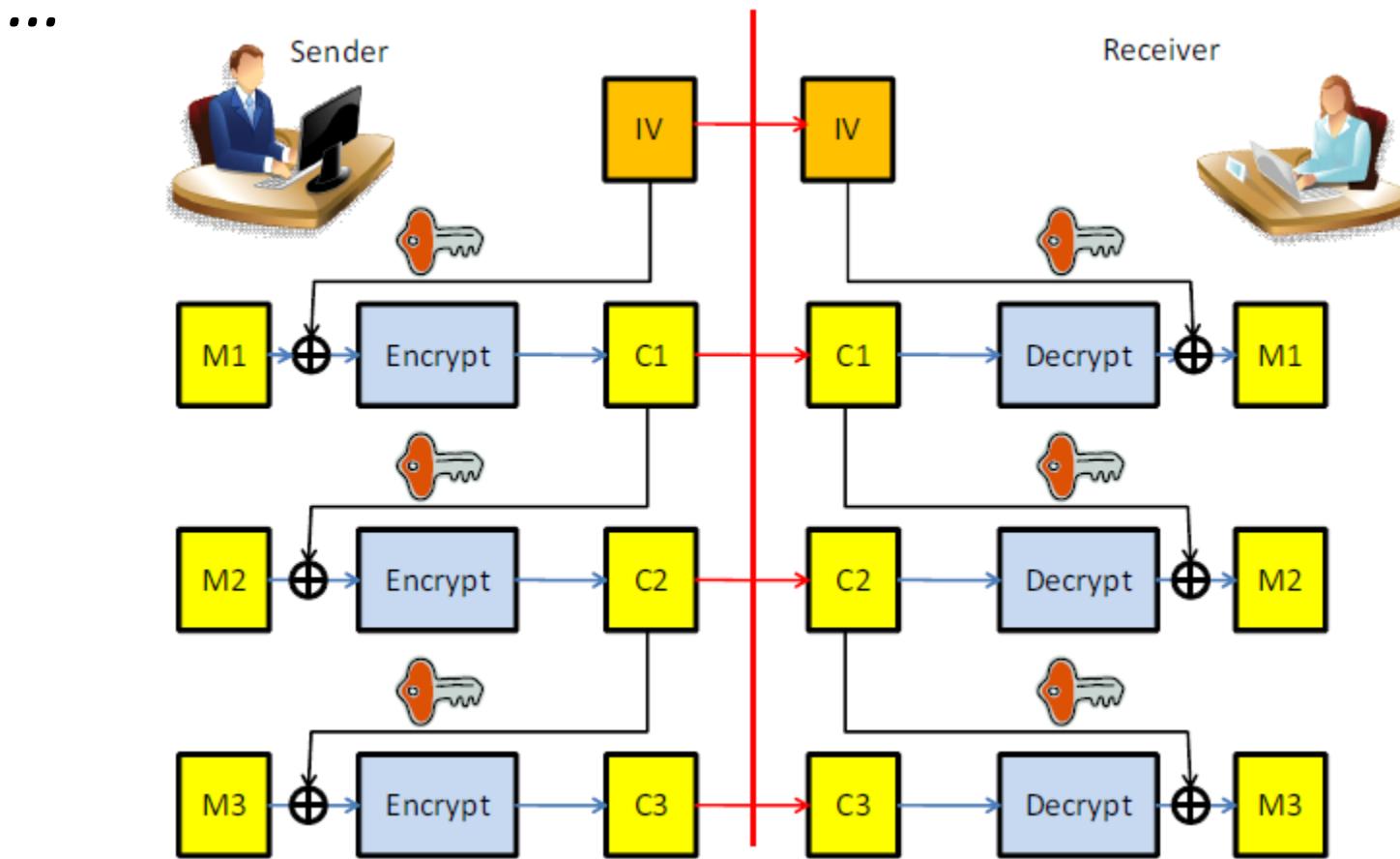


Electronic Code Book (ECB) Mode (2)

- For a given key, this mode behaves like we have a gigantic codebook, in which each plaintext block has an entry, hence the name Electronic Code Book
- Fast and simple but repeated input block creates repeated ciphertext block
- Vulnerable to replay attacks: if an attacker thinks block C_2 corresponds to \$ amount, then substitute another C_k
- Attacker can also build a codebook of
 $\langle C_k; \text{guessed } M_k \rangle$ pairs
- Application: secure transmission of short pieces of information (e.g. a temporary encryption key)

Cipher Block Chaining (CBC) Mode

- The plaintext is broken into blocks $M_1; M_2; M_3;$

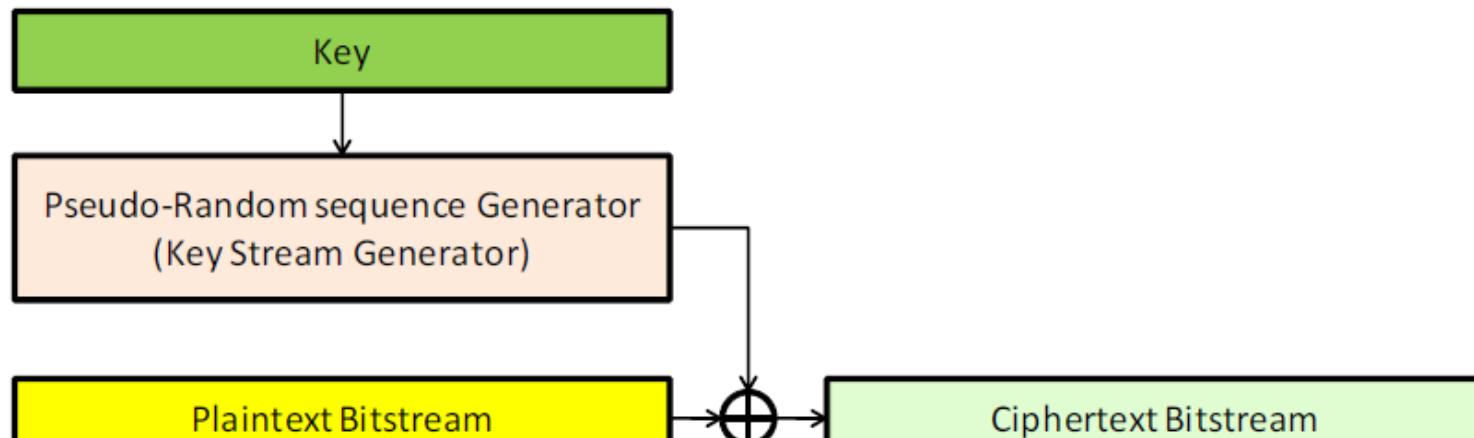


Cipher Block Chaining (CBC) Mode (2)

- Each plaintext block is XORed (chained) with the previous ciphertext block before encryption (hence the name CBC) $C_i = E_k(C_{i-1} \oplus M_i); C_0 = IV$
- The encryption of a block depends on the current and all blocks before it. . Then, the input plaintext $M_i = M_k$ will not result in the same output code due to memory-based chaining
- Use an Initial Vector (IV) (Use only once) to start the process
- Decryption: $M_i = D_k(C_i) \oplus C_{i-1}$
- Application: general block-oriented transmission.

Stream Ciphers

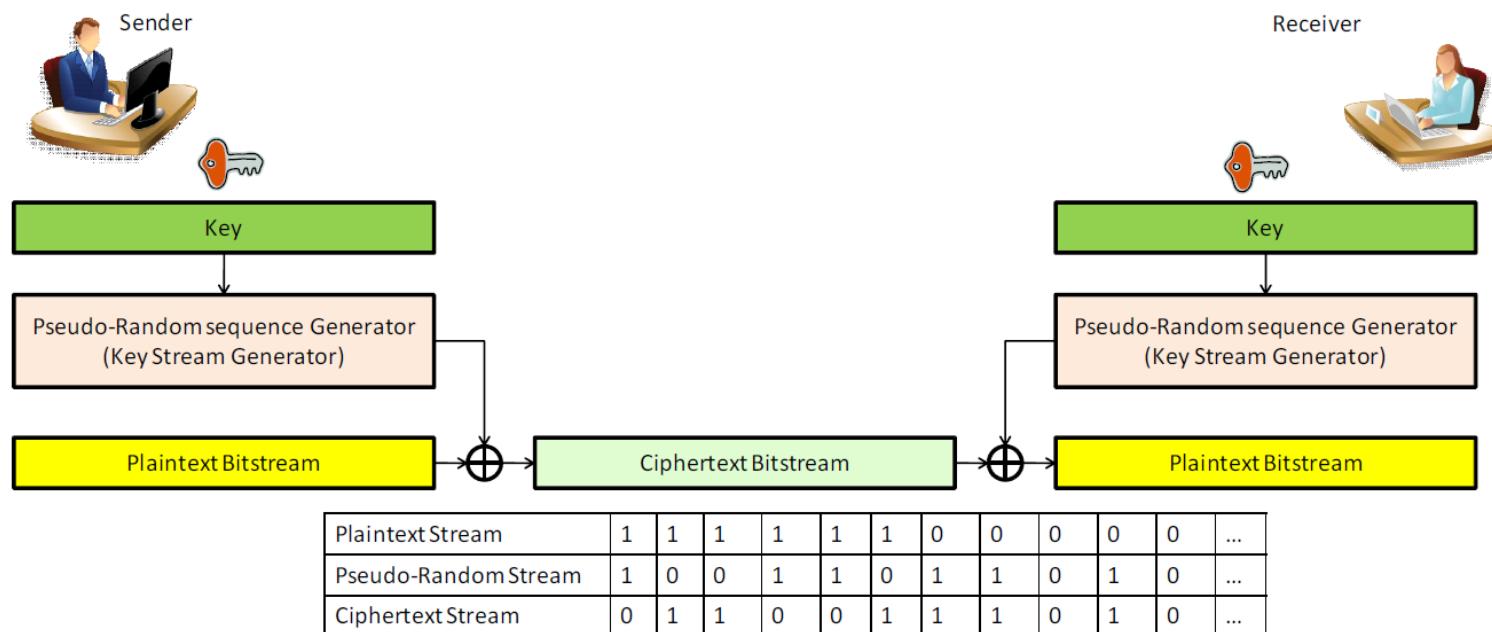
- Many times data is transmitted in serial form (one bit at a time)
- Cipher generates “Key-Stream” which is combined with “Message-Stream” to produce “Cipher-Stream”



Plaintext Stream	1	1	1	1	1	1	0	0	0	0	0	...
Pseudo-Random Stream	1	0	0	1	1	0	1	1	0	1	0	...
Ciphertext Stream	0	1	1	0	0	1	1	1	0	1	0	...

Stream Ciphers (2)

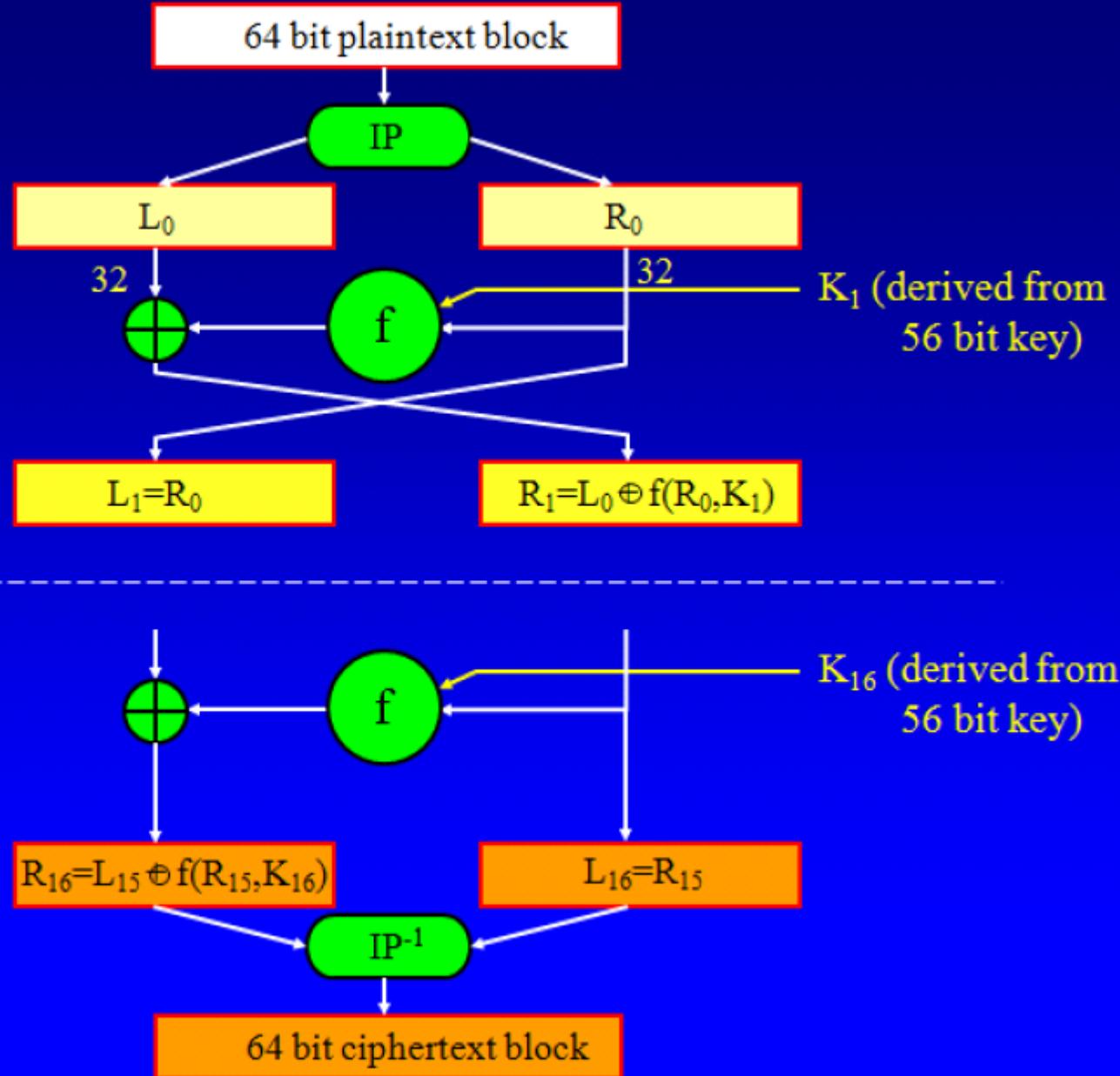
- Inverted at receiver by combining the same Key Stream
- Better than Block Ciphers for Serial Communication Channels
- Repeated input patterns do not produce repeated cipher stream sequences
- Only adds Confusion (no Diffusion)
- If Synchronization lost between sender and receiver (Cryptosync) – must resync



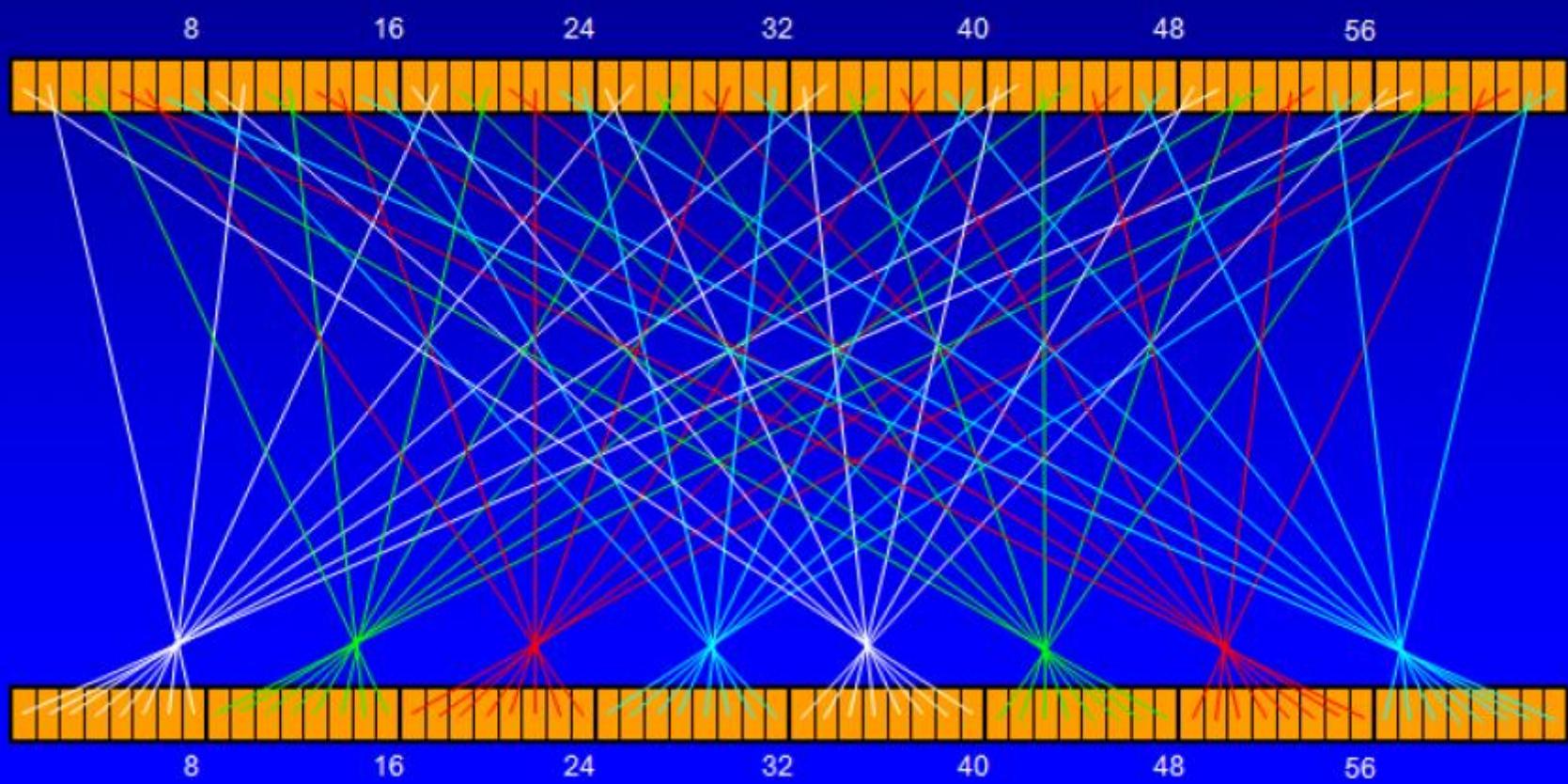
Example - DES: The Data Encryption Standard

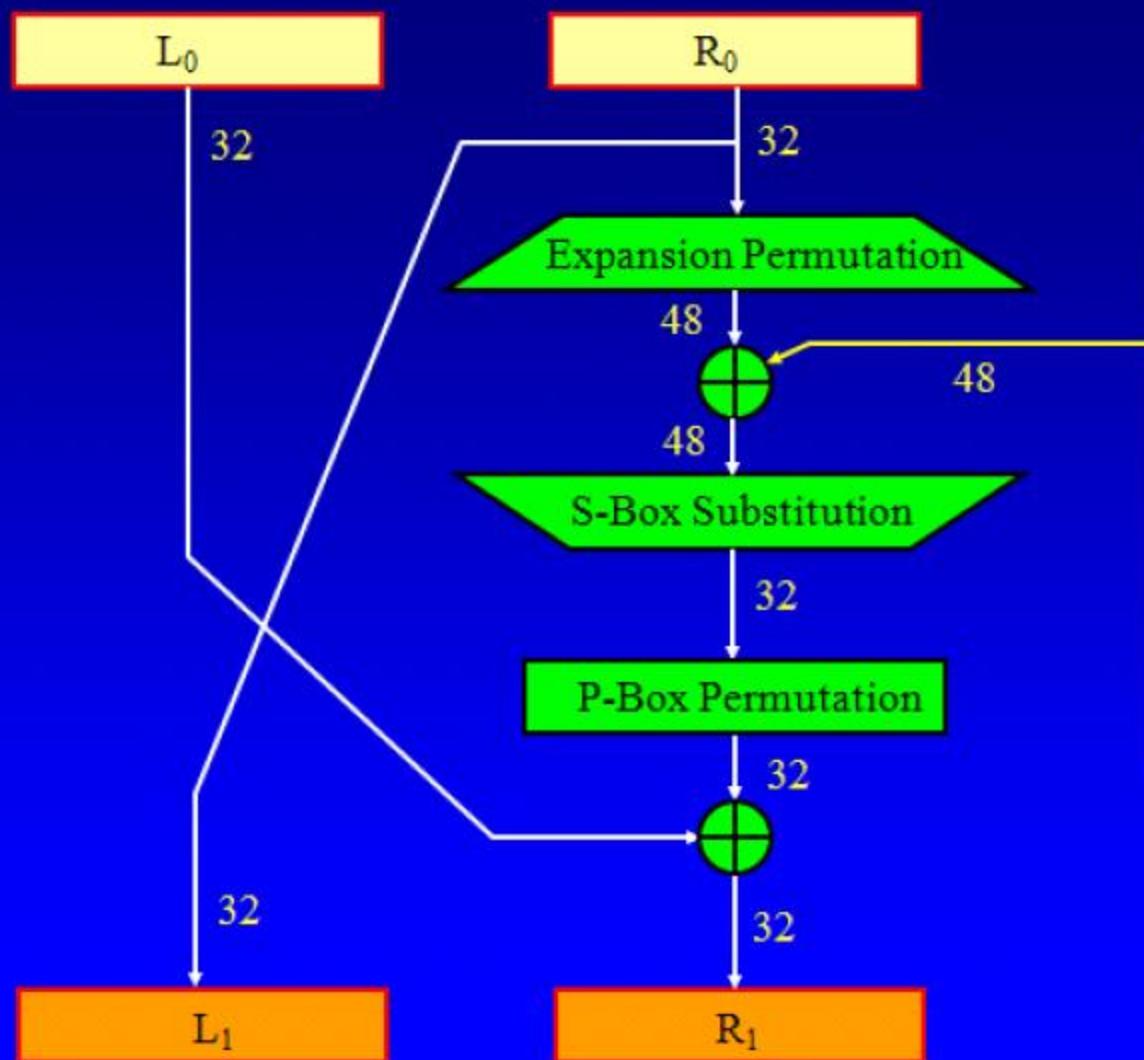
- Most widely used block cipher in the world.
- Adopted by NIST in 1977.
- Based on the Feistel cipher structure with 16 rounds of processing.
- Block = 64 bits, Key = 56 bits
- Design Principles of DES: To achieve high degree of **diffusion** and **confusion**.

DES



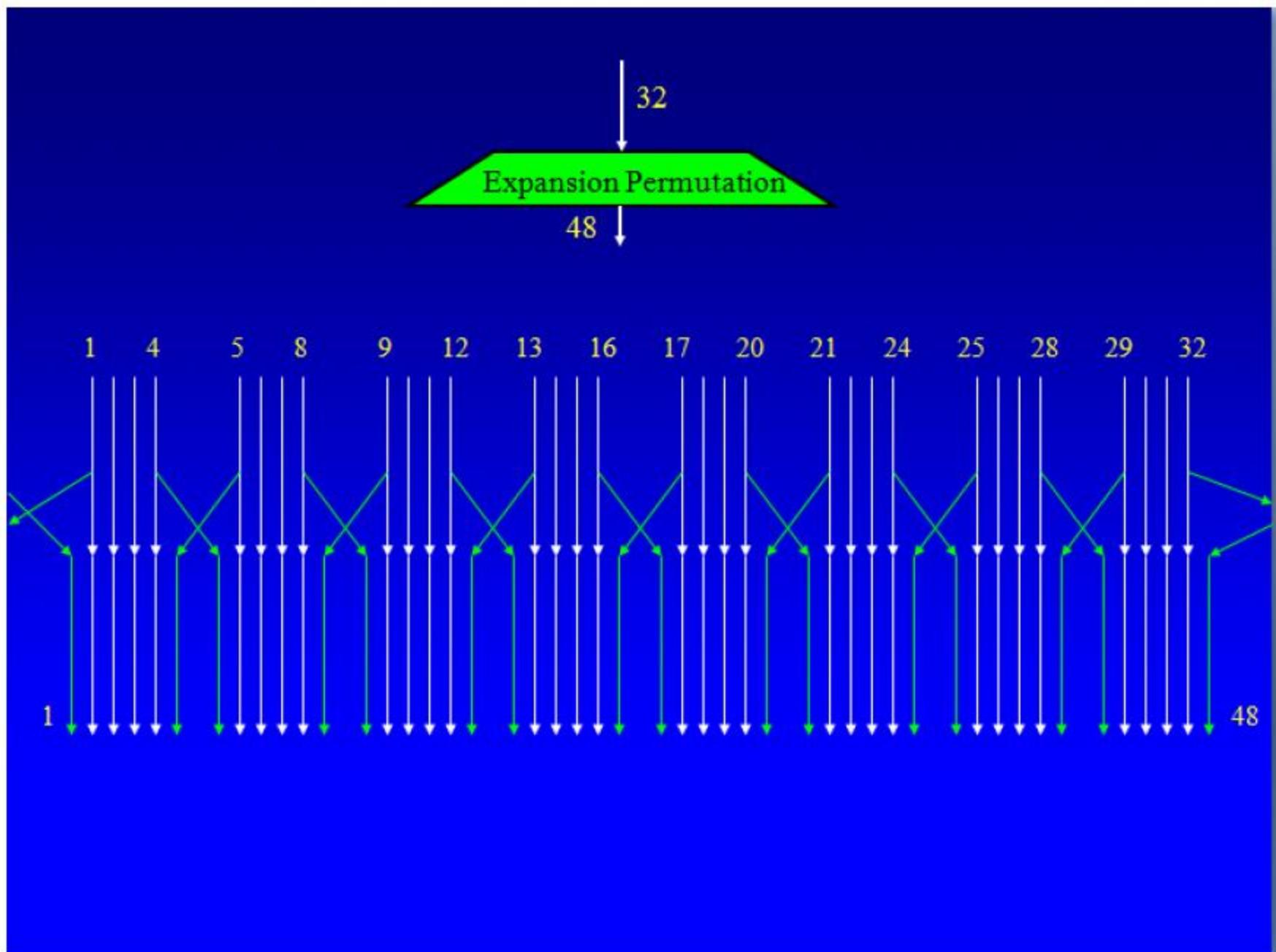
IP (Initial Permutation):

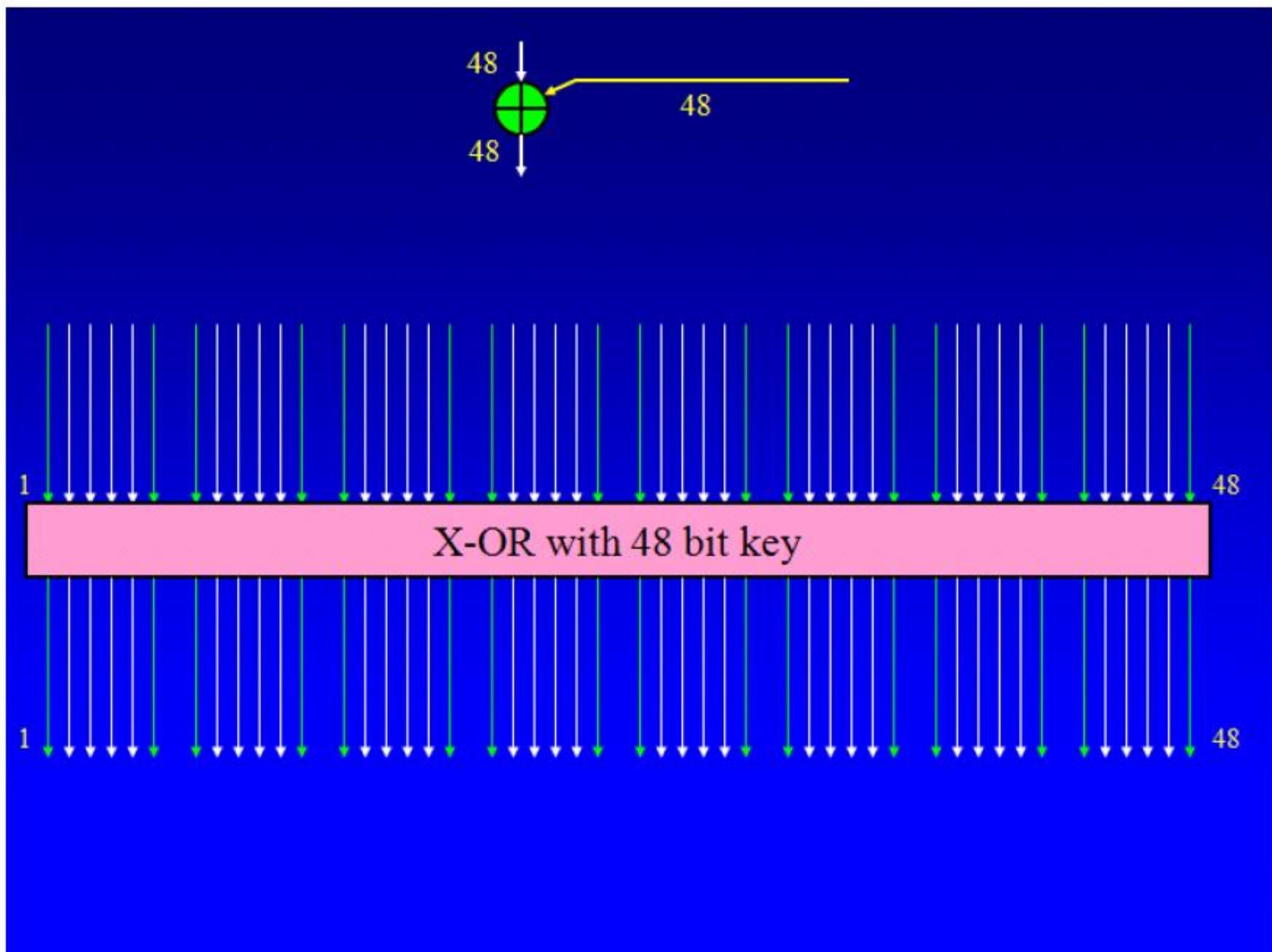


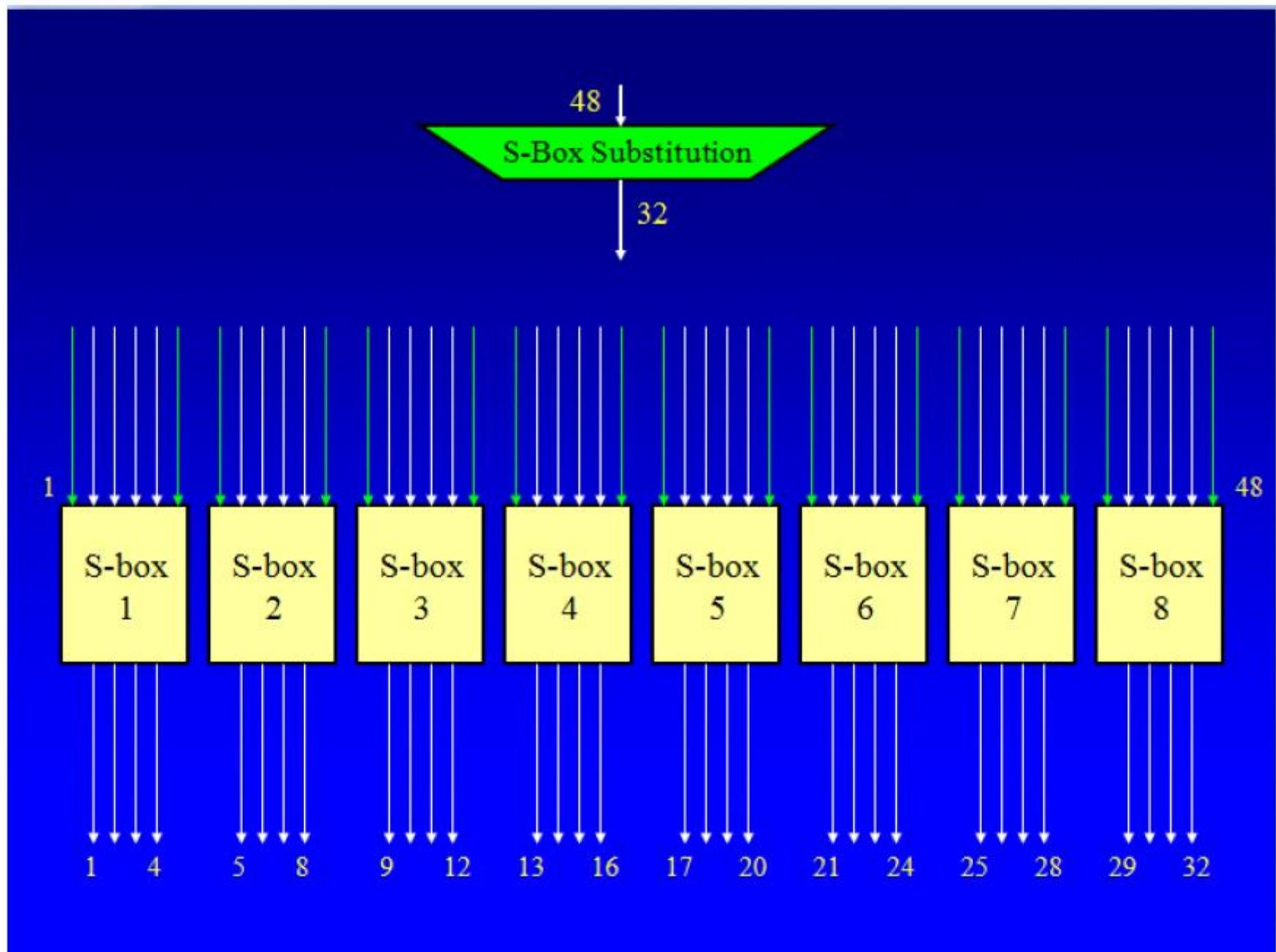


48 bit subkey
Generator
 $K_{48} = g(i, K_{56})$

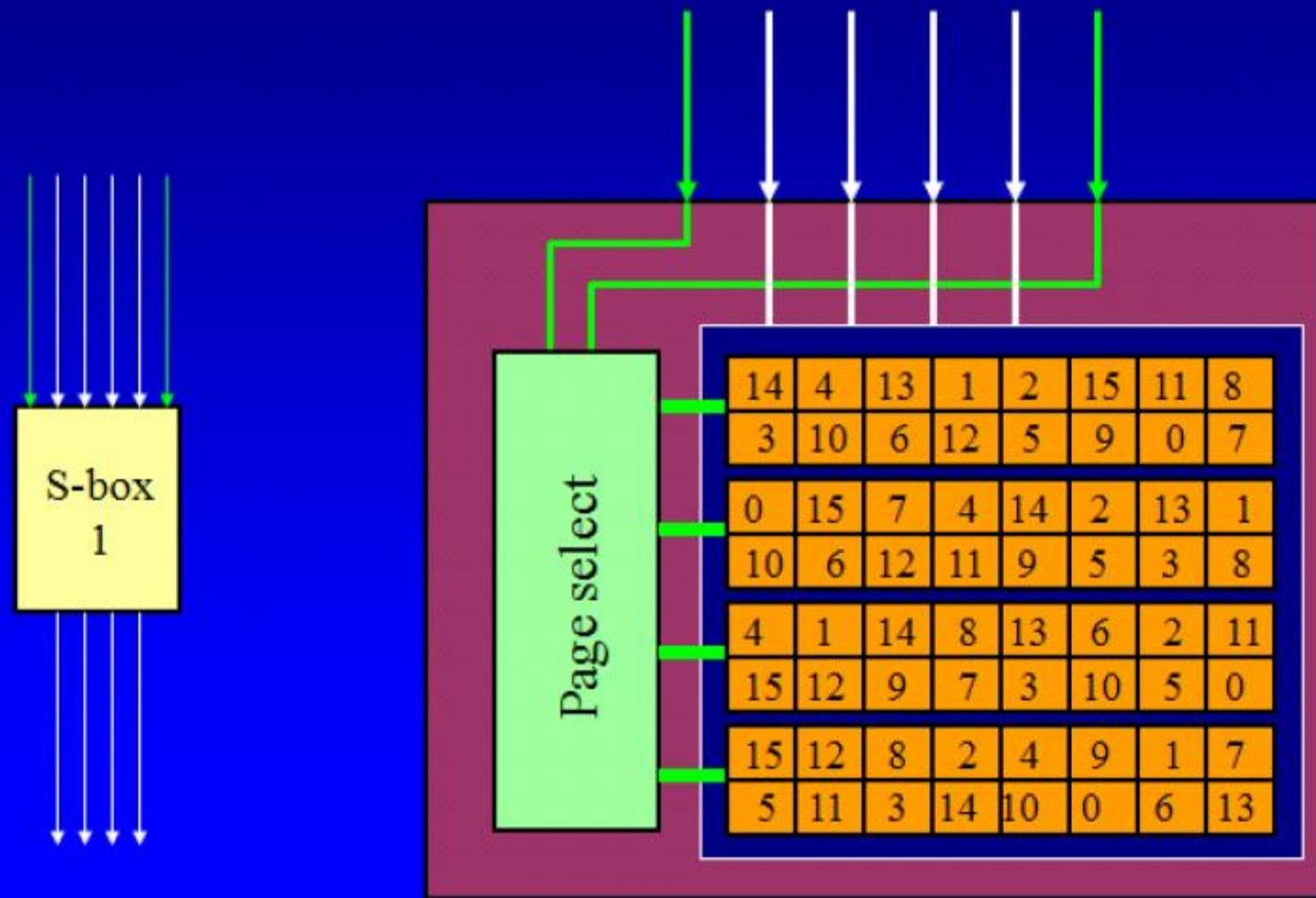
(The key for
each round is
deterministically
found from the
input 56 bit key).



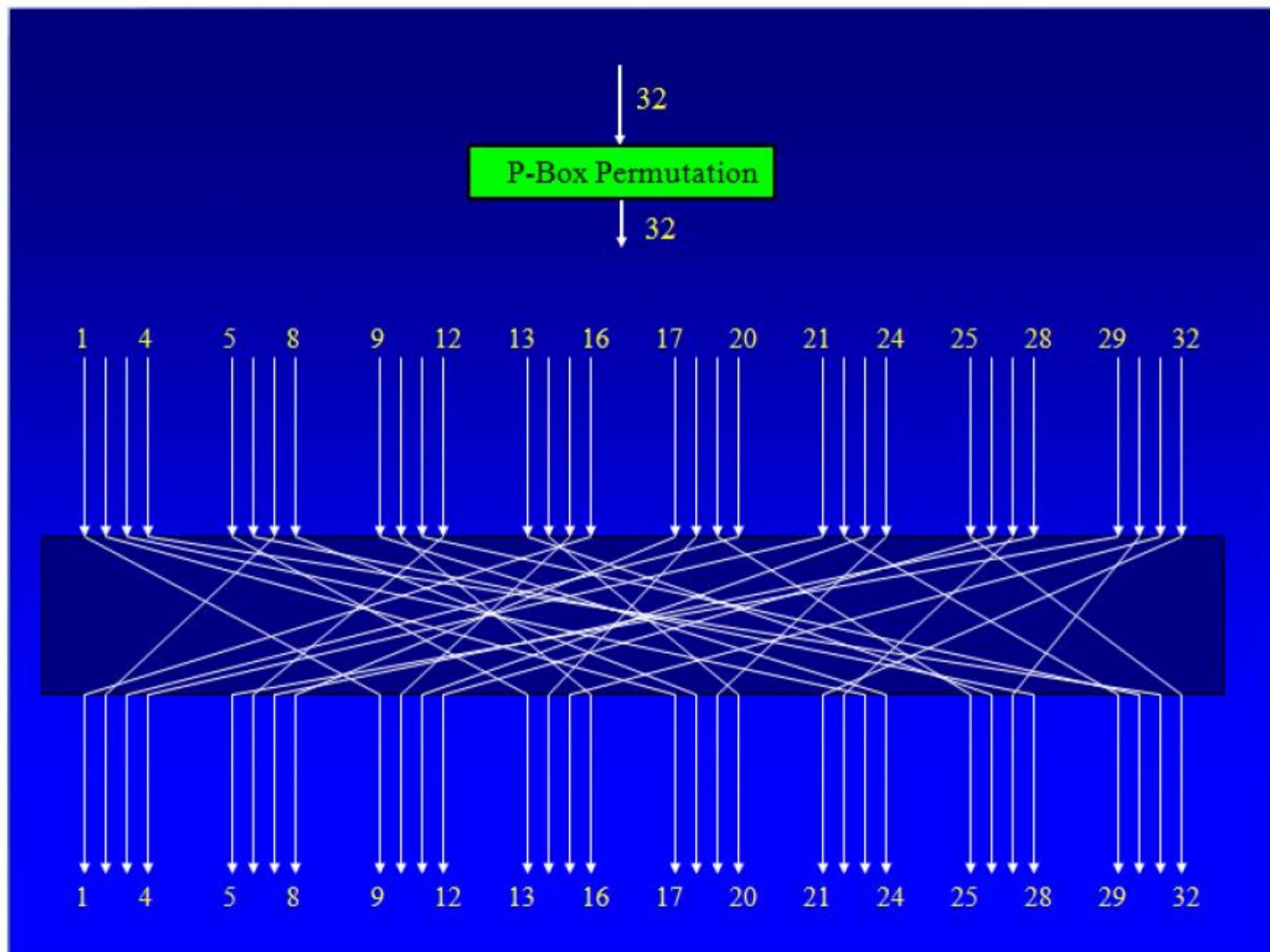




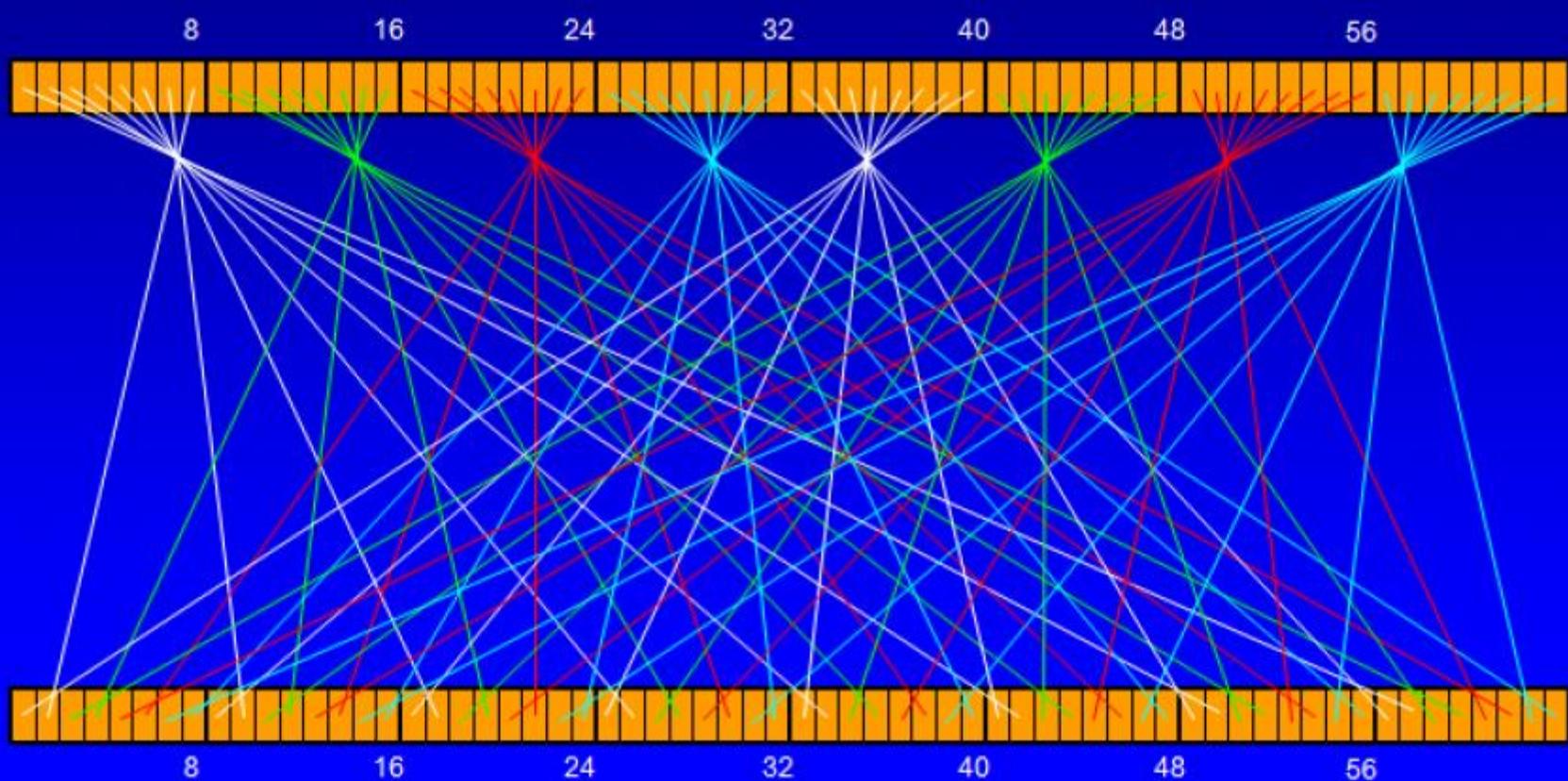
How an S-Box works



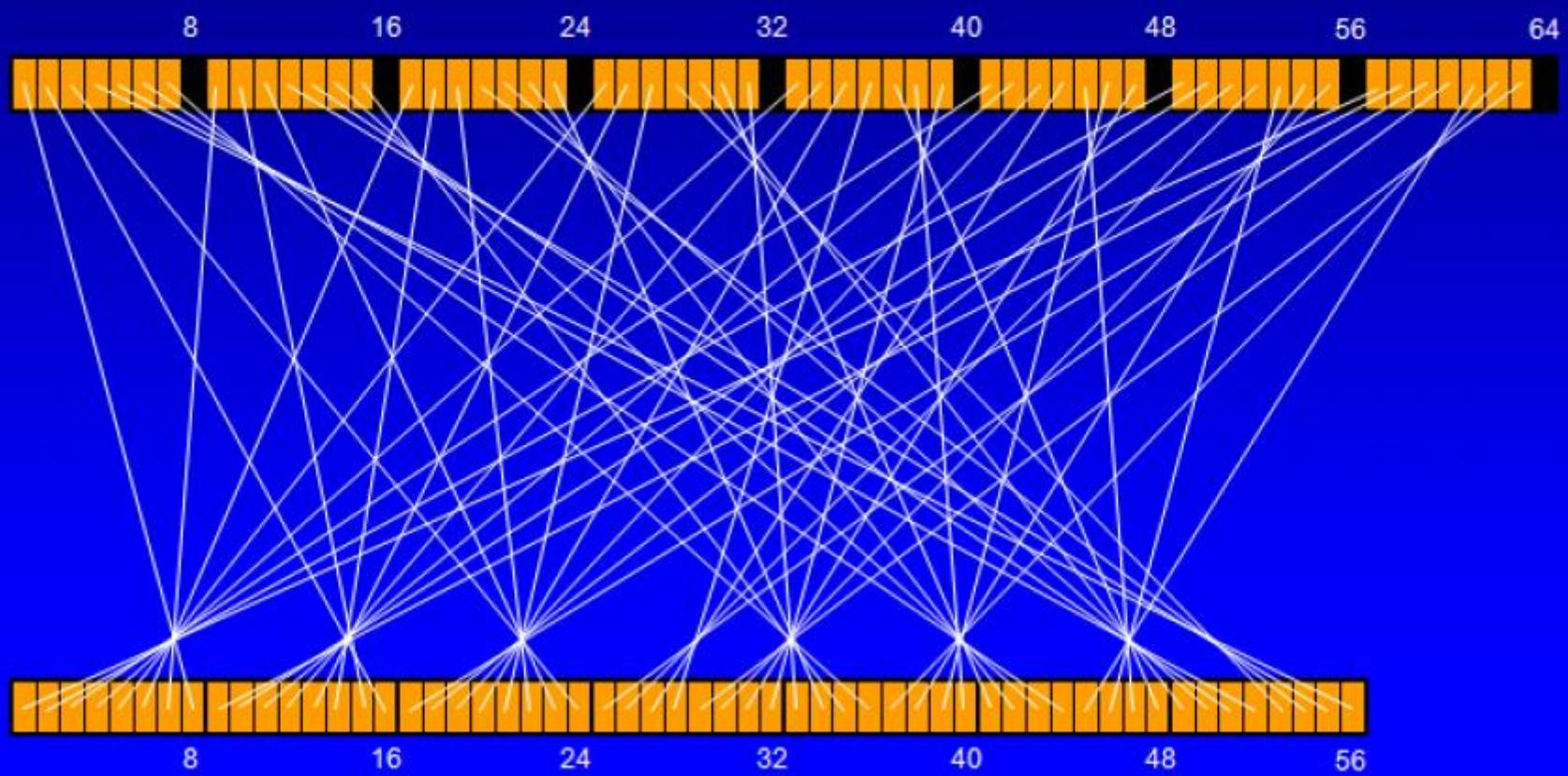
S ₅		Middle 4 bits of input																	
		0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111		
Outer bits	00	0010	1100	0100	0001	0111	1010	1011	0110	1000	0101	0011	1111	1101	0000	1110	1001		
	01	1110	1011	0010	1100	0100	0111	1101	0001	0101	0000	1111	1010	0011	1001	1000	0110		
	10	0100	0010	0001	1011	1010	1101	0111	1000	1111	1001	1100	0101	0110	0011	0000	1110		
	11	1011	1000	1100	0111	0001	1110	0010	1101	0110	1111	0000	1001	1010	0100	0101	0011		



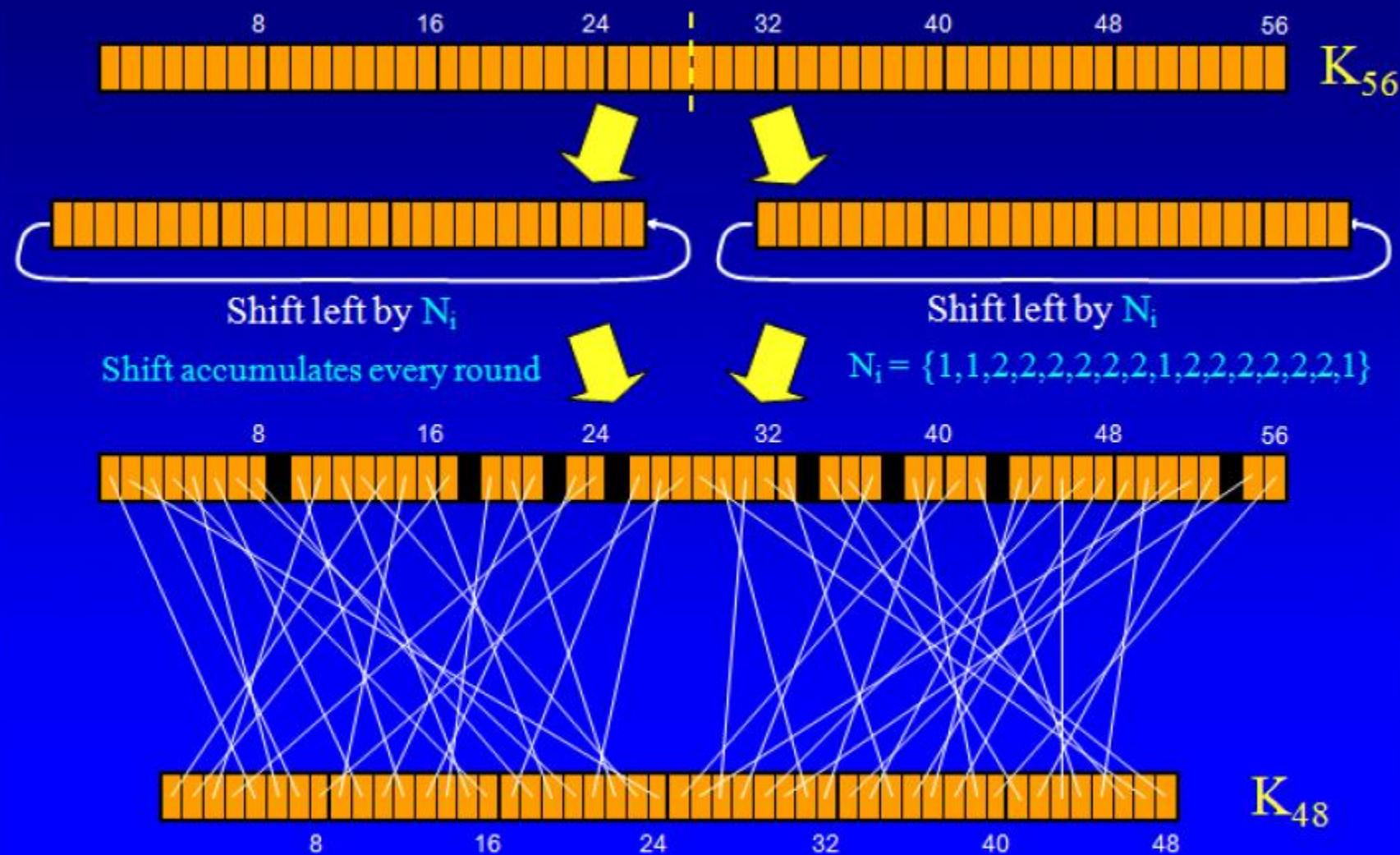
IP⁻¹ (Final Permutation):



Initial Key Permutation



Key Split & Shift & Compress

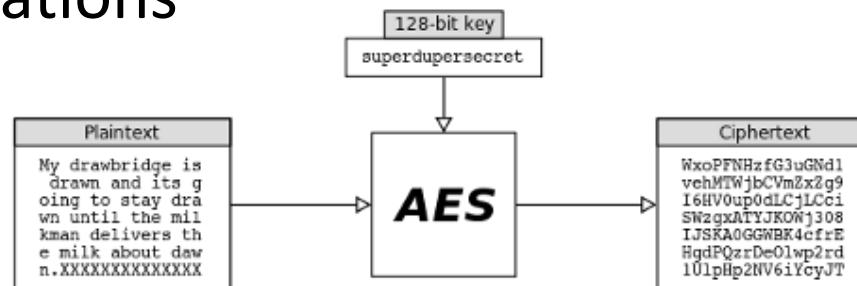


DES

- DES exhibits a strong avalanche effect
 - Changing 1 bit in the plaintext affects 34 bits in the ciphertext on average.
 - 1-bit change in the key affects 35 bits in the ciphertext on average.
- Attacks on DES
 - Brute-force key search
 - Needs only two plaintext-ciphertext samples
 - Trying 1 key per microsecond would take 1000+ years on average, due to the large key space size, $2^{56} \approx 7.2 \times 10^{16}$
 - Differential cryptanalysis
 - Known-plaintext attack
 - Possible to find a key with 2^{47} plaintext-ciphertext samples
 - Linear cryptanalysis
 - Known-plaintext attack
 - Possible to find a key with 2^{43} plaintext-ciphertext samples
- DES is feeling its age. A more secure cipher is needed.

Advanced Encryption Standard

- NIST created a program for the development of Advanced Encryption Standard (AES) (first call Sept. 97)
- “Winner” – **Rijndael** announced Oct. 2000
- Rijndael (Daemen and Rijmen) supports keys of 128, 192, or 256 bits and messages of 128, 192, or 256 bits (AES uses only 128 bit blocks)
- Designed to be resistant to linear or differential cryptanalysis
- Fast and efficient in hardware and software implementations

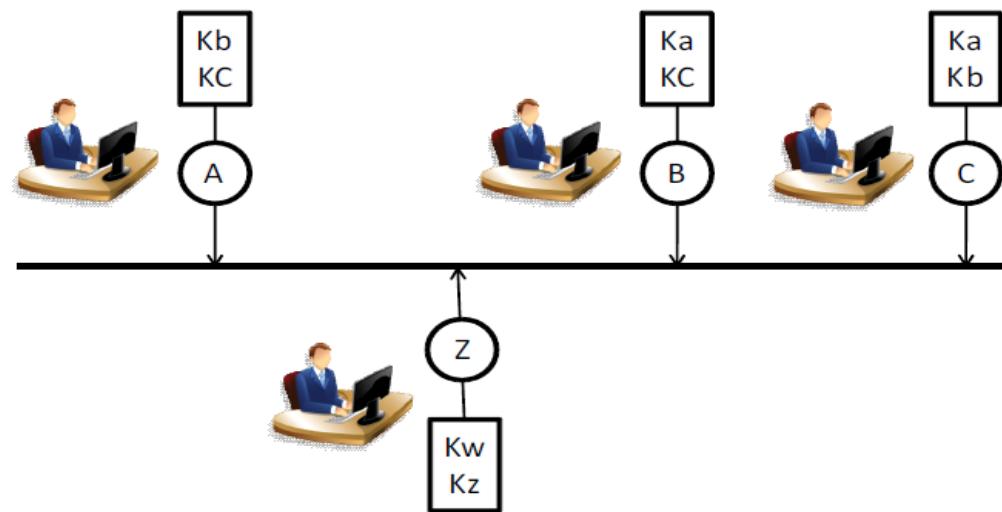


Symmetric Encryption Issues

- Symmetric Encryptions provide inherent user and data authentication
- Generally very fast
- Key Distribution and Management is a problem
- Digital signature are very difficult to realize

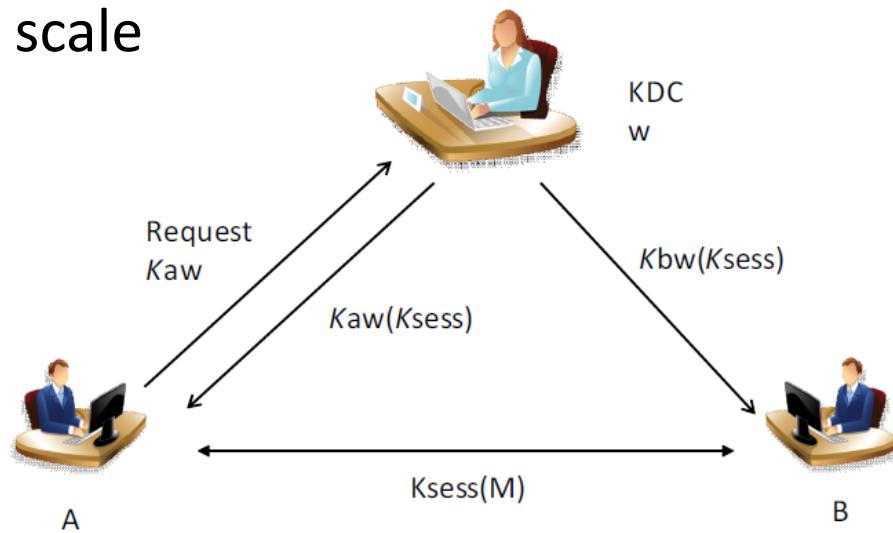
Key Management

- Requires approx. $\frac{N^2}{2}$ keys
 - Assume total N users, each user shares $N - 1$ keys with other users
 - Then, the total number of keys are $N(N - 1)$
 - Since a pair of users hold the same key, i.e. $K_{ac} = K_{ca}$,
$$\frac{N(N-1)}{2} \approx \frac{N^2}{2}$$



Alternative Key Management

- Each user shares a key with a Key Distribution Centre
- Users contact KDC requesting common session key
- KDC sends each party the session key
- Problem:
 - System is vulnerable to failure or compromise of KDC
 - KDC can “listen-in” to all users communications
 - Difficult to expand on large scale



Hash Function

-
- Hash Function Properties**
- Birthday Attack**
- Bloom Filter**

What are hash functions?

- Just a method of compressing strings
 - E.g., $H : \{0,1\}^* \rightarrow \{0,1\}^{160}$
 - Input is called “message”, output is “digest”
- Why would you want to do this?
 - Short, fixed-size better than long, variable-size
 - ❖ True also for non-crypto hash functions
 - Digest can be added for redundancy
 - Digest hides possible structure in message



- Features of Hash Functions
 - Hash functions form an unique image of a message
 - The hash can be applied to any size of message to produce a fixed size output
 - The hash is easy and efficient to compute but is computationally infeasible to invert

Classification of Hash Functions

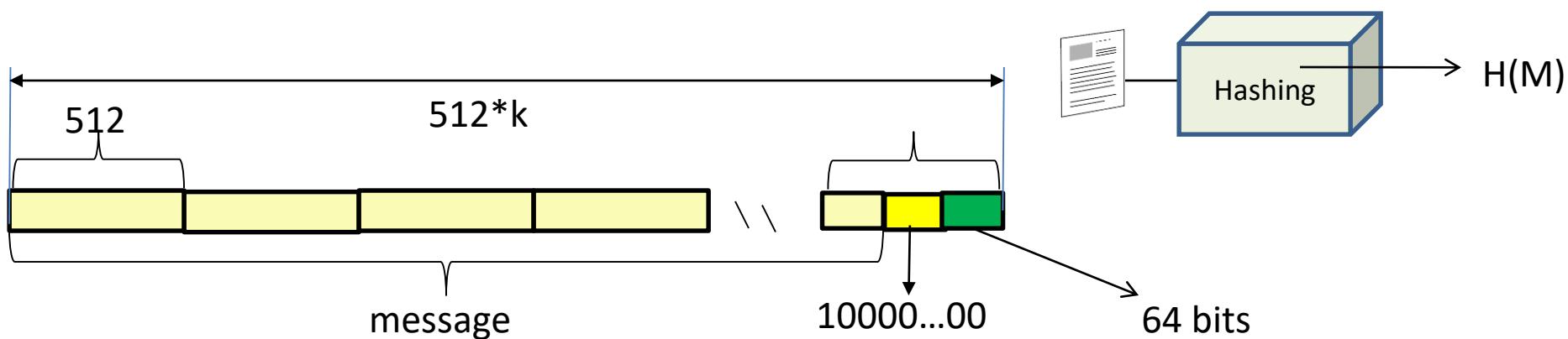
- MIC (message integrity codes)
 - Unkeyed: no keys required, anyone can generate or verify
 - One-Way Hash Functions (OWHFs)
 - Given y , hard to find x such that $H(x)=y$
 - Collision Resistant Hash Functions (CRHFs)
 - Hard to find $x \neq x'$ such that $H(x)=H(x')$
- MAC (message authentication codes)
 - both authentication and integrity
 - Keyed: usually based on a cipher – users must exchange secret key in order to authenticate
 - requires no additional mechanism

Properties of Hash Functions

- **Preimage resistance:** given y it's computationally infeasible to find a value x s.t. $h(x)=y$
- **2-nd preimage resistance:** given x and $y=h(x)$ it's computationally infeasible to find a value $x' \neq x$ s.t. $h(x')=h(x)$
- **Collision resistance:** it's computationally infeasible to find any two distinct values x', x s.t. $h(x')=h(x)$

MD5

- Introduced in 1992 by Ron Rivest (RSA fame)
- Un-keyed hash
- Processes 512 bit input blocks – 128 bit output hash
- Initially, message is padded so that is 64 bits shorter than a multiple of 512 bits by adding a single “1” and then “0”’s
- The last 64 bits are used to represent the length of the message prior to padding



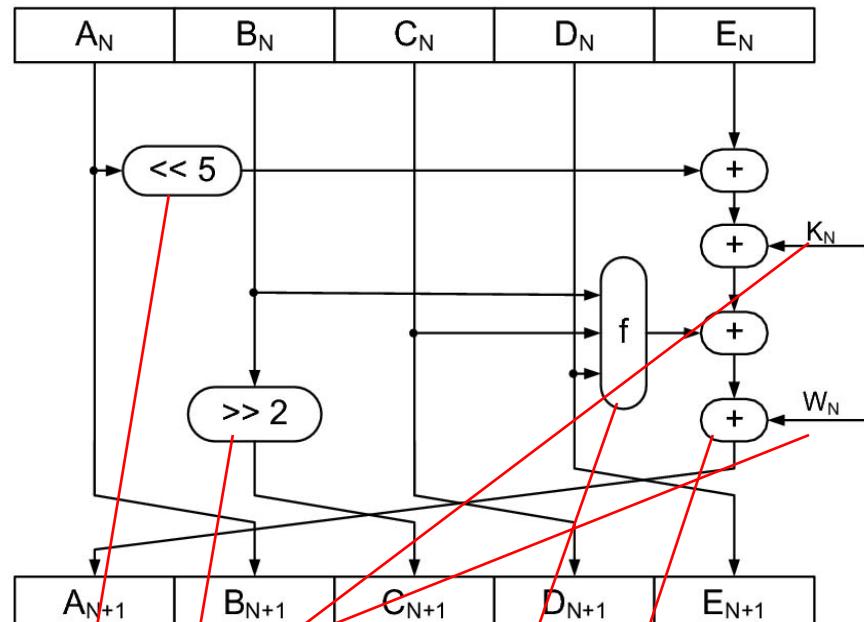
Secure Hash Algorithm (SHA)

- SHA-1(Secure Hash Algorithm)

- designed by the National Security Agency and published by the NIST as a U.S. Federal Information Processing Standard.
- iterated hash function
- 160-bit message digest
- word-oriented (32 bit) operation on bitstrings
- Padding scheme extends the input x by at most one extra 512-bit block
- The compression function maps 160+512 bits to 160 bits
- Make each input affect as many output bits as possible

expanded message word

Round constant



f is a nonlinear function

addition modulo 2^{32} .

a left bit rotation by 5 places

a right bit rotation by 2 places



SHA

- SHA was modified to SHA-1 in 1992
- May 2001 – FIPS (Federal Information Processing Standards) 180-2 proposal for expansion to 256, 384, and 512 bit hash
- **Designed to match security of AES at 128, 192, and 256 bits**
- July 2007, NIST (National Institute of Standards and Technology) announced a call for the development of a new hash function -- SHA-3 function to replace the older SHA
 - The list of 14 candidates accepted to Round 2 was published on July 24, 2009
 - The announcement of the final round candidates occurred on December 10, 2010
 - and the proclamation of a winner and publication of the new standard are scheduled to take place in 2012.



Birthday Attack

- Birthday paradox
 - 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%.
 - Finding two people with the same birthday is the same thing as finding a collision for this particular hash function.

$$P(n=23, m=365) = 1 - \frac{m^{(n)}}{m^n} = 1 - \frac{365^{(23)}}{365^{23}}$$

$$m^{(n)} = \frac{m!}{(m-n)!} = m \times (m-1) \times (m-2) \cdots (m-n+1)$$



Birthday Attack

- The probability that all 23 people have different birthdays is

$$1 \times \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \dots \left(1 - \frac{22}{365}\right) = 0.493$$

- Therefore, the probability of at least two having the same birthday is $1 - 0.493 = 0.507$
- Birthday Attack \rightarrow Square Root Attack

$$n \approx \sqrt{m} \Rightarrow 23 \approx \sqrt{365} \Rightarrow prob = 0.507$$



Birthday Attack

- The birthday attack (also called square root attacks) can be used to find collisions for hash functions if the output of the hash function is not sufficiently large.
- Suppose h is an n -bit hash function. Then there are $N = 2^n$ possible outputs. We have the situation of list of length $r \approx \sqrt{N}$ “people” with N possible “birthdays,” so there is a good chance of having two values with the same hash value
- (**Old**) Banking standards used 32bit DES MACs – If you had the key, it would only require 2^{16} tries to find a message that would produce the same hash value
- Today, hash value are at least 128 bits, preferably 160 bits to prevent square root attacks

Theorem: Birthday Bound

- 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

$$P(N, q) \leq \frac{q(q - 1)}{2N}$$

and

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

also if $1 \leq q \leq \sqrt{2N}$

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

Proposition

- **The inequality**

$$\left(1 - \frac{1}{e}\right) \cdot x \leq 1 - e^{-x} \leq x$$

Is true for any real number x with $0 \leq x \leq 1$.

Proof of Theorem

- 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,
- $P(N, q) = \Pr[C_1 \vee C_2 \vee \dots \vee C_q] \leq \Pr[C_1] \vee \Pr[C_2] \vee \dots \vee \Pr[C_q] \leq \frac{0}{N} + \frac{1}{N} + \dots \frac{q-1}{N} = \frac{(q-1)q}{2N}$
- This proves the upper bound.

Proof of Theorem (2)

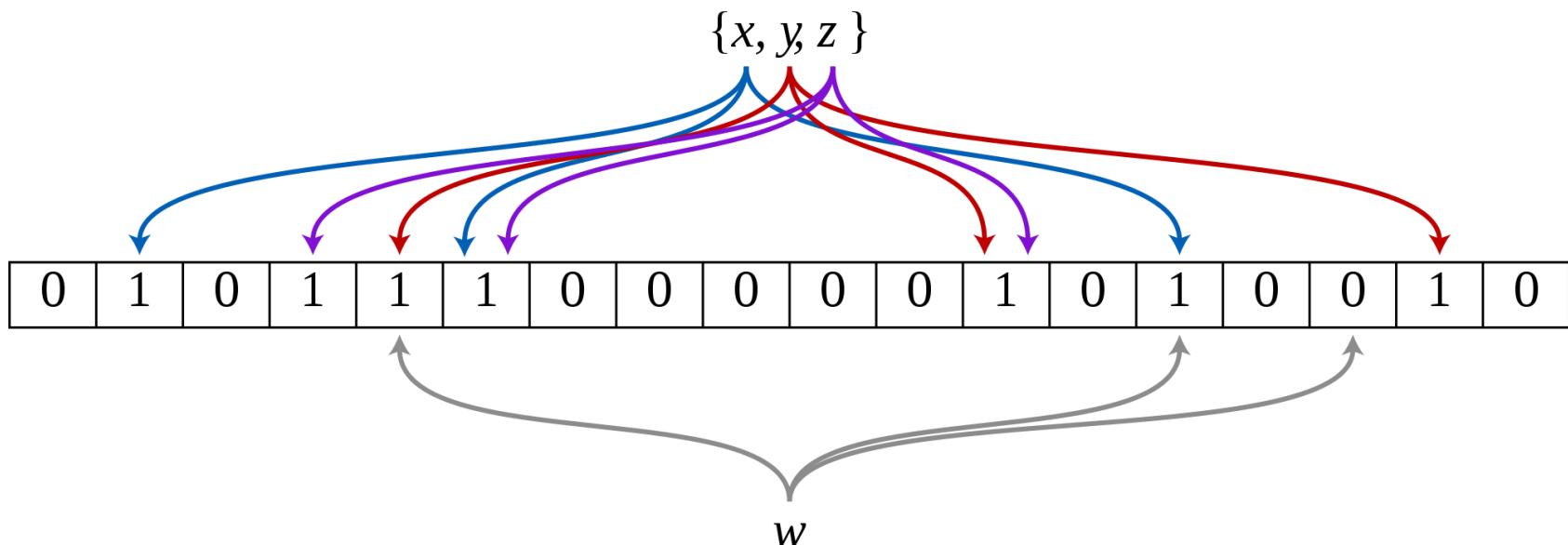
- For the lower bound we let D_i be the event that there is no collision after having thrown in the i -th ball. If there is no collision after throwing in i balls then they must all be occupying different slots, so the probability of no collision upon throwing in the $(i + 1)$ -st ball is exactly $\frac{(N-i)}{N}$. That is, $\Pr[D_{i+1}|D_i] = \frac{N-i}{N} = 1 - \frac{i}{N}$.
- Also note $\Pr[D_1] = 1$. The probability of no collision at the end of the game can now be computed via
- $1 - P(N, q) = \Pr[D_q] = \Pr[D_q|D_{q-1}] \cdot \Pr[D_{q-1}] = \dots = \prod_{i=1}^{q-1} \Pr[D_{i+1}|D_i] = \prod_{i=1}^{q-1} \left(1 - \frac{i}{N}\right)$
- Note that $\frac{i}{N} \leq 1$, 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 $\prod_{i=1}^{q-1} \left(e^{-\frac{i}{N}}\right) = e^{-\frac{1}{N}-\frac{2}{N}-\dots-\frac{q-1}{N}} = e^{-\frac{q(q-1)}{2N}}$
- Therefore $P(N, q) \geq 1 - e^{-\frac{q(q-1)}{2N}}$

Proof of Theorem (3)

- To get the last one, we need to make some more estimates. We know $\frac{(q-1)q}{2N} \leq 1$ because $q \leq \sqrt{2N}$, so we can use the inequality
$$\left(1 - \frac{1}{e}\right) \cdot x \leq 1 - e^{-x} \quad \text{to get } P(N, q) \geq \left(1 - \frac{1}{e}\right) \frac{(q-1)q}{2N}$$

Bloom Filter

- Approximate set membership problem .
- Trade-off between the space and the false positive probability .
- Generalize the hashing ideas.



Approximate set membership problem

- Suppose we have a set
 $S = \{s_1, s_2, \dots, s_m\} \subseteq \text{universe } U$
- Represent S in such a way we can quickly answer “Is x an element of S ?”
- To take as little space as possible, we allow false positive
 - i.e. sometimes $x \notin S$, we still answer yes
- But, if $x \in S$, we must answer yes .

Formal Description of Bloom Filter

Bloom Filter consists of an arrays $A[n]$ of n bits (space), and k independent random hash functions

$$h_1, \dots, h_k : U \rightarrow \{0, 1, \dots, n-1\}$$

1. Initially set the array to 0
2. $\forall s \in S, A[h_i(s)] = 1$ for $1 \leq i \leq k$
(an entry can be set to 1 multiple times, only the first times has an effect)
3. To check if $x \in S$, we check whether all location $A[h_i(x)]$ for $1 \leq i \leq k$ are set to 1

If not, clearly $x \notin S$.

If all $A[h_i(x)]$ are set to 1 ,we assume $x \in S$

Bloom Filter

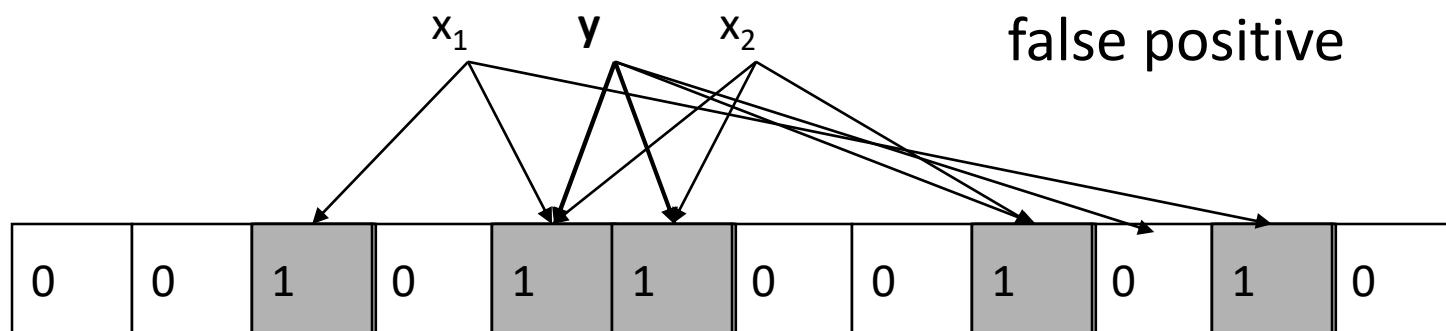
Initial with all 0

0	0	0	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---

Each element of S is hashed k times

Each hash location set to 1

If only 1s appear,
conclude that y is
in S, This may yield
false positive



To check if y is in S , check the k hash location. If a 0 appears , y is not in S

False positive

- After all the elements of S are hashed into the bloom filters ,the probability that a specific bit is still 0 is (Assume all hash functions are random)

$$p = \left(1 - \frac{1}{n}\right)^{km} \approx e^{-km/n}$$

- To simplify the analysis ,we can assume a fraction p of the entries are still 0 after all the elements of S are hashed into bloom filters.

False positive

- The probability of a false positive f is

$$f = (1 - p)^k \approx (1 - e^{-km/n})^k$$

- To find the optimal k to minimize f .

Minimize f iff minimize $g = \ln(f)$

$$\frac{dg}{dk} = \ln(1 - e^{-km/n}) + \frac{km}{n} \frac{e^{-km/n}}{1 - e^{-km/n}}$$

$\Rightarrow k = \ln 2 \cdot \frac{n}{m}$ (the number of hash functions)

$\Rightarrow f = (1/2)^k = (0.6185..)^{n/m}$

The false positive probability falls exponentially in n/m , the number bits used per item.

$$f = (1-p)^k \approx (1-e^{-km/n})^k \quad \rightarrow \quad y = \frac{1}{2}$$

$$g = \ln(f) = \ln\left(1 - e^{-\frac{km}{n}}\right)^k = k \cdot \ln\left(1 - e^{-\frac{km}{n}}\right) \quad y = e^x = \frac{1}{2}$$

$$\frac{dg}{dk} = \ln\left(1 - e^{-\frac{km}{n}}\right) + \frac{e^{-\frac{km}{n}}}{1 - e^{-\frac{km}{n}}} \cdot \frac{km}{n} = 0 \quad x = -\ln 2$$

$$\frac{dg}{dk} = \ln\left(1 - e^{-\frac{km}{n}}\right) + \frac{e^{-\frac{km}{n}}}{1 - e^{-\frac{km}{n}}} \cdot \frac{km}{n} = 0 \quad x = -\ln 2 = -\frac{km}{n}$$

$$(1 - e^{-\frac{km}{n}}) \ln\left(1 - e^{-\frac{km}{n}}\right) = e^{-\frac{km}{n}} \cdot \left(-\frac{km}{n}\right) \quad k = \ln 2 \cdot \frac{n}{m}$$

Let $x = -\frac{km}{n}$

$$(1 - e^x) \ln(1 - e^x) = e^x \cdot x$$

$$\ln(1 - e^x)^{(1-e^x)} = e^x \cdot x$$

$$(1 - e^x)^{(1-e^x)} = e^{x \cdot e^x}$$

Let $y = e^x$

$$(1 - y)^{(1-y)} = y^y$$

Summary of Bloom Filter

- A Bloom filters is like a hash table ,and simply uses one bit to keep track whether an item hashed to the location.
- If $k = 1$, it's equivalent to a hashing based fingerprint system.
- It's interesting that when k is optimal $k = \ln 2 \cdot \frac{n}{m}$, then $p = \frac{1}{2}$.
- **Exercise:** Alice wants to send $m=10$ files to Bob, how to use minimal overhead to assure the data integrity of these files?

Public Key Encryption

- Number Theory
- RSA Encryption, Signature
- ElGamal Encryption
- Diffie-Hellman Key Exchange

Number Theory

- Extended Euclidean Algorithm
- Modular Arithmetic
- Euler Totient Function $\phi(n)$
- Fermat's Theorem
- Euler's Theorem
- Chinese Remainder Theorem

Extended Euclidean Algorithm

- Given two integers a and b , we often need to find other two integers, u and v , such that

$$u \times a + v \times b = \gcd(a, b)$$

- The extended Euclidean algorithm can calculate the $\gcd(a, b)$ and at the same time calculate the value of u and v .



Euclid

Extended Euclidean Algorithm

Dividend	Divisor	Quotient	Reminder
a=60	= b=13	× 4	+ 8
b=13	= 8	× 1	+ 5
8	= 5	× 1	+ 3
5	= 3	× 1	+ 2
3	= 2	× 1	+ 1

$$1 = 3 - 2 \times 1$$

$$1 = 3 - (5 - 3 \times 1) \times 1 = 3 \times 2 - 5 \times 1$$

$$1 = (8 - 5 \times 1) \times 2 - 5 \times 1 = 8 \times 2 - 5 \times 3$$

$$1 = 8 \times 2 - (13 - 8 \times 1) \times 3 = 8 \times 5 - 13 \times 3$$

$$1 = (60 - 13 \times 4) \times 5 - 13 \times 3 = \underline{60 \times 5} - \underline{13 \times 23}$$

$$GCD(a=60,b=13)$$

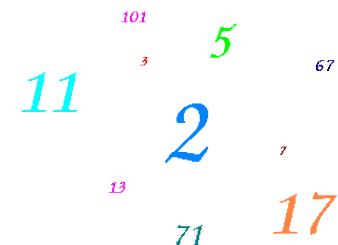
$$GCD(60,13)=1$$

$$1 = 60 \times 5 + 13 \times (-23)$$

$$U=5$$

$$V=-23$$

1723



Modular Arithmetic

- $a \equiv b \pmod{n}$
 - n is the modulus
 - a is “congruent” to b, modulo n
 - a - b is divisible by n $n | (a-b)$
 - $a \% n = b \% n$ $7 \% 12 = 19 \% 12 = 7$
- $a \equiv b \pmod{n}$, $c \equiv d \pmod{n}$
 - Addition
 $a + c \equiv b + d \pmod{n}$
 - Multiplication
 $ac \equiv bd \pmod{n}$

$$a - b = jn$$

$$c - d = kn$$

$$a + c - (b + d) = (j + k)n$$

Modular Arithmetic (Cont.)

- Power

➤ $a \equiv b \pmod{n} \rightarrow a^k \equiv b^k \pmod{n}$

If $a^k \equiv b^k \pmod{n}$,

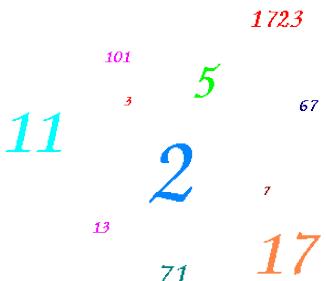
According to multiplication rule

$$a \cdot a^k \equiv b \cdot b^k \pmod{n},$$

$$\therefore a^{k+1} \equiv b^{k+1} \pmod{n}$$

- Going n times around the clock

➤ $a + kn \equiv b \pmod{n}$



Modular Arithmetic (Cont.)

- If a, b have no common factors, there exists a^{-1} such that $a \cdot a^{-1} \equiv 1 \pmod{b}$
 - a^{-1} is called the “multiplicative inverse”
 - We can calculate a^{-1} with extended Euclidean algorithm

Dividend	Divisor	Quotient	Remainder
$a=60$	$= b=13$	$\times 4$	$+ 8$
$b=13$	$= 8$	$\times 1$	$+ 5$
8	$= 5$	$\times 1$	$+ 3$
5	$= 3$	$\times 1$	$+ 2$
3	$= 2$	$\times 1$	$+ 1$

$$1 = 3 - 2 \times 1$$

$$1 = 3 - (5 - 3 \times 1) \times 1 = 3 \times 2 - 5 \times 1$$

$$1 = (8 - 5 \times 1) \times 2 - 5 \times 1 = 8 \times 2 - 5 \times 3$$

$$1 = 8 \times 2 - (13 - 8 \times 1) \times 3 = 8 \times 5 - 13 \times 3$$

$$1 = (60 - 13 \times 4) \times 5 - 13 \times 3 = 60 \times 5 - 13 \times 23$$

$$1 = 60 \times 5 - 13 \times 23 \pmod{60} = -13 \times 23 \pmod{60}$$

$$1 = 13 \times (-23) = 13 \times 37 \pmod{60}$$

For example:

$$13 \times ? \equiv 1 \pmod{60}$$

$$13 \times 37 = 481 = 8 \times 60 + 1 \pmod{60} = 1 \pmod{60}$$

Since $37 + 23 = 60 \pmod{60}$, $37 = -23 \pmod{60}$

1723

101
3
5
67
11
2
7
13
71
17

Exercises

1. If $p|10a - b, p|10c - d$, then $p|ad - bc$
2. if n is odd, then $3|2^n + 1$
3. $k = 0,1,2, \dots$ for $n \in \mathbb{Z}$, we have $2n + 1|1^{2k+1} + 2^{2k+1} + \dots + (2n)^{2k+1}$
4. if $m - p|mn + pq$ then $m - p|mq + np$
5. if $x \equiv 1 \pmod{m^k}$ then $x^m \equiv 1 \pmod{m^{k+1}}$

Euler Totient Function $\phi(n)$

- when doing arithmetic modulo n
- **complete set of residues** is: 0..n-1
- **reduced set of residues** is those numbers (residues) which are relatively prime to n
 - E.g., for n=10,
 - complete set of residues is {0,1,2,3,4,5,6,7,8,9}
 - reduced set of residues is {1,3,7,9}
- number of elements in reduced set of residues is called the **Euler Totient Function $\phi(n)$**



Euler

Euler Totient Function $\phi(n)$ (2)

- to compute $\phi(n)$ need to count number of residues to be excluded
- in general need prime factorization, but
 - for p (p prime) $\phi(p) = p-1$
 - for p,q (p,q prime)
 $\phi(pq) = \phi(p) \times \phi(q) = (p-1) \times (q-1)$
- eg.
 - $\phi(37) = 36, \phi(11) = 10$
 - $\phi(21) = (3-1) \times (7-1) = 2 \times 6 = 12,$
 - $\phi(10) = (2-1) \times (5-1) = 1 \times 4 = 4 \quad \{1,3,7,9\}$

Fermat's Theorem

- $a^{p-1} = 1 \pmod{p}$
 - where p is prime and $\gcd(a,p)=1$
- also known as Fermat's Little Theorem
- also $a^p = a \pmod{p}$
- useful in public key and primality testing
- $\phi(p)=p-1$



Fermat

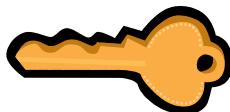
Euler's Theorem

- a generalisation of Fermat's Theorem
- $a^{\phi(n)} = 1 \pmod{n}$
 - for any a, n where $\gcd(a, n) = 1$
- eg.
 - $a=3; n=10; \phi(10)=4;$
hence $3^4 = 81 = 1 \pmod{10}$
 - $a=2; n=11; \phi(11)=10;$
hence $2^{10} = 1024 = 1 \pmod{11}$

Public Key Cryptography

- **Public Key/Asymmetric** cryptography involves the use of **two keys**:

➤ **public-key**, which may be known by anybody, and can be used to **encrypt messages**, and **verify signatures**

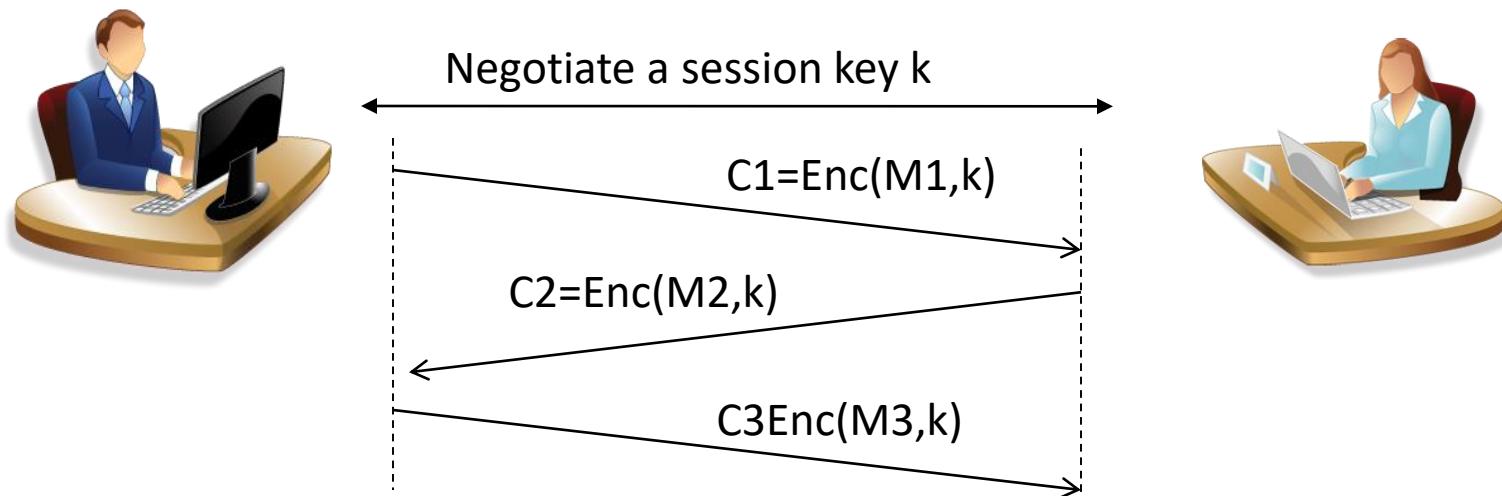


➤ **private-key**, known only to the recipient, used to **decrypt messages**, and **sign (create) signatures**



Applications of Public Key Cryptography

- Public-Key algorithms
 - encryption/decryption (provide confidentiality)
 - digital signatures (provide authentication)
 - key exchange (of session keys)



RSA Encryption

P & Q PRIME
N = PQ
 $ED \equiv 1 \pmod{(P-1)(Q-1)}$
 $C = M^e \pmod{N}$
 $M = C^d \pmod{N}$
RSA Algorithm



ACM Turing Award 2002

RSA Encryption

- p and q are large primes.
- $n = p \cdot q$
- $\phi(n) = (p - 1)(q - 1)$
- e is an integer, $1 < e < \phi(n)$, $\gcd(e, \phi(n)) = 1$
- d is an integer, $1 < d < \phi$, $e \cdot d \equiv 1 \pmod{\phi}$
(d is inverse of e in modulo $\phi(n)$, calculated by the extended Euclidean algorithm)
- (n, e) is public key
- (n, d) is private key

RSA Encryption (Cont.)

- Encryption : $c=m^e \text{ mod } n$
- Decryption : $m=c^d \text{ mod } n$
- m is plaintext, c is ciphertext.

$$\begin{aligned}m &= c^d \text{ mod } n = m^{ed} \text{ mod } n \\&= m^{1+k\varphi(n)} \text{ mod } n = m \cdot (m^{\varphi(n)})^k \text{ mod } n \\&= m \cdot 1 \text{ mod } n = m\end{aligned}$$

RSA Example



$$m = 15$$

$$C = m^e \bmod n = 15^{13} \bmod 77 = 64$$

$$\begin{aligned} PK : & (n, e) \\ SK : & d \end{aligned}$$

C

$$\begin{aligned} p &= 7, q = 11, n = pq = 7 \times 11 = 77 \\ \phi(n) &= (p-1)(q-1) = 6 \times 10 = 60 \\ e &= 13, d = ? \\ \therefore \gcd(60, 13) &= 1 \end{aligned}$$

$$\begin{aligned} m &= C^d \bmod n = 64^{37} \bmod 77 = 15 \\ &= 15^{13 \times 37} \bmod 77 = 15^{1+8 \times 60} \bmod 77 = 15 \end{aligned}$$

Dividend	Divisor	Quotient	Remainder
$\phi(n) = 60$	$= e = 13$	$\times 4$	$+ 8$
$e = 13$	$= 8$	$\times 1$	$+ 5$
8	$= 5$	$\times 1$	$+ 3$
5	$= 3$	$\times 1$	$+ 2$
3	$= 2$	$\times 1$	$+ 1$

$$\begin{aligned} 1 &= 3 - 2 \times 1 \\ 1 &= 3 - (5 - 3 \times 1) \times 1 = 3 \times 2 - 5 \times 1 \\ 1 &= (8 - 5 \times 1) \times 2 - 5 \times 1 = 8 \times 2 - 5 \times 3 \\ 1 &= 8 \times 2 - (13 - 8 \times 1) \times 3 = 8 \times 5 - 13 \times 3 \\ 1 &= (60 - 13 \times 4) \times 5 - 13 \times 3 = 60 \times 5 - 13 \times 23 \\ 1 &= 60 \times 5 - 13 \times 23 \bmod 60 = -13 \times 23 \bmod 60 \\ 1 &= 13 \times (-23) = 13 \times 37 \bmod 60 \end{aligned}$$

Since $37 + 23 = 60 \bmod 60$, $37 = -23 \bmod 60$

$d = 37$

Security of RSA Cryptosystem

- Since n and e are in public key, in order to recover $m=c^d \bmod n$ from c , we need to know the private key d .
- While in order to compute d , we need to factorize $n=p \cdot q$ to compute $\phi(n)=(p-1)(q-1)$.
- Therefore, the security of RSA cryptosystem is dependent upon the hardness of factor large integer.
- Currently, 1024 bit-RSA is commonly used.
- For other applications, like military application, 4096 bit-RSA is preferred.

The modulus $n=p*q$ can not be common in the system

Why

- Suppose two entities share the common modulus $n=p*q$, but have different public-private key pairs (e_1, d_1) and (e_2, d_2) with $\gcd(e_1, e_2)=1$. If the same message m is encrypted with $c_1=m^{e_1} \bmod n$, $c_2=m^{e_2} \bmod n$, then the message m can be recovered without knowing the private keys.
- Since $\gcd(e_1, e_2)=1$, we can use the extended Euclidean algorithm to calculate r, s such that $r*e_1+s*e_2=1$.
- Then, $c_1^r * c_2^s = m^{r*e_1+s*e_2} = m \bmod n$, we can get m
- If $r < 0$, we should calculate $(c_1^{-1})^{-r} * c_2^s = m \bmod n$

Example

- If $p=7$, $q=11$, $n=p*q=77$, $\phi(n)=(p-1)(q-1)=60$.
- $(e_1=13, d_1=37)$, $(e_2=17, d_2=53)$ extended Euclidean algorithm
- For $m=15$, $c_1=m^{e_1} \text{ mod } n=15^{13} \text{ mod } 77=64$, $c_2=m^{e_2} \text{ mod } n=15^{17} \text{ mod } 77=71$
- Since $\gcd(e_1, e_2)=1$, we have $4*13 + (-3)*17 = 1$ extended Euclidean algorithm
- Because $s=-3$, we compute $(c_2)^{-1} \text{ mod } 77=-13 \text{ mod } 77=64$ extended Euclidean algorithm $-13*71 + 12*77 = 1$
- $c_1^{e_2} * (c_2^{-1})^s = 64^4 * 64^3 \text{ mod } 77 = 15$

common modulus attack

Chinese Remainder Theorem (CRT)

- CRT is used to speed up modulo computations
- Let m_1, m_2, \dots, m_k be pairwise relatively prime integers, i.e., $\gcd(m_i, m_j) = 1$ for $1 \leq i < j \leq k$. Let $a_i \in \mathbb{Z}_{m_i}$ for $1 \leq i \leq k$ and set $M = m_1 m_2 \dots m_k$. There exists a unique $A \in \mathbb{Z}_M$ such that $A \equiv a_i \pmod{m_i}$ for $i = 1 \dots k$.
- A can be computed as:

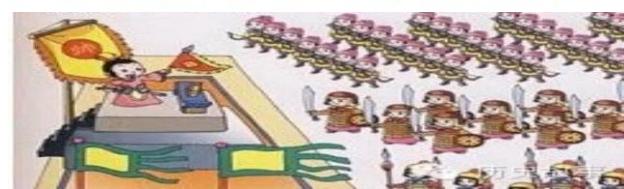
$$A \equiv \left(\sum_{i=1}^k a_i c_i \right) \pmod{M}$$

- Where, for $1 \leq i \leq k$,

$$c_i = M_i \times (M_i^{-1} \pmod{m_i}) \text{ and } M_i = \frac{M}{m_i}$$



中國剩餘定理（韓信點兵）



CRT Proof

- Since $M_i = \frac{M}{m_i} = \frac{m_1 \times m_2 \times \cdots \times m_k}{m_i} = m_1 \times m_2 \times \cdots \times m_{i-1} \times m_{i+1} \times \cdots \times m_k$

$$c_i = [M_i \times (M_i^{-1} \bmod m_i)] \bmod m_i = [M_i \bmod m_i \times (M_i^{-1} \bmod m_i)] \bmod m_i = 1 \bmod m_i$$

$$c_i = [M_i \times (M_i^{-1} \bmod m_i)] \bmod m_j$$

$$= [m_1 \times m_2 \times \cdots \times m_{i-1} \times m_{i+1} \times \cdots \times m_k \times (M_i^{-1} \bmod m_i)] \bmod m_j = 0$$

- We have

$$c_i = M_i \times (M_i^{-1} \bmod m_i) = \begin{cases} 1 \bmod m_i \\ 0 \bmod m_j \text{ where } j \neq i \end{cases}$$

- Therefore,

$$A \equiv \left(\sum_{i=1}^k a_i c_i \right) \bmod M$$

$$= a_1 c_1 + a_2 c_2 + \cdots + a_k c_k + l \times M = a_i \bmod m_i$$

- where $1 \leq i \leq k$

CRT Proof

- A is unique in Z_M
- Suppose A is not the unique solution, there must exist another answer $A' \equiv a_i \pmod{m_i}$ in Z_M .

$$\begin{cases} A \equiv a_1 \pmod{m_1} \\ \vdots \\ A \equiv a_i \pmod{m_i} \\ \vdots \\ A \equiv a_k \pmod{m_k} \end{cases}$$

$$\begin{cases} A' \equiv a_1 \pmod{m_1} \\ \vdots \\ A' \equiv a_i \pmod{m_i} \\ \vdots \\ A' \equiv a_k \pmod{m_k} \end{cases}$$

- Then $A \equiv A' \pmod{m_i}$
 - $A - A' = r_1 * m_1 = r_2 * m_2 = \dots r_k * m_k$
 - $r_i | m_j$ where $i \neq j$ (since m_i 's are relatively prime)
 - $m_1 | A - A', m_2 | A - A', \dots, m_k | A - A'$ ----(1)
 - $M | A - A'$. Suppose $M \nmid A - A'$, there exist at least one $m_i \nmid A - A'$, which contradicts with (1). Therefore,
 - $M | A - A' \Rightarrow A \equiv A' \pmod{M}$, which shows A is unique in Z_M

Example

- Find a number x such that have remainders of 1 when divided by 3, 2 when divided by 5 and 3 when divided by 7. i.e.

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

- First, we know $M=3*5*7=105$, $M_1=M/m_1=35$, $M_2=M/m_2=21$, $M_3=M/m_3=15$

$$\begin{aligned} x &= \sum_{i=1}^3 a_i \cdot [M_i \cdot (M_i^{-1} \pmod{m_i})] \pmod{M} \\ &= 1 \times 35 \times 2 + 2 \times 21 \times 1 + 3 \times 15 \times 1 \pmod{105} \\ &= 52 \end{aligned}$$

CRT Speed Up RSA Decryption

- p and q are large primes.
 - $n = p \cdot q$ $\varphi(n) = (p - 1)(q - 1)$
 - e is an integer, $1 < e < \varphi(n)$, $\gcd(e, \varphi(n)) = 1$
 - d is an integer, $1 < d < \varphi$, $e \cdot d \equiv 1 \pmod{\varphi}$
 - (n, e) is public key (n, d) is private key
 - Encryption : $c = m^e \pmod{n}$
 - Decryption : $m = c^d \pmod{n}$
 - m is plaintext, c is ciphertext.
- $p=179, q=197$
 - $n = 35263 \quad \varphi(n) = 34888$
 - $e = 17, d = 8209$
 - $d_1 = d \pmod{p-1} = 8209 \pmod{178} = 21$
 - $d_2 = d \pmod{q-1} = 8209 \pmod{196} = 173$
 - $c = m^e \pmod{n} = 168^{17} \pmod{35263} = 28657$
 - $m = c^d \pmod{n} = 28657^{8209} \pmod{35263} = 168$

$$c = 28657 \quad c \pmod{p} = 28657 \pmod{179} = 17$$

$$c \pmod{q} = 28657 \pmod{197} = 92$$

$$\begin{cases} m \equiv c^{d_1} \equiv 17^{21} \pmod{179} \equiv 168 \pmod{179} \\ m \equiv c^{d_2} \equiv 92^{173} \pmod{197} \equiv 168 \pmod{197} \end{cases}$$

$$\begin{cases} c \equiv 17 \pmod{179} \\ c \equiv 92 \pmod{197} \end{cases}$$

CRT Speed Up RSA Decryption

$$\begin{aligned} c &= 28657 \quad c \bmod p = 28657 \bmod 179 = 17 \\ &\quad c \bmod q = 28657 \bmod 197 = 92 \end{aligned} \quad \left\{ \begin{array}{l} c \equiv 17 \pmod{179} \\ c \equiv 92 \pmod{197} \end{array} \right.$$
$$\left\{ \begin{array}{l} m \equiv c^{d_1} \equiv 17^{2^1} \pmod{179} \equiv 168 \pmod{179} \\ m \equiv c^{d_2} \equiv 92^{17^3} \pmod{197} \equiv 168 \pmod{197} \end{array} \right. \xrightarrow{\text{Coincidence}}$$

$$m_1 = 179, m_2 = 197, M = m_1 * m_2 = 35263$$

$$M_1 = M / m_1 = 197, M_2 = M / m_2 = 179$$

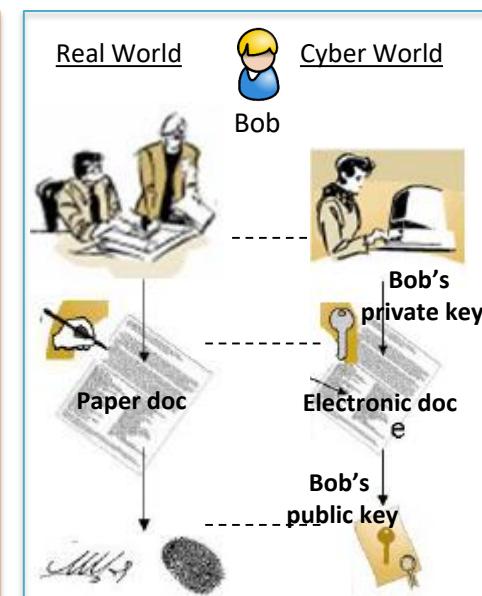
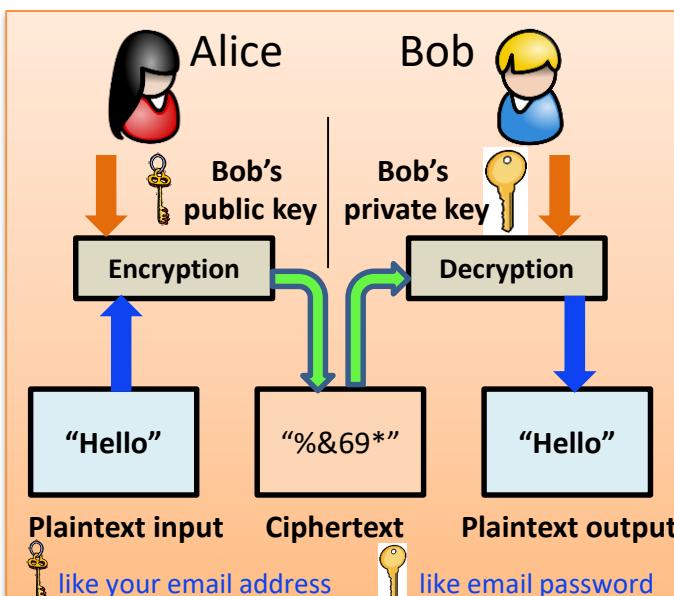
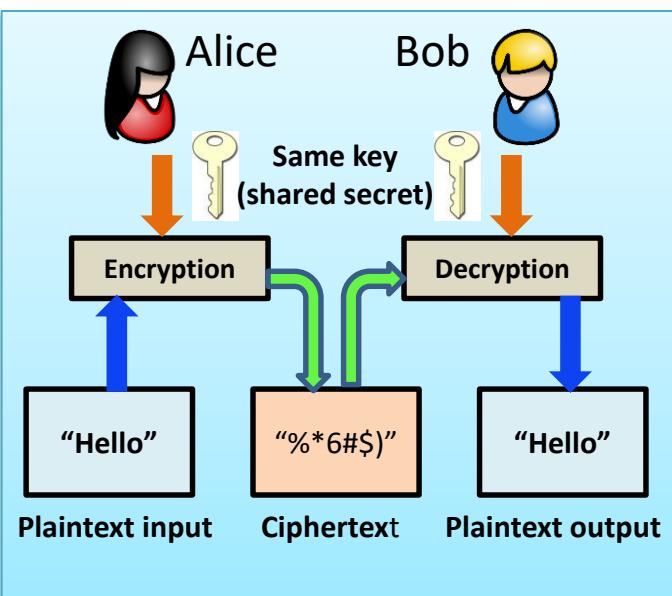
$$M_1^{-1} \pmod{m_1} = 197^{-1} \pmod{179} = 10$$

$$M_2^{-1} \pmod{m_2} = 179^{-1} \pmod{197} = 186$$

$$x = \sum_{i=1}^3 a_i \cdot [M_i \cdot (M_i^{-1} \pmod{m_i})] \pmod{M}$$

$$\begin{aligned} &= 168 \times 197 \times 10 + 168 \times 179 \times 186 \pmod{35263} \\ &= 168 \end{aligned}$$

Comparison Symmetric Encryption and Public Key Encryption



Symmetric Encryption (AES, DES, ...)

- Fast
- Need share the same key
- Achieve Confidentiality

Public key Encryption (RSA)

- Slow (due to exponential operation)
- Do not need share the same key
- Achieve Confidentiality

Digital Signature

- + Use private key to sign
- + Use public key to verify
- + Achieve authentication, data integrity, non repudiation

RSA Digital Signature



Hash function

$$H(.)$$

$$PK : (n, e)$$

$$SK : d$$



$$m$$

Signing:

$$\sigma = H(m)^d \bmod n$$

$$(m, \sigma)$$

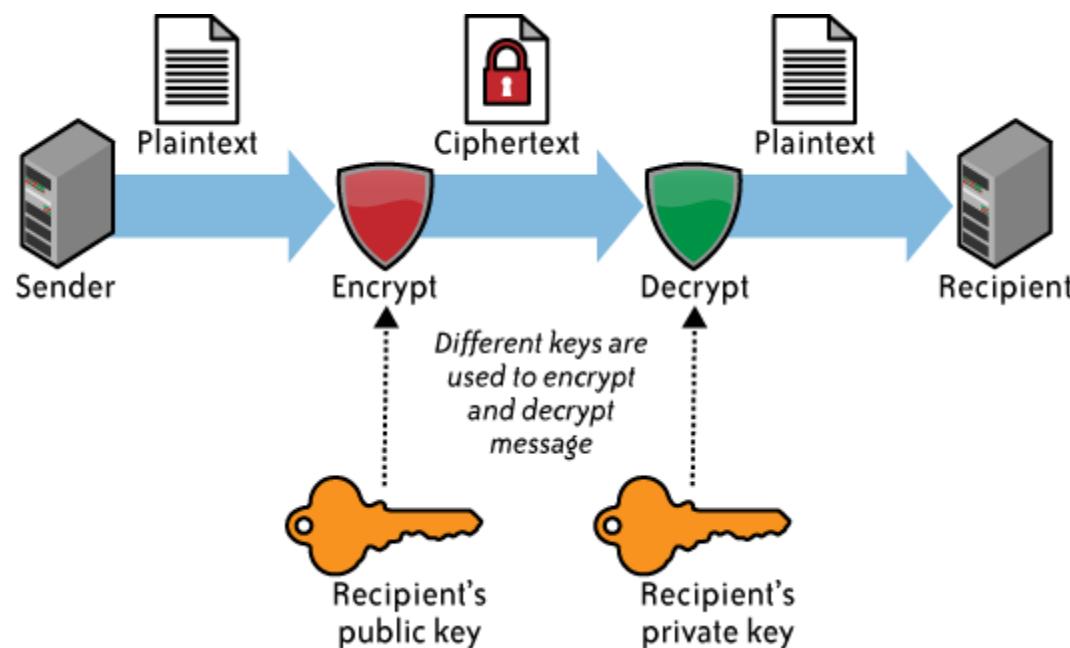
Verification:

$$\sigma^e = ? = H(m) \bmod n$$

$$\sigma^e \bmod n = H(m)^{ed} \bmod n = H(m)^{1+k\varphi(n)} \bmod n$$

$$= H(m) \cdot (H(m)^{\varphi(n)})^k \bmod n = H(m) \bmod n$$

ElGamal Encryption



Taher Elgamal

Taher El Gamal. 1985. A public key cryptosystem and a signature scheme based on discrete logarithms. In *Proceedings of CRYPTO 84 on Advances in cryptology*, G R Blakley and David Chaum (Eds.). Springer-Verlag New York, Inc., New York, NY, USA, 10-18.

Primary Root

- Consider $Z_7^* = \{1, 2, 3, 4, 5, 6\}$, we know $3^1 \equiv 3 \pmod{7}$, $3^2 \equiv 2 \pmod{7}$, $3^3 \equiv 6 \pmod{7}$, $3^4 \equiv 4 \pmod{7}$, $3^5 \equiv 5 \pmod{7}$, $3^6 \equiv 1 \pmod{7}$. We obtain all the nonzero elements modulo 7 as powers of 3. This means that 3 is a primitive root modulo 7.
- However, $3^3 \equiv 1 \pmod{13}$, so only 1, 3, 9 are the powers of 3. Therefore, 3 is not a primitive root modulo 13.
- In general, when p is a prime, a primitive root modulo p is a number α whose powers yield every nonzero elements modulo p .

$$\alpha^m \equiv n \pmod{p}, \text{ for } 1 \leq n < p$$
- There are $\phi(p-1)$ primitive root modulo p .

Discrete Logarithm Problem (DLP)

- Let p a large prime, g is primitive root of \mathbb{Z}_p^*
- DLP: Given p, g and $(g^a \bmod p)$, determine a .
- DLP in \mathbb{Z}_p^* is hard, that is, there is no an efficient algorithm which can determine a in a polynomial time.

Given $(p, g, x \in [1, p-1])$, it is easy to compute

$$Y = g^x \bmod p$$

However, given $(p, g, Y = g^x \bmod p)$, it is hard to compute

$$x \text{ such that } Y = g^x \bmod p$$

ElGamal Encryption

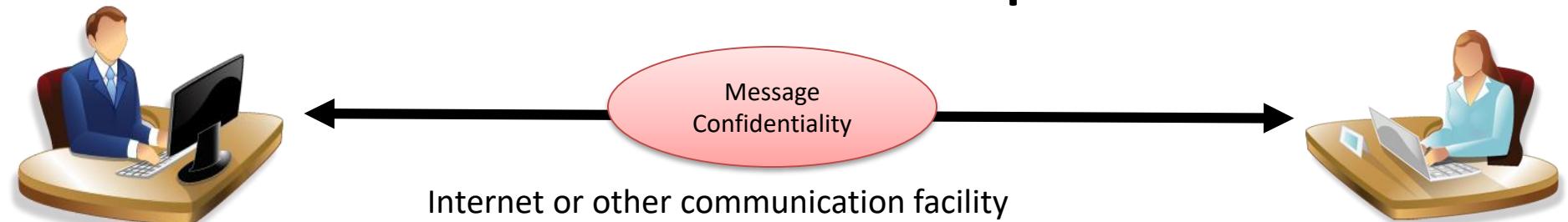
- p is a large prime, g is a generator (primitive root) of \mathbb{Z}_p^*
- x is a random number in $[1, p-1]$
- $Y = g^x \text{ mod } p$
- (p, g) are public parameters
- Y is a public key
- x is a private key (Because of the hardness of DLP, it is hard to get the private key x from the public key Y)

ElGamal Encryption (Cont.)

- Encryption :
 - Choose a random number k in $(1, p-1)$, and compute $K=Y^k \bmod p$
 - Encrypt a message M in Z_{p^*} as $C=(C_1, C_2)$, where ($C_1=g^k \bmod p$, $C_2=M*K \bmod p$)
- Decryption :
 - Use the private key x to compute $C_1^x=g^{kx}=Y^k=K \bmod p$
 - Recover $M=C_2/K \bmod p$

Because of the hardness of DLP, it is hard to get x , k from $Y=g^x \bmod p$ and $K=Y^k \bmod p$

ElGamal Example



$$m = 7$$

$$p = 13, g = 2$$

$$SK : x = 5$$

$$PK : Y = g^x \bmod 13 = 2^5 \bmod 13 = 6$$

$$k = 4, K = Y^k \bmod 13 = 6^4 \bmod 13 = 9$$

$$C_1 = g^k \bmod 13 = 2^4 \bmod 13 = 3$$

$$C_2 = m \cdot K \bmod 13 = 7 \cdot 9 \bmod 13 = 11$$

$$C = (C_1, C_2) = (3, 11)$$

In real system, p
should be large

$$C_1^x \bmod 13 = 3^5 \bmod 13 = 9 = K$$

$$\begin{aligned} m &= C_2 / K \bmod 13 = 11 \times 9^{-1} \bmod 13 \\ &= 11 \times 3 \bmod 13 = 7 \end{aligned}$$

A.M. TURING CENTENARY CELEBRATION WEBCAST

[MORE ACM AWARDS](#)

Search

TYPE HERE



A.M. TURING AWARD WINNERS BY...

[ALPHABETICAL LISTING](#)[YEAR OF THE AWARD](#)[RESEARCH SUBJECT](#)

2015 AWARD WINNERS:

**Whitfield Diffie and
Martin Hellman**

Cryptography Pioneers Receive 2015 ACM A.M. Turing Award

Whitfield Diffie, former Chief Security Officer of Sun Microsystems and Martin E. Hellman, Professor Emeritus of Electrical Engineering at Stanford University, are the recipients of the 2015 ACM A.M. Turing Award, for critical contributions to modern cryptography. The ability for two parties to communicate privately over a secure channel is fundamental for billions of people around the world. On a daily basis, individuals establish secure online

Diffie-Hellman Key Exchange



Ralph Merkle, Martin Hellman, Whitfield Diffie (1977)



p: a large prime, g: a generator of order p-1

$$(p, g)$$



Choose a random number

$$x \in \{1, \dots, p-1\}$$

Compute

$$g^x \bmod p$$

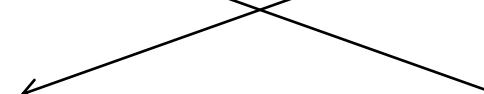


Choose a random number

$$y \in \{1, \dots, p-1\}$$

Compute

$$g^y \bmod p$$



$$K = (g^y)^x = g^{xy} \bmod p$$

$$K = (g^x)^y = g^{xy} \bmod p$$

D-H Key Exchange Example



$$(p = 11, g = 7)$$



Choose a random number

$$x = 6$$

Compute

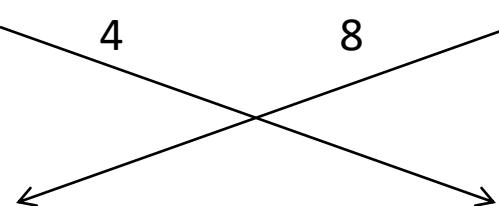
$$g^x \bmod p = 7^6 \bmod 11 = 4$$

Choose a random number

$$y = 9$$

Compute

$$g^y \bmod p = 7^9 \bmod 11 = 8$$



$$K = (g^y)^x \bmod p = 8^6 \bmod 11 = 3$$

$$K = (g^x)^y \bmod p = 4^9 \bmod 11 = 3$$

$$K = g^{xy} \bmod p = 7^{6 \times 9} \bmod 11 = 3$$

Is D-H Key Exchange Secure?

p: a large prime, g: a generator of order p-1



$$(p, g)$$

Choose a random number

$$x \in \{1, \dots, p-1\}$$

Compute

$$g^x \bmod p$$



Choose a random number

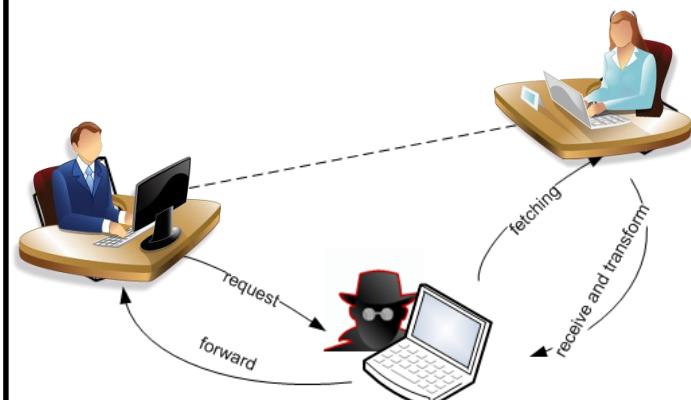
$$y \in \{1, \dots, p-1\}$$

Compute

$$g^y \bmod p$$

$$K = (g^y)^x = g^{xy} \bmod p$$

$$K = (g^x)^y = g^{xy} \bmod p$$



Man-in-the-Middle attack

- When p is a large prime, the Discrete Logarithm Problem is hard. Therefore, given $g^x \bmod p$, $g^y \bmod p$, the adversary cannot compute x, y. Then, the key $K=g^{xy} \bmod p$ is secure.
- However, the key is not authenticated. (Man-in-the-Middle attack)

Man-In-The-Middle Attack

p: a large prime, g: a generator of order p-1

(p, g)



Choose a random number

$$x \in \{1, \dots, p-1\}$$

Compute

$$g^x \bmod p$$



Choose two random numbers

$$y', x' \in \{1, \dots, p-1\}$$



Choose a random number

$$y \in \{1, \dots, p-1\}$$

Compute

$$g^y \bmod p$$

$$g^{y'} \bmod p, g^{x'} \bmod p$$

$$K_1 = (g^{y'})^x = g^{xy'} \bmod p$$

$$K_2 = (g^{x'})^y = g^{x'y} \bmod p$$

$$K_1 = (g^x)^{y'} = g^{xy'} \bmod p$$

$$K_2 = (g^y)^{x'} = g^{x'y} \bmod p$$

M

$$C_1 = AES(M, K_1)$$

K₁

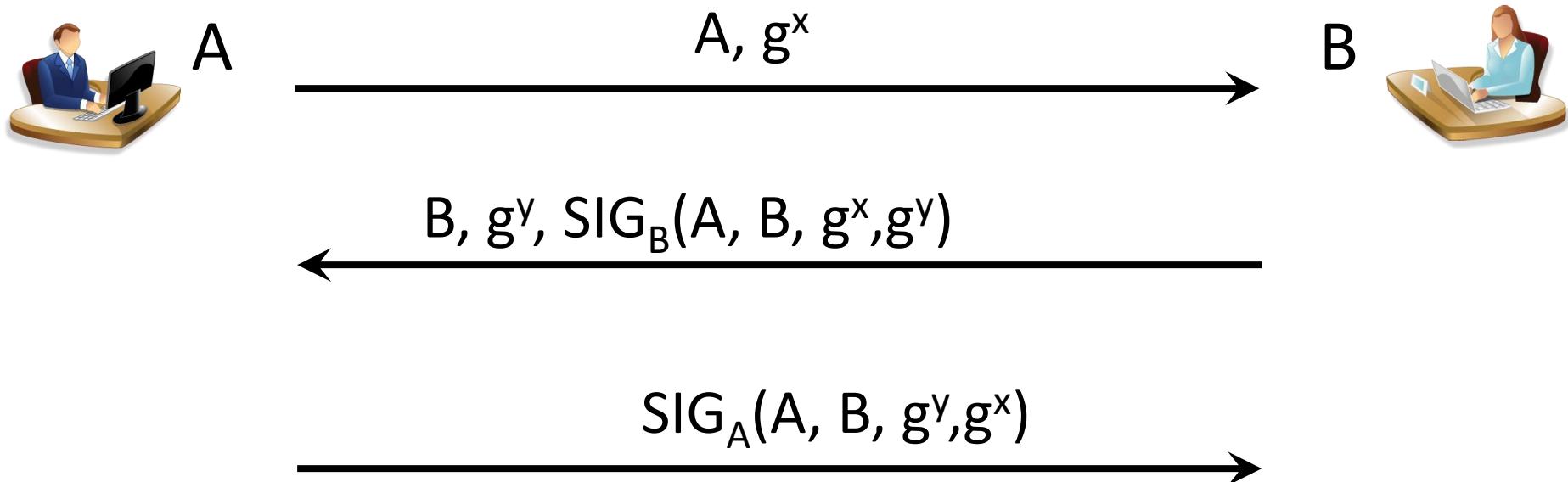
M

$$C_2 = AES(M, K_2)$$

K₂

M

Authenticated D-H Key Exchange



Each party signs its own DH value to prevent man-in-the-middle attack (and the peer's DH value as a freshness guarantee against replay)

A: "Shared K= g^{xy} with B" ($K \leftrightarrow B$) B: "Shared K= g^{xy} with A" ($K \leftrightarrow A$)



Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 4: Homomorphic cryptographic techniques
for privacy

Lecturer: Rongxing LU

Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

How to achieve data privacy in cloud, dropbox?

- One solution
 - Encrypt data + Computing on Encrypted Data



Computing on Encrypted Data

- It would be very nice to
 - Encrypt my data in the cloud
 - While still allowing the cloud to search/sort/edit/... this data on my behalf
 - Keeping the data in the cloud in encrypted form
 - Without needing to ship it back and forth to be decrypted
 - Encrypt my queries to the cloud
 - While still allowing the cloud to process them
 - Cloud returns encrypted answers
 - that I can decrypt



Computing on Encrypted Data

Directions

- From: 19 Skyline Drive,
Hawothorne, NY 10532,
USA
- To: Columbia University



@ ...



Computing on Encrypted Data



typo

(A) Did you mean:

1 19 Skyline Dr.
19 Skyline Dr, Hawthorne, NY 10532

These directions are for planning purposes only. You may find that construction projects, traffic, weather, or other events may cause conditions to differ from the map results, and you should plan your route accordingly. You must obey all signs or notices regarding your route.

Map data ©2010

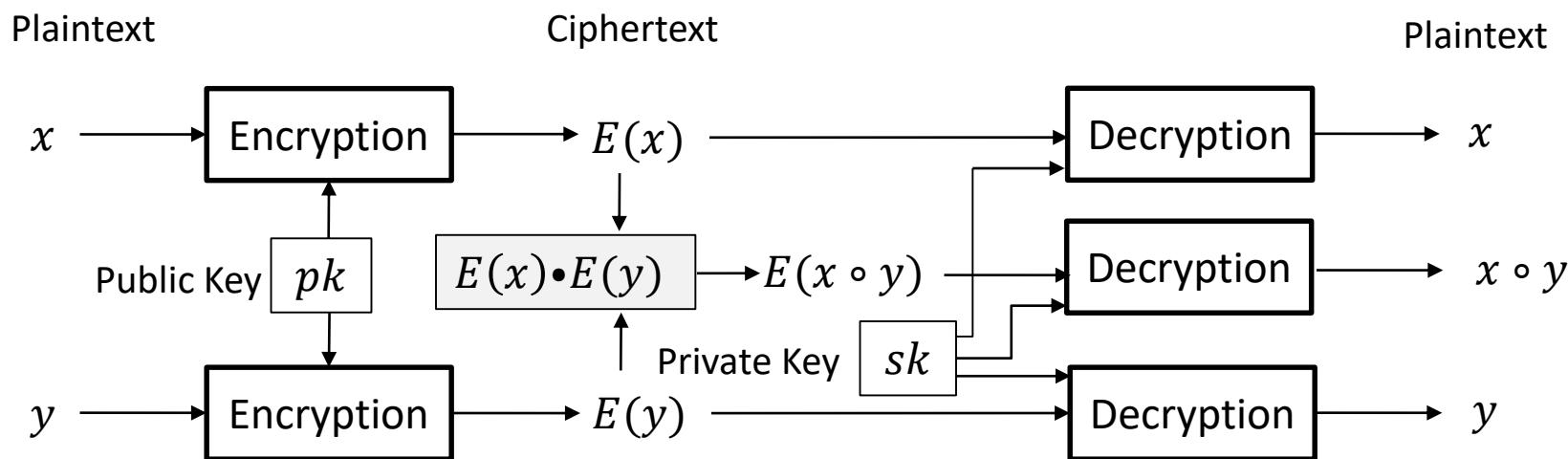


\$kjh9*mslt@na0
&maXxjq02bflx
m^00a2nm5,A4.
pE.abxp3m58bsa
(3saM%w,snanba
nq~mD=3akm2,A
Z,ltnhde83|3mz{nd
dewiunb4]gnbTa*
kjew^bwJ^mdns0



How to achieve computing on Encrypted Data?

- One Solution
 - Homomorphic Encryption



RSA Is A Multiplication Homomorphic Encryption

- RSA Encryption
 - Key Generation: Public Key (n, e) , Private Key d
 - Encrypt: $c = m^e \bmod n$
 - Decrypt: $m = c^d \bmod n$
- Multiplication Homomorphic Property
 - $c_1 = m_1^e \bmod n, c_2 = m_2^e \bmod n,$
 - $E(m_1) \cdot E(m_2) = c_1 \cdot c_2 = (m_1 \cdot m_2)^e \bmod n = E(m_1 \cdot m_2)$

ElGamal Is A Multiplication Homomorphic Encryption

- ElGamal Encryption
 - Key Generation: Public Key $y = g^x \text{ mod } p$, Private Key x
 - Encrypt: $c = g^r \text{ mod } p$, $\tilde{c} = m \cdot y^r \text{ mod } p$
 - Decrypt: $m = \frac{\tilde{c}}{c^x} \text{ mod } p$
- Multiplication Homomorphic Property
 - $E(m_1) : c_1 = g^{r_1} \text{ mod } p$, $\tilde{c}_1 = m_1 \cdot y^{r_1} \text{ mod } p$
 - $E(m_2) : c_2 = g^{r_2} \text{ mod } p$, $\tilde{c}_2 = m_2 \cdot y^{r_2} \text{ mod } p$
 - $E(m_1) \cdot E(m_2) : \rightarrow c_1 \cdot c_2 = g^{r_1+r_2} \text{ mod } p$, $\tilde{c}_1 \cdot \tilde{c}_2 = m_1 \cdot m_2 \cdot y^{r_1+r_2} \text{ mod } p : \rightarrow E(m_1 \cdot m_2)$

Paillier Is An Addition Homomorphic Encryption

- Pallier Encryption
 - Key Generation
 - Input: two prime numbers p, q
 - Compute $n = pq, \lambda = \text{lcm}(p - 1, q - 1)$
 - Choose $g \in \mathbb{Z}_{n^2}^*$ such that $\gcd(L(g^\lambda \bmod n^2), n) = 1$ with $L(u) = \frac{u-1}{n}$,
compute $\mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$
 - Output: public key $pk = (n, g)$ private key $sk = (\lambda, \mu)$
 - Encrypt: Input $m \in \mathbb{Z}_n$
 - Choose $r \in \mathbb{Z}_n^*$, Compute $c = g^m \cdot r^n \bmod n^2$
 - Decrypt: Input $c \in \mathbb{Z}_{n^2}$
 - Compute $m = L(c^\lambda \bmod n^2) \cdot \mu \bmod n$
- Addition Homomorphic Encryption
 - $E(m_1, r_1) : c_1 = g^{m_1} \cdot r_1^n \bmod n^2$
 - $E(m_2, r_2) : c_2 = g^{m_2} \cdot r_2^n \bmod n^2$
 - $E(m_1, r_1) \cdot E(m_2, r_2) = c_1 \cdot c_2 = g^{m_1+m_2} \cdot (r_1 r_2)^n \bmod n^2 = E(m_1 + m_2, r_1 r_2)$

Boneh-Goh-Nissim (BGN) Is An Addition Homomorphic Encryption with Limited Multiplication Homomorphic Property

- Subgroup decision problem
 - Given $h \in G$ decide whether h has order q or order n
 - Subgroup decision assumption:
 \forall poly-time Adv:

$$\Pr[(p, q, G, G_T, e, g) \leftarrow \text{Gen}(1^k); h \leftarrow G^* : \text{Adv}(n, G, G_T, e, g, h) = 1] \\ \approx \Pr[(p, q, G, G_T, e, g) \leftarrow \text{Gen}(1^k); h \leftarrow G_q^* : \text{Adv}(n, G, G_T, e, g, h) = 1]$$

Boneh-Goh-Nissim (BGN) Is An Addition Homomorphic Encryption with Limited Multiplication Homomorphic Property

- Composite order bilinear group
 - Gen(1^k) generates (p, q, G, G_T, e, g)
 - G, G_T finite cyclic groups of order $n = pq$
 - G has generator g
 - Pairing $e: G \times G \rightarrow G_T$
 - $e(g, g)$ generates G_T
 - $e(g^a, g^b) = e(g, g)^{ab}$
 - Deciding group membership, group operations, and bilinear pairing efficiently computable

Boneh-Goh-Nissim (BGN) Is An Addition Homomorphic Encryption with Limited Multiplication Homomorphic Property

- BGN Encryption:
 - KeyGen:
 - Public key: $g, h \in G$ h order q
 - Secret key: p, q $n = pq$
 - Encrypt: $c = g^m h^r$ $r \leftarrow \mathbb{Z}_n$
 - Decrypt: $c^q = (g^m h^r)^q = g^{qm} h^{qr} = (g^q)^m$
 - **Note: m is in a small message space according to the concrete scenarios.**
- Addition Homomorphic Encryption
 - $E(m_1, r_1) = c_1 = g^{m_1} h^{r_1}$ $E(m_2, r_2) = c_2 = g^{m_2} h^{r_2} \rightarrow c_1 * c_2 = g^{m_1+m_2} h^{r_1+r_2} = E(m_1+m_2, r_1+r_2)$
- Multiplication Homomorphic Encryption
 - $E(m_1, r_1) = c_1 = g^{m_1} h^{r_1}$ $E(m_2, r_2) = c_2 = g^{m_2} h^{r_2}$
 - $e(E(m_1, r_1), E(m_2, r_2)) = e(g, g)^{m_1 m_2} e(g, g)^{m_1 r_2 + r_1 m_2 + q r_1 r_2} = E(m_1 m_2, m_1 r_2 + r_1 m_2 + q r_1 r_2)$

(x,+)-Homomorphic Encryption

It will be really nice to have...

- Plaintext space \mathbb{Z}_2 (with operations +,x)
- Ciphertext space some ring \mathcal{R} (with operations +,x)
- Homomorphic for both + and x
 - $\text{Enc}(x_1) + \text{Enc}(x_2)$ in $\mathcal{R} = \text{Enc}(x_1 + x_2 \bmod 2)$
 - $\text{Enc}(x_1) \times \text{Enc}(x_2)$ in $\mathcal{R} = \text{Enc}(x_1 \times x_2 \bmod 2)$
- Then we can compute any function on the encryptions
 - Since every binary function is a polynomial

Details of Paillier Homomorphic Encryption



Pascal Paillier:

Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. [EUROCRYPT 1999](#): 223-238

Mathematical Background of Paillier

- Let $p = 2p' + 1$ and $q = 2q' + 1$ be two safe primes, where p' and q' are also two primes. Compute $n = p q$, we need to prove the following two results:
 - for any $x \in \mathbb{Z}_n$, we have $(1 + n)^x = 1 + x \cdot n \bmod n^2$
 - let $\lambda = \text{lcm}(p - 1, q - 1) = 2p'q'$ be the least common multiple of $p - 1$ and $q - 1$. For any $x \in \mathbb{Z}_{n^2}^*$, we have $x^{n\lambda} = 1 \bmod n^2$

For the first result, we use the following theorem to prove it.

- For any $x \in Z_n$, we have $(1 + n)^x = 1 + x \cdot n \text{ mod } n^2$.
- When $x = 0$, the result obviously holds.
- When $0 < x < n$, we have

$$\begin{aligned}(1 + n)^x &= \sum_{i=0}^x \binom{x}{i} 1^{n-i} \cdot n^i \text{ mod } n^2 \\&= 1 + n \cdot x + \binom{x}{2} n^2 + \dots + \binom{x}{x} n^x \text{ mod } n^2 \\&= 1 + n \cdot x \text{ mod } n^2\end{aligned}$$

For the second result, we should use some lemma and theorem as below.

- **(Euler Totient Function)** Let $P = \prod_i p_i^{l_i}$ with p_i pairwise different primes and $l_i > 0$. Then, Euler Totient Function is defined as

$$\phi(P) = P \cdot \prod_i \left(1 - \frac{1}{p_i}\right)$$

- **Lemma**

$$\phi(p) = p - 1,$$

$$\phi(q) = q - 1,$$

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

$$\phi(p^2) = p\phi(p), \phi(q^2) = q\phi(q), \phi(n^2) = n\phi(n).$$

- According to the Euler Theorem, for any $x \in Z_{n^2}^*$, we have $x^{\phi(n^2)} = x^{n\phi(n)} = x^{n \cdot 2\lambda} = 1 \text{ mod } n^2$, but we still cannot determine whether $x^{n\lambda} = 1 \text{ mod } n^2$. In order to obtain $x^{n\lambda} = 1 \text{ mod } n^2$, we need to use the result of the Chinese Remainder Theorem.

Theorem (Chinese Remainder Theorem). Suppose that m_1, m_2, \dots, m_k are pairwise relatively prime positive integers, and let a_1, a_2, \dots, a_k be integers. Then, the system of congruences, $x \equiv a_i \pmod{m_i}$ for $1 \leq i \leq k$, has a unique solution modulo $M = m_1 \times m_2 \times \dots \times m_k$, which is given by

$$x \equiv a_1 M_1 y_1 + a_2 M_2 y_2 + \dots + a_k M_k y_k \pmod{M}$$

where $M_i = \frac{M}{m_i}$ and $y_i \equiv \frac{1}{M_i} \pmod{m_i}$ for $1 \leq i \leq k$.

Let $x \in \mathbb{Z}_{n^2}^*$, from the Euler Theorem, we have

$$\begin{cases} x^{\phi(p^2)} = x^{p(p-1)} = x^{p \cdot 2p'} = 1 \pmod{p^2} \Rightarrow x^{pq \cdot 2p'q'} = 1^{qq'} \pmod{p^2} \Rightarrow x^{n\lambda} = 1 \pmod{p^2} \\ x^{\phi(q^2)} = x^{q(q-1)} = x^{q \cdot 2q'} = 1 \pmod{q^2} \Rightarrow x^{pq \cdot 2p'q'} = 1^{pp'} \pmod{q^2} \Rightarrow x^{n\lambda} = 1 \pmod{q^2} \end{cases}$$

Because $\gcd(p^2, q^2) = 1$, we can apply the Extended Euclidean Algorithm to find two integers s, t such that $s \cdot p^2 + t \cdot q^2 = 1$. Let $m_1 = p^2$, $m_2 = q^2$, then $M = m_1 m_2 = n^2$, $M_1 = q^2$, $M_2 = p^2$, $y_1 = t$, and $y_2 = s$. Based on the Chinese Remainder Theorem, where $a_1 = a_2 = 1$, we have

$$x^{n\lambda} = a_1 M_1 y_1 + a_2 M_2 y_2 \pmod{M} = 1 \cdot q^2 \cdot t + 1 \cdot p^2 \cdot s \pmod{n^2} = 1 \pmod{n^2}$$

n -th Residues Modulo n^2

Definition (**n -th Residues Modulo n^2**). A number $y \in \mathbb{Z}_{n^2}^*$ is said to be an n -th residues modulo n^2 if there exists a number $x \in \mathbb{Z}_{n^2}^*$ such that $y = x^n \pmod{n^2}$.

Let **NR** be the set of n -th residues modulo n^2 . It has been proved that the size of **NR** is exactly $\phi(n)$, i.e., $|\mathbf{NR}| = \phi(n)$. In addition, it has also been proved that “given $y \in \mathbb{Z}_{n^2}^*$, decide whether or not y is n -th residue modulo n^2 ” is a hard problem, i.e., there does not exist an algorithm that solves the problem in a polynomial time.

Let $\mathcal{G} = \{u \in \mathbb{Z}_{n^2}^* \mid \text{ord}(u) = kn, 1 \leq k \leq \lambda\}$. A special case $g = (1+n)^a \pmod{n^2}$ with $\text{ord}(g) = n$ for some random integer $a \geq 1$ belongs to \mathcal{G} , because $g^n = (1+n)^{an} = 1 \pmod{n^2}$. Define a function $f(m, r) = g^m r^n \pmod{n^2}$ over $\mathbb{Z}_n \times \mathbb{Z}_n^* \rightarrow \mathbb{Z}_{n^2}^*$, we can show that it is a bijective function. First, because $|\mathbb{Z}_n| = n$, $|\mathbb{Z}_n^*| = \phi(n)$, and $|\mathbb{Z}_{n^2}^*| = \phi(n^2) = n\phi(n)$, we have $|\mathbb{Z}_n \times \mathbb{Z}_n^*| = |\mathbb{Z}_{n^2}^*|$, and thus $\mathbb{Z}_n \times \mathbb{Z}_n^* \rightarrow \mathbb{Z}_{n^2}^*$ is injective. On the other hand, if for some $(m_0, r_0), (m_1, r_1) \in \mathbb{Z}_n \times \mathbb{Z}_n^*$ with $f(m_0, r_0) = f(m_1, r_1)$, we have

n -th Residues Modulo n^2

$$\begin{aligned} g^{m_0} r_0^n &= g^{m_1} r_1^n \pmod{n^2} \Rightarrow g^{m_0 - m_1} r_0^n = r_1^n \pmod{n^2} \\ \Rightarrow g^{(m_0 - m_1)\lambda} r_0^{n\lambda} &= r_1^{n\lambda} \pmod{n^2} \Rightarrow g^{(m_0 - m_1)\lambda} = 1 \pmod{n^2} \end{aligned}$$

which means $\text{ord}(g)|(m_0 - m_1)\lambda$. By choosing $g = (1+n)^a \pmod{n^2}$ with $\text{ord}(g) = n$, we have $n|(m_0 - m_1)\lambda$. Because $\text{gcd}(n, \lambda) = 1$, we have $m_0 - m_1 = 0 \pmod{n}$. As a result, we have $m_0 = m_1 \pmod{n} \Rightarrow m_0 = m_1$. Once we have $m_0 = m_1$, we can further obtain $r_0^n = r_1^n \pmod{n^2}$ from $g^{m_0 - m_1} r_0^n = r_1^n \pmod{n^2}$. It has been easily proved that $f(x) = x^n \pmod{n^2}$ over $\mathbb{Z}_n^* \rightarrow \mathbf{NR}$ is bijective. Therefore, given $r_0^n = r_1^n \pmod{n^2}$, we have $r_0 = r_1$. With this nice bijective function $f(m, r) = g^m r^n \pmod{n^2}$, the Paillier PKE was proposed



Security Analysis of Paillier

- **The decisional composite residuosity assumption (DCRA) :**
- given a composite n and an integer z , it is hard to decide whether z is a n -residue modulo n^2 or not, i.e., whether there exists y such that

$$z \equiv y^n \pmod{n^2}$$

- semantic security against chosen-plaintext attacks (IND-CPA)



Decide whether a number $z \in Z_{n^2}$ is a random or there exists $y \in Z_{n^2}$ such that $z \equiv y^n \pmod{n^2}$



attacker



$$z \equiv y^n \pmod{n^2}$$

Step 1: choose two message $m_0, m_1 \in Z_n$

m_0, m_1

Step 2: randomly choose a bit
 $b \in \{0,1\}$

$$c = g^{m_b} \cdot z \pmod{n^2}$$

c

Step 3: guess $b' \in \{0,1\}$, because we consider IND-CPA

b'

Step 4: check if $b' = b$

decide z



Step 1: randomly choose a bit $\alpha \in \{0,1\}$, if $\alpha = 0$, set $z \equiv y^n \pmod{n^2}$,
else randomly choose $z \in Z_{n^2}$, i.e., $z \neq y^n \pmod{n^2}$

Step 2: give z to Challenger, and the challenger will guess one bit 0 or 1 on α



Step 0: know an attacker have a non-negligible advantage ε to attack the Paillier encryption under IND-CPA, $\text{Adv}^{\text{IND-CPA}} = \varepsilon = 2 \cdot \Pr[b' = b] - 1$, $\Pr[b' = b] = \frac{1}{2} + \frac{\varepsilon}{2}$

Step 1: run the previous protocol

Step 2: check if $b' = b$, return his guess $\alpha' \in \{0,1\}$ on α , i.e., if $b' = b$, set $\alpha' = 0$

challenger

If $\alpha = 0$, $z \equiv y^n \pmod{n^2}$. Then, $c = g^{m_b} \cdot z \pmod{n^2}$ is a valid Paillier ciphertext, and the attacker can exert his attack advantage to correctly guess $b' = b$ with probability $\frac{1}{2} + \frac{\varepsilon}{2}$, i.e., $\Pr[\alpha' = 0 | \alpha = 0] = \Pr[b' = b | \alpha = 0] = \frac{1}{2} + \frac{\varepsilon}{2}$

If $\alpha = 1$, a random $z \in Z_{n^2}$. Then, $c = g^{m_b} \cdot z \pmod{n^2}$ is a not valid Paillier ciphertext, and the attacker cannot exert his attack advantage to correctly guess $b' = b$, i.e., the probability is just $\frac{1}{2}$, i.e., $\Pr[\alpha' = 0 | \alpha = 1] = \Pr[b' = b | \alpha = 1] = \frac{1}{2}$

$$\text{Adv}^{\text{DCRA}} = |\Pr[\alpha' = 0 | \alpha = 0] - \Pr[\alpha' = 0 | \alpha = 1]| = |\Pr[b' = b | \alpha = 0] - \Pr[b' = b | \alpha = 1]| = \frac{\varepsilon}{2}$$

Adv^{DCRA} is also non-negligible. Because we know decisional composite residuosity is hard, which causes contradiction.

Security Reduction

If Paillier is not secure, we can use it to solve DCRA. However, DCRA is hard, we conclude Paillier is secure!

Security Analysis of Boneh-Goh-Nissim (BGN)

- Subgroup decision problem
 - Given $h \in G$ decide whether h has order q or order n
 - Subgroup decision assumption:

\forall poly-time Adv:

$$\Pr[(p, q, G, G_T, e, g) \leftarrow \text{Gen}(1^k); h \leftarrow G^* : \text{Adv}(n, G, G_T, e, g, h) = 1] \\ \approx \Pr[(p, q, G, G_T, e, g) \leftarrow \text{Gen}(1^k); h \leftarrow G_q^* : \text{Adv}(n, G, G_T, e, g, h) = 1]$$



$(p, q, G, G_T, e, g) \leftarrow \text{Gen}(1^k)$; randomly choose a bit $\alpha \in \{0, 1\}$, if $\alpha = 0$, $h \leftarrow G^*$, else if $\alpha = 1$, $h \leftarrow G_q^*$
Publish n, G, G_T, e, g, h , to determine $h \leftarrow G^*$ or $h \leftarrow G_q^*$



attacker



n, G, G_T, e, g, h

Step 1: choose two message $m_0, m_1 \in \text{small message space}$

m_0, m_1

c

Step 2: randomly choose a bit $b \in \{0, 1\}$ and a random r

$$c = g^{m_b} \cdot h^r$$

Step 3: guess $b' \in \{0, 1\}$, because we consider IND-CPA

b'

Step 4: check if $b' = b$

decide h



$(p, q, G, G_T, e, g) \leftarrow \text{Gen}(1^k)$; randomly choose a bit $\alpha \in \{0, 1\}$, if $\alpha = 0$, $h \leftarrow G_q^*$, else if $\alpha = 1$, $h \leftarrow G^*$
Publish n, G, G_T, e, g, h , to determine $h \leftarrow G^*$ or $h \leftarrow G_q^*$



challenger

Step 0: know an attacker have a non-negligible advantage ε to attack the BGN under IND-CPA , $\text{Adv}^{\text{IND-CPA}} = \varepsilon = 2 \cdot \Pr[b' = b] - 1$, $\Pr[b' = b] = \frac{1}{2} + \frac{\varepsilon}{2}$

Step 1: run the previous protocol

Step 2: check if $b' = b$, return his guess $\alpha' \in \{0, 1\}$ on α , i.e., if $b' = b$, set $\alpha' = 0$



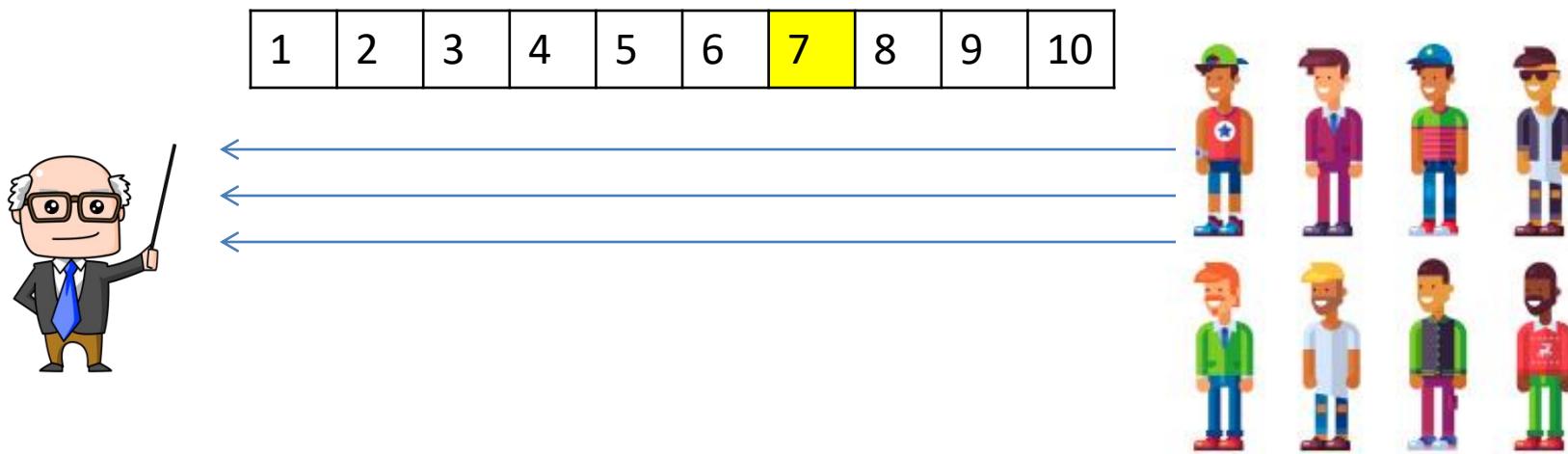
If $\alpha = 0, h \leftarrow G_q^*$. Then, $c = g^{m_b} \cdot h^r$ is a valid BGN ciphertext, and the attacker can exert his attack advantage to correctly guess $b' = b$ with probability $\frac{1}{2} + \frac{\varepsilon}{2}$,

$$\text{i.e., } \Pr[\alpha' = 0 | \alpha = 0] = \Pr[b' = b | \alpha = 0] = \frac{1}{2} + \frac{\varepsilon}{2}$$

If $\alpha = 1, h \leftarrow G^*$. Then, $c = g^{m_b} \cdot h^r$ is a not valid Paillier ciphertext, and the attacker cannot exert his attack advantage to correctly guess $b' = b$, i.e., the probability is just $\frac{1}{2}$, i.e., $\Pr[\alpha' = 0 | \alpha = 1] = \Pr[b' = b | \alpha = 1] = \frac{1}{2}$

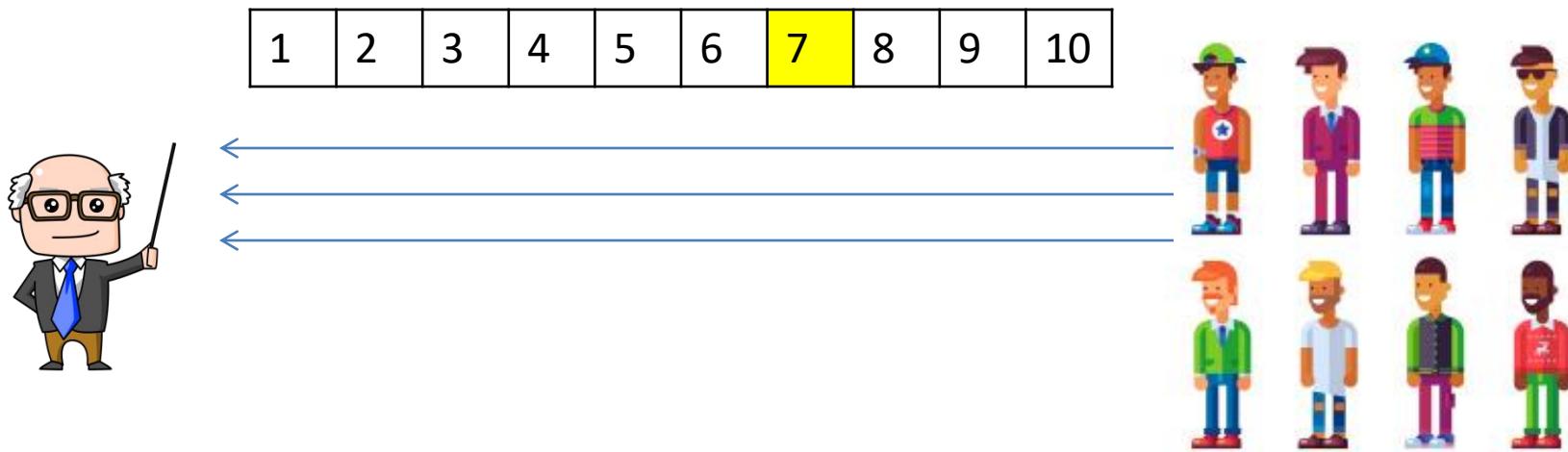
$$\text{Adv}^{\text{DCRA}} = |\Pr[\alpha' = 0 | \alpha = 0] - \Pr[\alpha' = 0 | \alpha = 1]| = |\Pr[b' = b | \alpha = 0] - \Pr[b' = b | \alpha = 1]| = \frac{\varepsilon}{2}$$

Application Discussion: How to design a Privacy-Preserving Lecturer Evaluation



- **Requirements:** Lecturer can calculate the average evaluation $E(x)$, but cannot know each individual student's evaluation.

Application Discussion: How to design a Privacy-Preserving Lecturer Evaluation



- **Requirements:** Lecturer can calculate the average evaluation $E(x)$ and Variance $\text{Var}(x) = E(x^2) - (E(x))^2$, but cannot know each individual student's evaluation.

Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 5: Anonymous communication network
techniques

Lecturer: Rongxing LU

Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

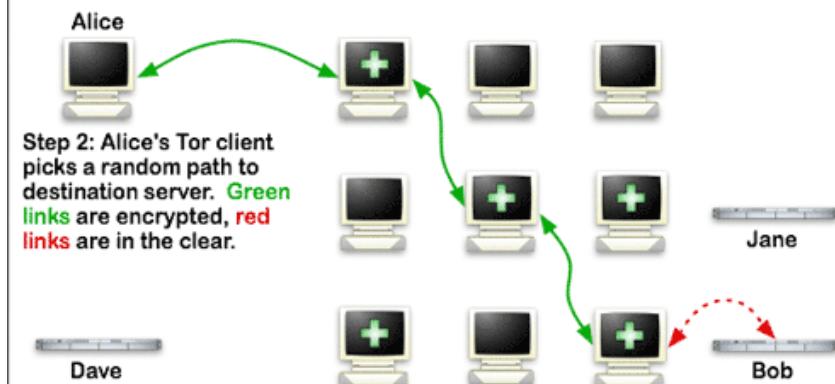
How to achieve anonymous communications over Internet?

- One Solution: Tor

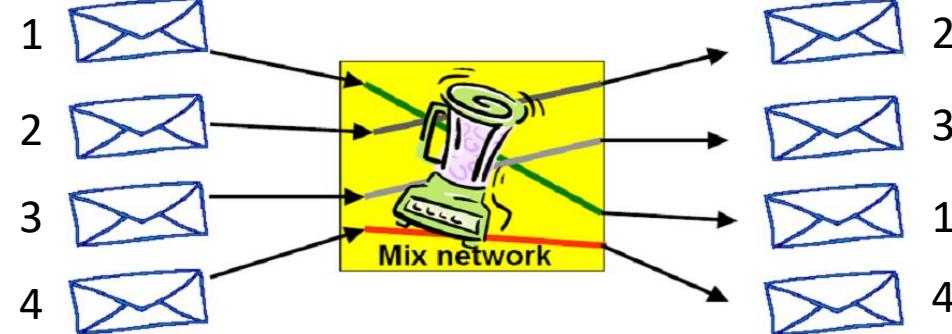
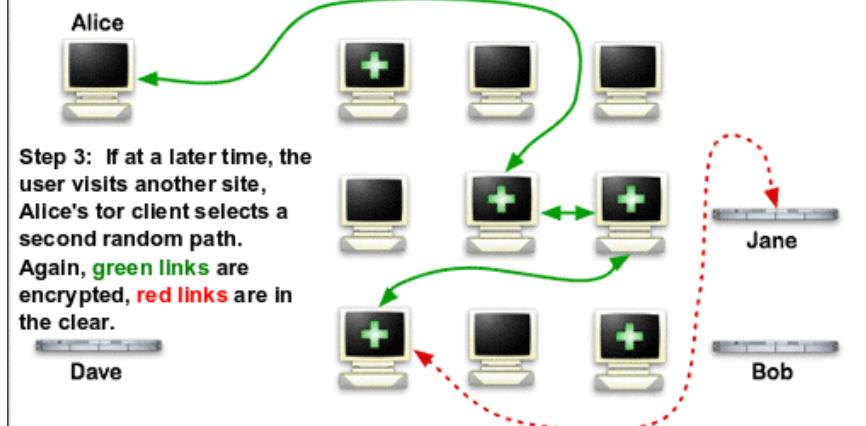
E How Tor Works: 1



E How Tor Works: 2



E How Tor Works: 3



Why Anonymous Communication & Network?

- Privacy issue
- Some covert missions may require anonymous communication
- In hostile environments, end-hosts may need to hide their communications to prevent being captured

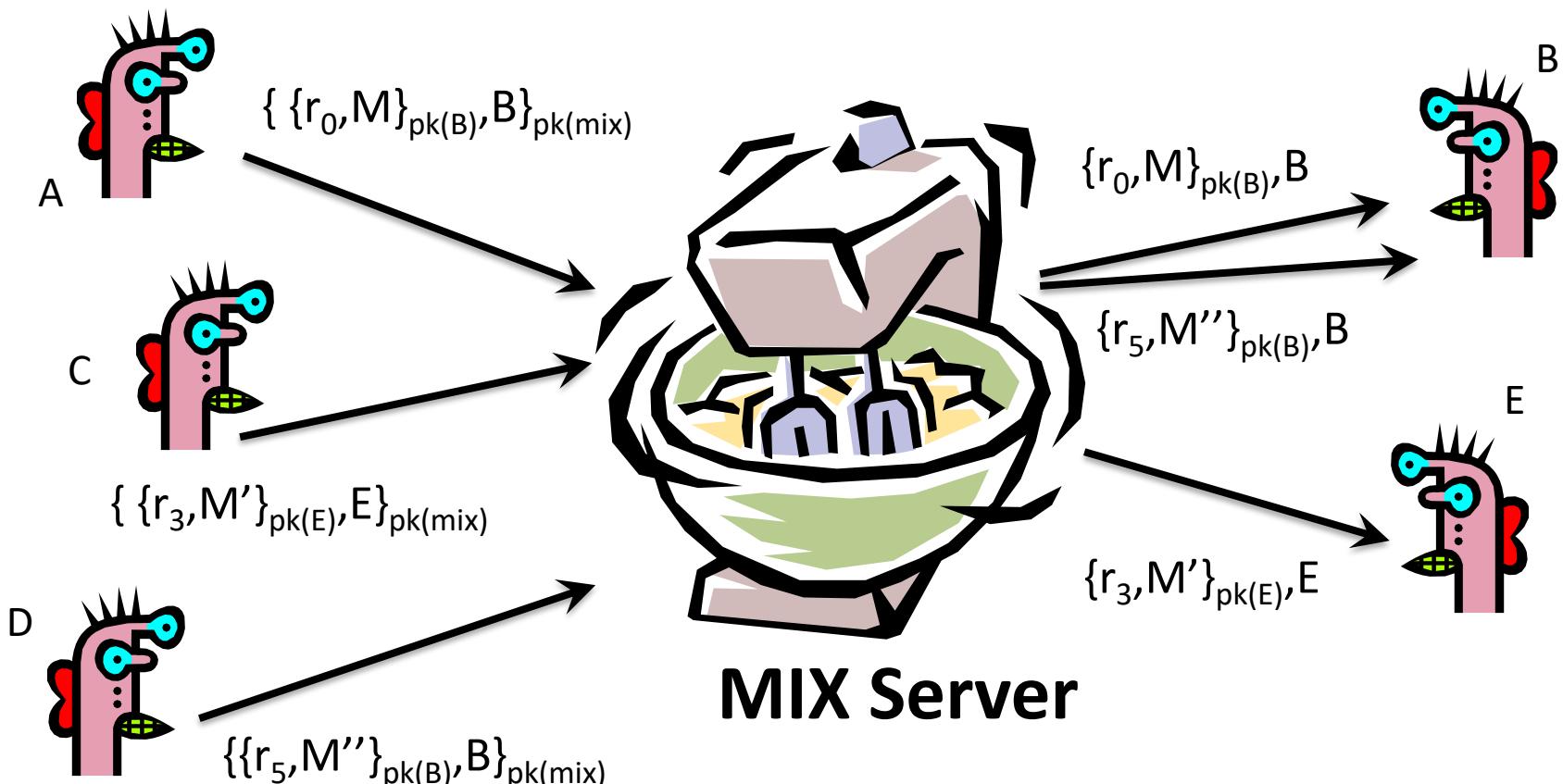
Anonymity in terms of unlinkability

- Sender anonymity
 - A particular message is not linkable to any sender and that to a particular sender, no message is linkable
- Recipient anonymity
 - A particular message cannot be linked to any recipient and that to a particular recipient, no message is linkable
- Relationship anonymity
 - The sender and the recipient cannot be identified as communicating with each other, even though each of them can be identified as participating in some communication.
- A. Pfizmann and M. Waidner, *Networks without User Observability*. Computers & Security 6/2 (1987) 158-166

Re-encryption techniques for Mix-Network

- Mix Network
 - ElGamal Re-encryption
- Universal Mix Network
 - Universal ElGamal Re-encryption

Mix Network



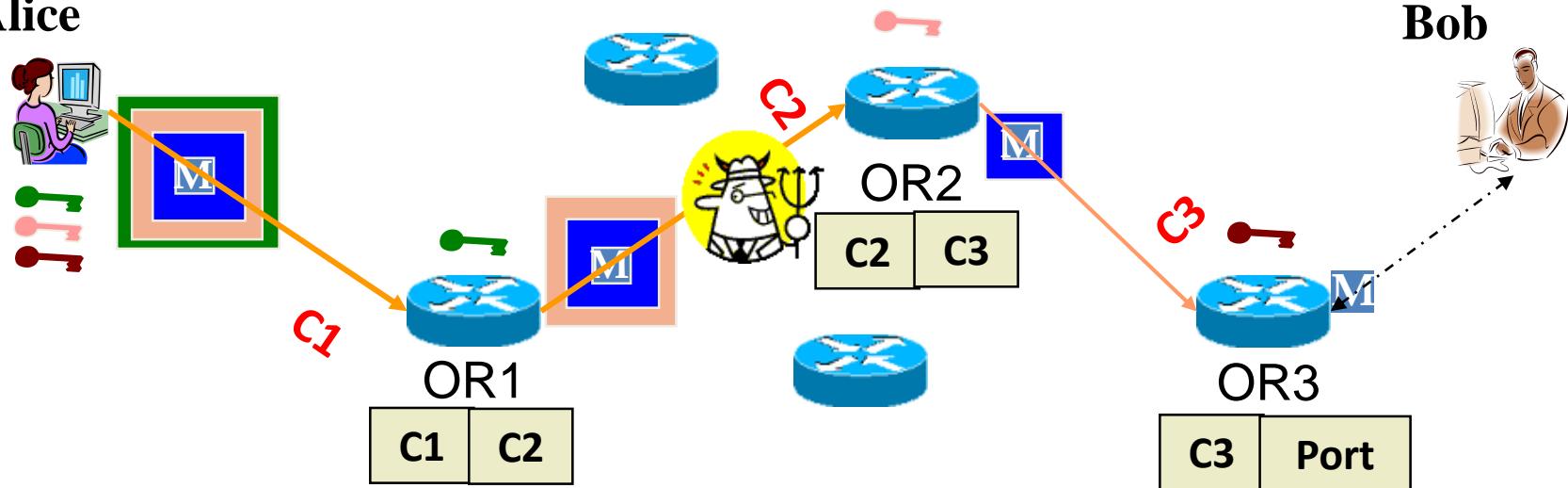
Adversary knows all senders and all receivers, but cannot link a sent message with a received message

Each packet: before being sent to the mix server, all content should be encrypted with the mix server's public key.

Is it possible to use mix server's public key to only encrypt destination info ? Yes.

Onion Routing

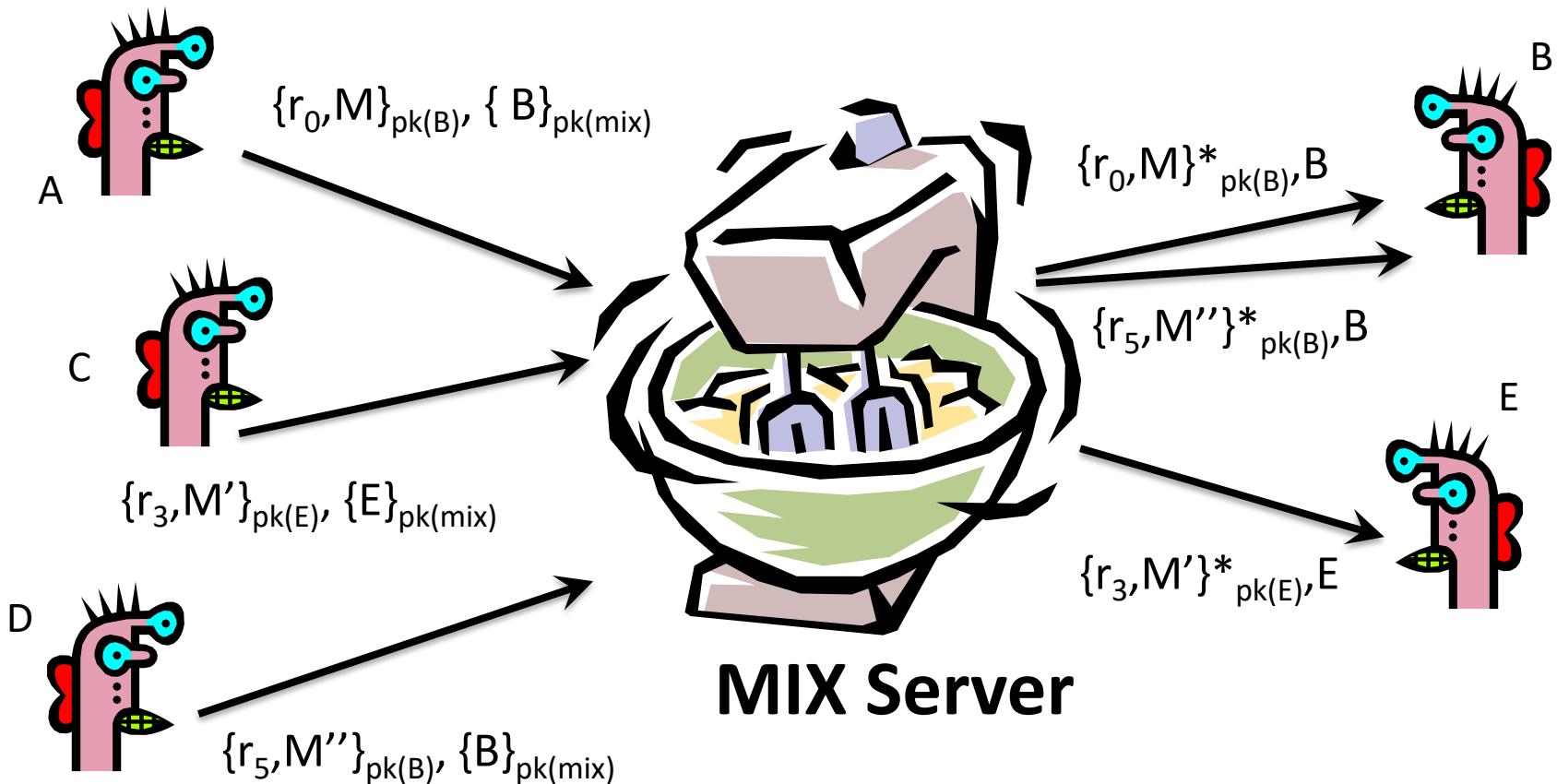
Alice



Bob

- A circuit is built incrementally one hop by one hop
- Onion-like encryption
 - Alice negotiates an AES key with each router
 - Messages are divided into equal sized **cells**
 - Each router knows only its predecessor and successor
 - Only the Exit router (OR3) can see the message, however it does not know where the message is from

Mix Network (2)



In this case, senders only need to use mix server's public key to encrypt receivers' information.

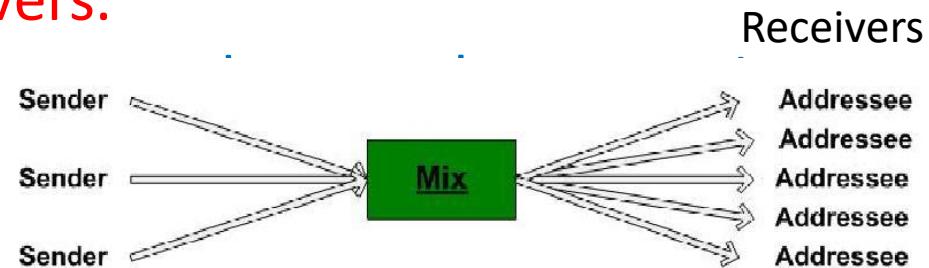
With the receiver's public key, the mix server needs to re-pack each incoming packet, i.e., $\{r_0, M\}_{pk(B)} \rightarrow \{r_0, M\}^*_{pk(B)}$, so as to make them unlinkable, but the transformed ciphertext can still be recovered.

How to achieve this goal? We can apply the **re-encryption technique**.

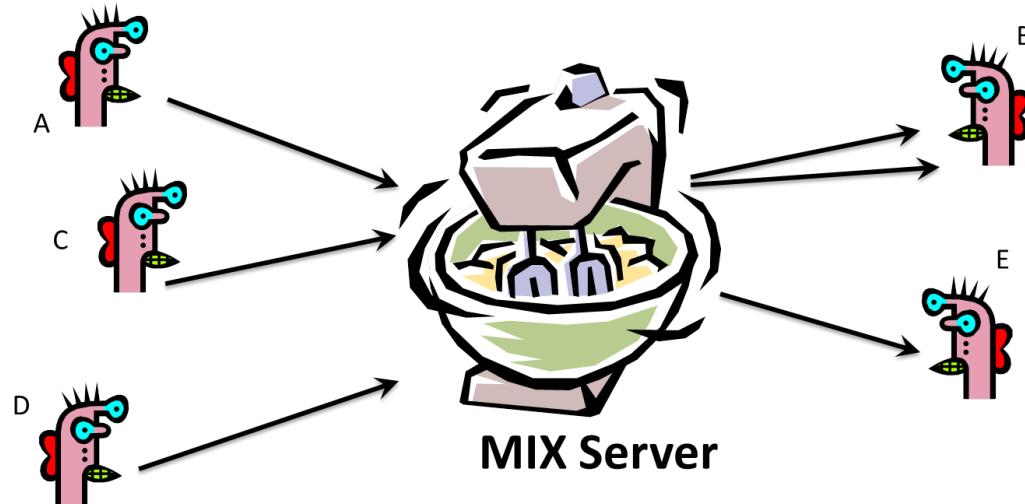
ElGamal Re-encryption

If $A = g^a \text{ mod } p$ is a public key
and the pair $(X, Y) = (A^r M \text{ mod } p, g^r \text{ mod } p)$
is an encryption of message M , then for any value c , the pair
 $(X', Y') = (A^c X, g^c Y) = (A^{c+r} M \text{ mod } p, g^{c+r} \text{ mod } p)$
is a transformed encryption of the same message M , for any
value c .

- Note that, the mix server must know the public key $A = g^a \text{ mod } p$ of the receiver. Otherwise, it does not work. So this is the reason why the senders need encrypt the receivers with the mix server's public key so that only the mix server knows the receivers.
- Now, the question is “if mix how to process?”



Universal Mix Network



1. **Submission of inputs.** Senders post to a bulletin board messages that are universally encrypted under the public key of the recipient for whom they are intended.
2. **Universal mixing.** Any server can be called upon to mix the contents of the bulletin board, **but the server does not need to know the public key of the recipient.**
3. **Retrieval of the outputs.** Potential recipients **must try to decrypt every encrypted message output by the universal mixnet.** Successful decryptions correspond to messages that were intended for that recipient. The others (corresponding to decryption output ' \perp ') are discarded by the party attempting to perform the decryption.

Universal ElGamal Re-encryption

If $A = g^a \bmod p$ is a public key

and the pair $(X, Y, W, Z) = (A^r M \bmod p, g^r \bmod p, A^s \bmod p, g^s \bmod p)$ is an encryption of message M , then for any value r_1, r_2 , the pair

$$\begin{aligned} & (X', Y', W', Z') \\ &= (X \cdot W^{r_1} \bmod p, Y \cdot Z^{r_1} \bmod p, W^{r_2} \bmod p, Z^{r_2} \bmod p) \\ &= (A^{(r+s \cdot r_1)} M \bmod p, g^{(r+s \cdot r_1)} \bmod p, A^{s \cdot r_2} \bmod p, g^{s \cdot r_2} \bmod p) \\ &= (A^{r'} M \bmod p, g^{r'} \bmod p, A^{s'} \bmod p, g^{s'} \bmod p) \end{aligned}$$

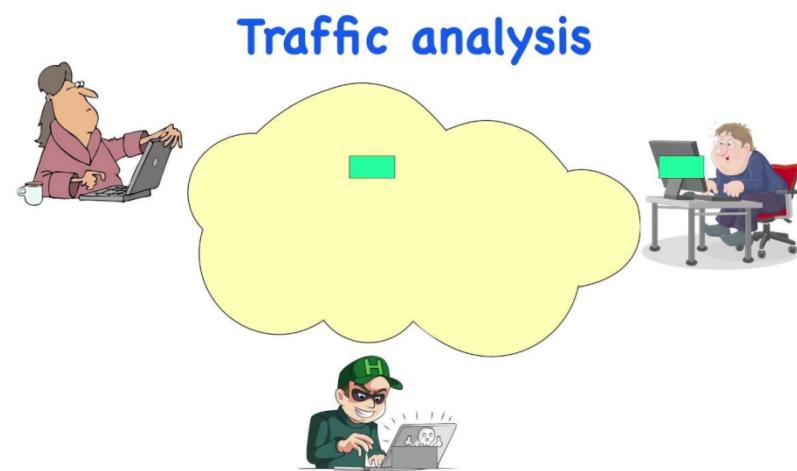
For $r' = r + s \cdot r_1$, $s' = s \cdot r_2$

Decryption:

- $M = \frac{X}{Y^a}$, check whether $\frac{W}{Z^a} = 1 \bmod p$. If $\frac{W}{Z^a} = 1 \bmod p$, M is accepted. Otherwise, rejected.

Traffic Analysis Attacks against an Anonymous Communication System

- Contextual attacks
 - Communication pattern attacks
 - Packet counting attacks
 - Intersection attack



Communication pattern attacks

- By simply looking at the communication patterns (when users send and receive), one can find out a lot of useful information.
- Communicating participants normally do not "talk" at the same time, that is, when one party is sending, the other is usually silent.
- The longer an attacker can observe this type of communication synchronization, the less likely it is just an uncorrelated random pattern.
- This attack can be mounted by a passive adversary that can monitor entry and exit mix nodes. Law Enforcement officials might be quite successful mounting this kind of attack as they often have a-priori information: they usually have a hunch that two parties are communicating and just want to confirm their suspicion.



Packet counting attacks

- 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.
- A partial solution is to have parties only send standard numbers of messages but this isn't a viable option in many settings.
- The packet counting and communication pattern attacks can be combined to get a "message frequency" attack



Intersection attack

- An attacker having information about what users are active at any given time can, through repeated observation, determine what users communicate with each other. This attack is based on the observation that users typically communicate with a relatively small number of parties.
- For example, the typical user usually queries the same websites in different sessions (his queries are not random). By performing an operation similar to an intersection on the sets of active users at different time it is probable that, the attacker can gain interesting information.
- The intersection attack is well known open problem and seems extremely difficult to solve in an efficient manner.

K-Anonymous Message Transmission

- Oct, 2003, Luis von Ahn, Andrew Bortz and Nicholas J. Hopper
- present a scheme for anonymous communication that is efficient and requires no trusted third parties

The Model

- Reliable Communication
- The adversary can see all communications in network
- The adversary can own some of the participants
- A participant owned by the adversary may act arbitrarily

DC Nets

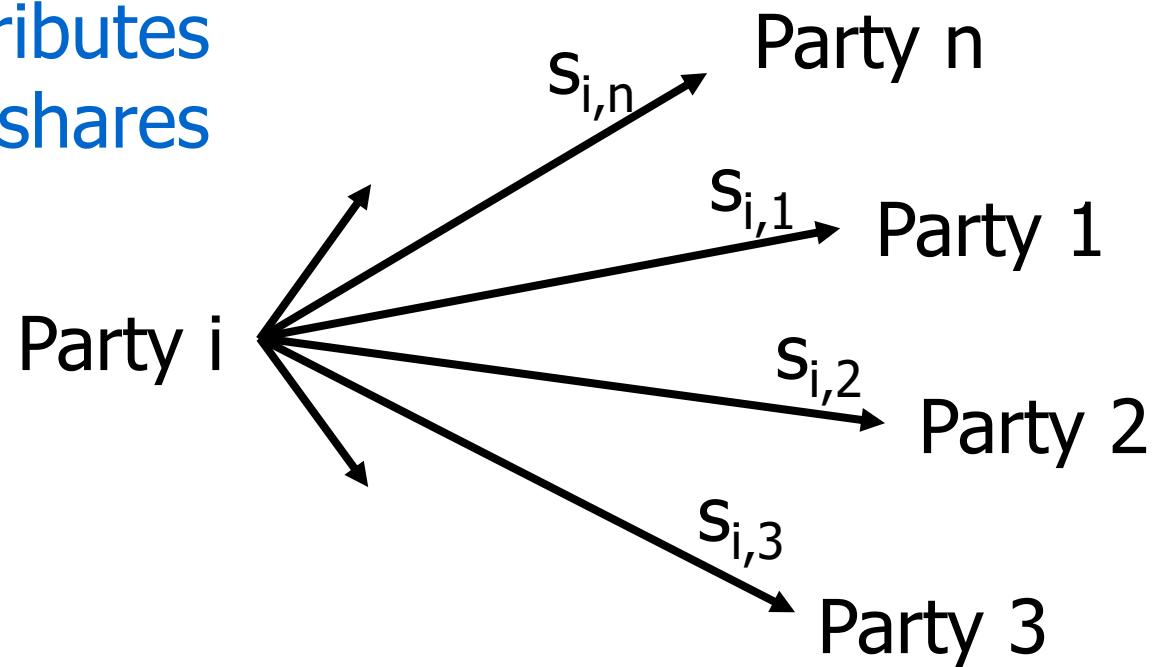
(Dining cryptographers networks)

- Key Idea:
 - Divide time into small steps
 - At step t , party i wants to send message $M_i \in Z_m$
 - If party j doesn't want to send a message at step t , they must send $M_j=0$
 - Each party i splits M_i into n random shares

$$M_i = s_{i,1} + s_{i,2} + \dots + s_{i,n-1} + \underbrace{(M_i - (s_{i,1} + \dots + s_{i,n-1}))}_{S_{i,n}}$$

DC Nets

Each party distributes their n shares



DC Nets

- All parties add up every share that they have received and broadcast the result
 - (Let B_i denote Party i's broadcast)
- $B_i = s_{1,i} + s_{2,i} + \dots + s_{n,i}$

$$M_i = s_{i,1} + s_{i,2} + \dots + s_{i,n-1} + s_{i,n}$$

$$B_i = s_{1,i} + s_{2,i} + \dots + s_{n,i}$$

$$B_1 + B_2 + \dots + B_n = M_1 + M_2 + \dots + M_n$$

DC Nets

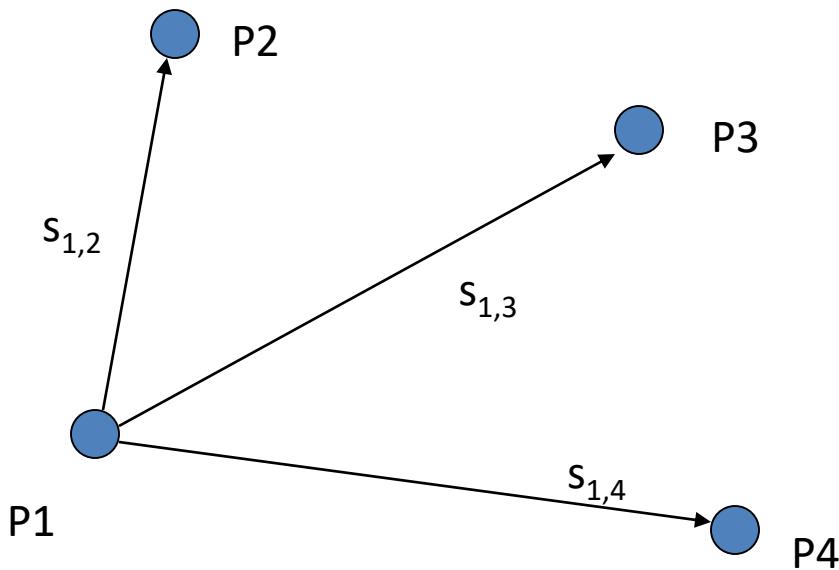
- If only one of the M_i is nonzero, then:
 - $B_1 + B_2 + \dots + B_n = M_i$
- DC Nets Challenge
 - It is very easy for the adversary to jam the channel!
 - Communication complexity is $O(n^2)$

Full Anonymity Versus k-Anonymity

- We will relax the requirement that the adversary learns **nothing** about the origin of a given message
- We will accept **k-anonymity**, in which the adversary can only narrow down his search to **k** participants

k-anonymous message transmission (k-AMT)

Idea: Divide N parties into “small” DC-Nets of size $O(k)$. Encode M_t as (group, msg) pair



$$s_{1,1} + s_{1,2} + s_{1,3} + s_{1,4} = (G_t, M_t)$$

How to compromise k-anonymity

- If everyone follows the protocol, it's *impossible* to compromise the anonymity guarantee.
- So instead, don't follow the protocol: if Alice can never send anonymously, she will have to communicate using onymous means.

How to break k-AMT (I)

- Don't follow the protocol: after receiving shares $s_{1,i}, \dots, s_{k,i}$, instead of broadcasting s_i , generate a random value r and broadcast that instead.
- This will randomize the result of the DC-Net protocol, preventing Alice from transmitting.

Stopping the “randomizing” attack

- Solution: Use *Verifiable Secret Sharing*. Every player in the group announces (by broadcast) a commitment to all of the shares of her input.
- These commitments allow verification of her subsequent actions.

Pedersen Commitment Scheme

- **Setup:** receiver chooses...
 - Large primes p and q such that $q|p-1$
 - Generator g of the order- q subgroup of \mathbb{Z}_p^*
 - Random secret a from \mathbb{Z}_q
 - $h = g^a \text{ mod } p$
 - Values p, q, g, h are public, a is secret
- **Commit:** to commit to some $x \in \mathbb{Z}_q$, sender chooses random $r \in \mathbb{Z}_q$ and sends $c = g^x h^r \text{ mod } p$ to receiver
 - This is simply $g^x (g^a)^r = g^{x+ar} \text{ mod } p$
- **Reveal:** to open the commitment, sender reveals x and r , receiver verifies that $c = g^x h^r \text{ mod } p$

Security of Pedersen Commitments

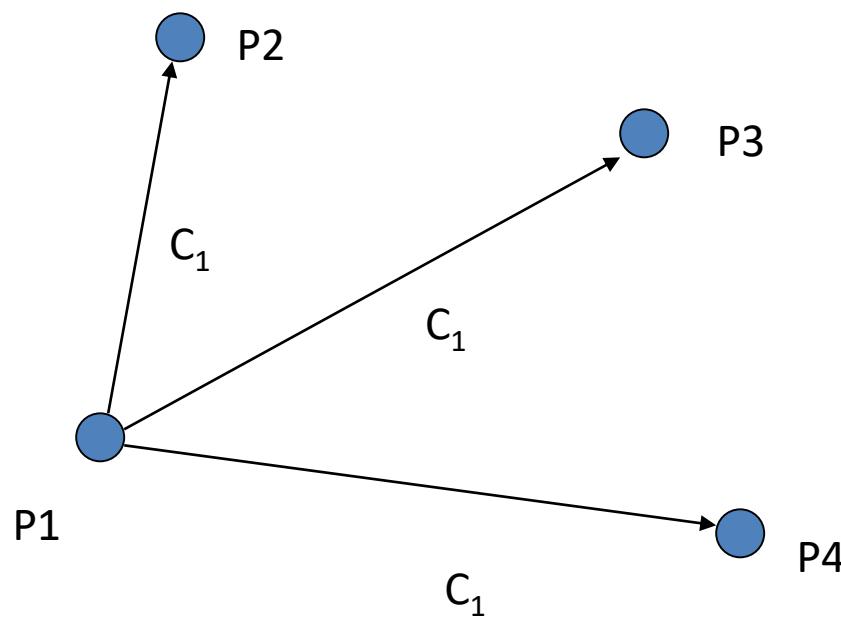
- Perfectly hiding
 - Given commitment c , every value x is equally likely to be the value committed in c
 - Given x, r and any x' , exists r' such that $g^x h^r = g^{x'} h^{r'}$
 $r' = (x-x')a^{-1} + r \pmod{q}$ (but must know a to compute r')
- Computationally binding
 - If sender can find different x and x' both of which open commitment $c=g^x h^r$, then he can solve discrete log
 - Suppose sender knows x, r, x', r' s.t. $g^x h^r = g^{x'} h^{r'} \pmod{p}$
 - Because $h=g^a \pmod{p}$, this means $x+ar = x'+ar' \pmod{q}$
 - Sender can compute a as $(x'-x)(r-r')^{-1}$
 - But this means sender computed discrete logarithm of h !

k-anonymous message transmission (k-AMT) with VSS

Before starting, each player *commits* to $s_{i,1} \dots s_{i,k}$ via *Pedersen commitment* $C(s,r)=g^s h^r$

$$s_{1,1} + s_{1,2} + s_{1,3} + s_{1,4} = x_1 = (G_i, M_i)$$

C	1
1	$g^{s_{1,1}} h^{r_{1,1}}$
2	$g^{s_{1,2}} h^{r_{1,2}}$
3	$g^{s_{1,3}} h^{r_{1,3}}$
4	$g^{s_{1,4}} h^{r_{1,4}}$

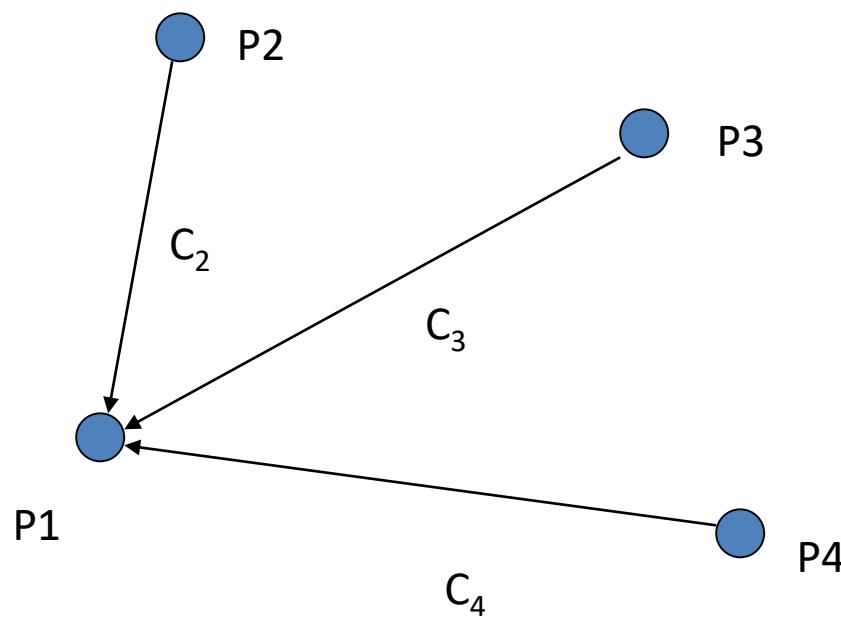


k-anonymous message transmission (k-AMT) with VSS

Before starting, each player *commits* to $s_{i,1} \dots s_{i,k}$ via *Pedersen commitment* $C(s,r)=g^s h^r$

$$s_{1,1} + s_{1,2} + s_{1,3} + s_{1,4} = x_1 = (G_i, M_i)$$

C	1...	k	
1	$g^{s_{1,1}} h^{r_{1,1}}$	$g^{s_{k,1}} h^{r_{k,1}}$	$g^{s_1} h^r$
2	$g^{s_{1,2}} h^{r_{1,2}}$	$g^{s_{k,2}} h^{r_{k,2}}$	$g^{s_2} h^r$
3	$g^{s_{1,3}} h^{r_{1,3}}$	$g^{s_{k,3}} h^{r_{k,3}}$	$g^{s_3} h^r$
4	$g^{s_{1,4}} h^{r_{1,4}}$	$g^{s_{k,4}} h^{r_{k,4}}$	$g^{s_4} h^r$
	$g^{x_1} h^r$	$g^{x_k} h^r$	



How to break k-AMT (II)

- The multiparty sum protocol gives k participants a single shared channel: at most one person can successfully transmit each turn.
- So: Transmit every turn! VSS still perfectly hides the value of each input; no one will know who is hogging the line.

Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 6: Private information retrieval techniques

Lecturer: Rongxing LU

Email: RLU1@unb.ca Office: GE 114

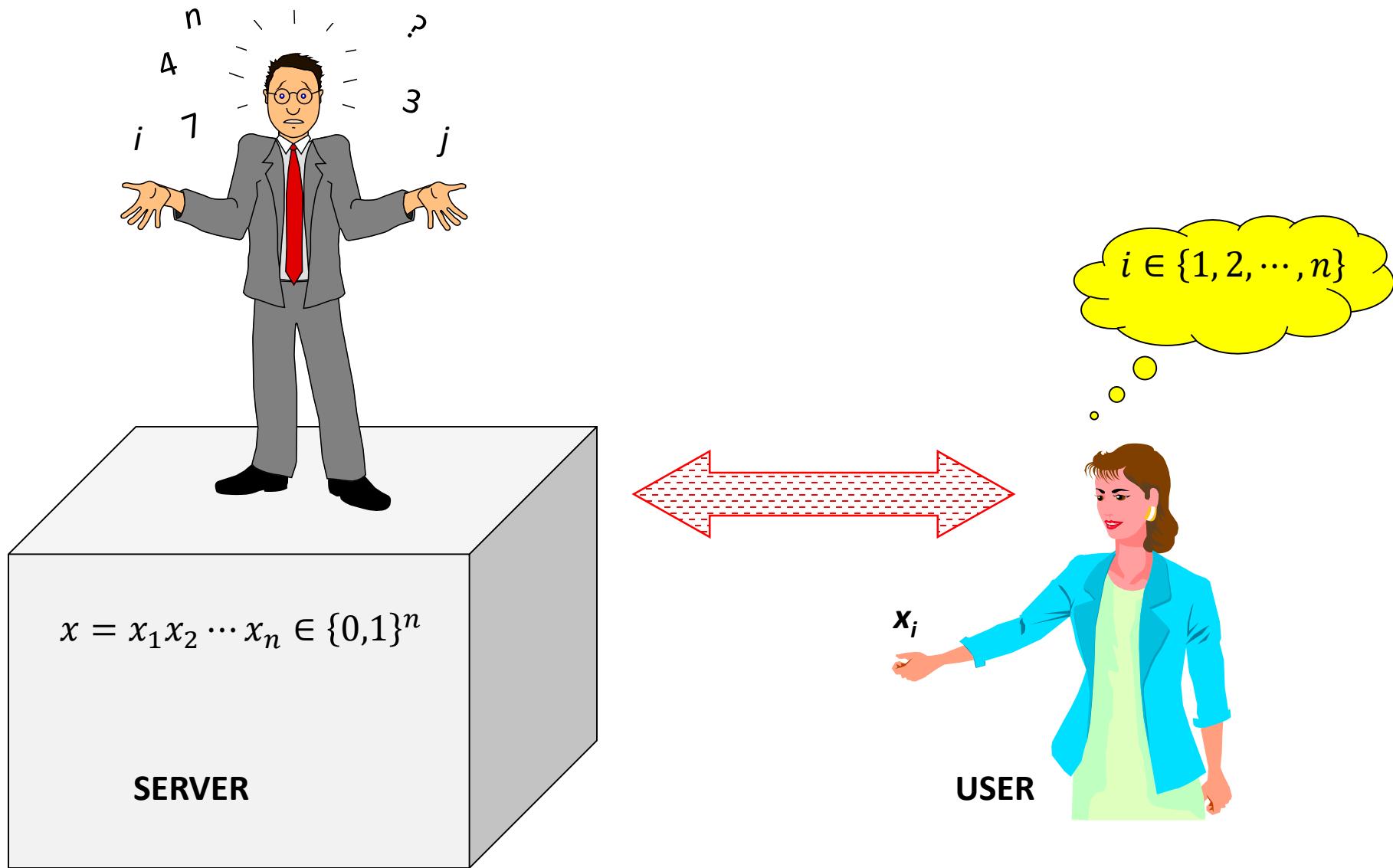
Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

Private information retrieval (PIR)

- What is PIR?
 - 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.

Private information retrieval (PIR)

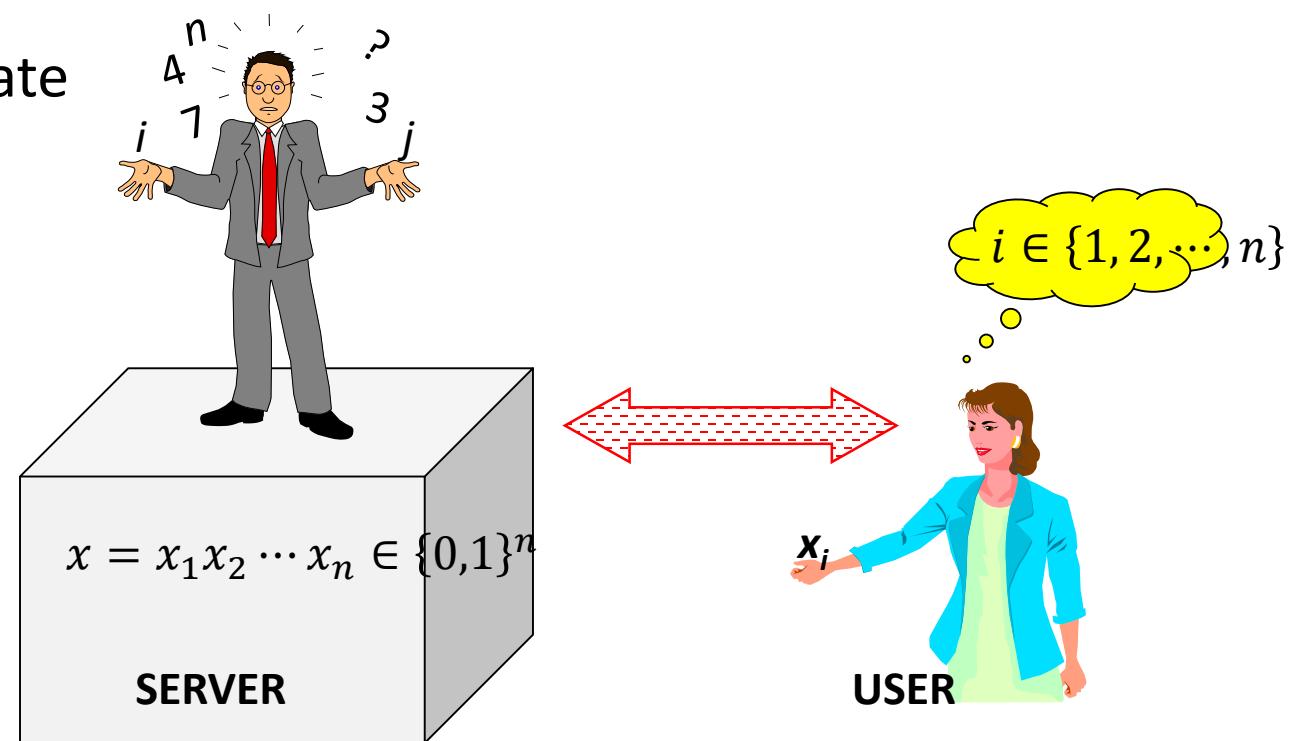


Why do we need to study PIR?

- Motivation
 - patent databases; stock quotes; web access; many more....
 - e.g., imagine buying in a store without the seller knowing what you buy.
- Goal
 - allow user to query database while hiding the identity of the data-items she is after.
 - Note: Encrypting requests is useful against third parties; not against owner of data.

PIR Modeling

- **Server:** holds n -bit string x
 - n should be thought of as **very large**
- **User:** wishes
 - to retrieve x_i
 - to keep i private

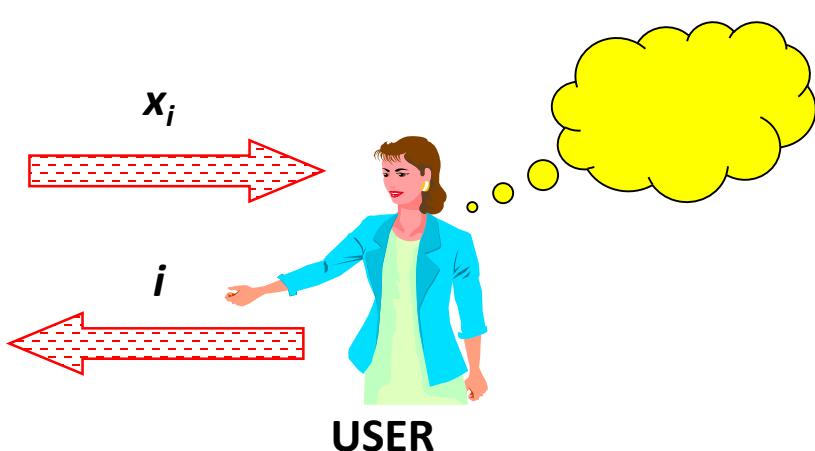
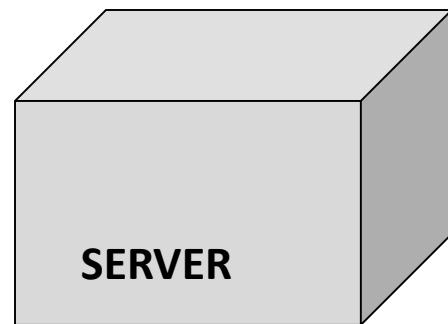


PIR Evolving

- 1) Non-Private Protocol

$$i \in \{1, 2, \dots, n\}$$

$$x = x_1 x_2 \cdots x_n \in \{0,1\}^n$$

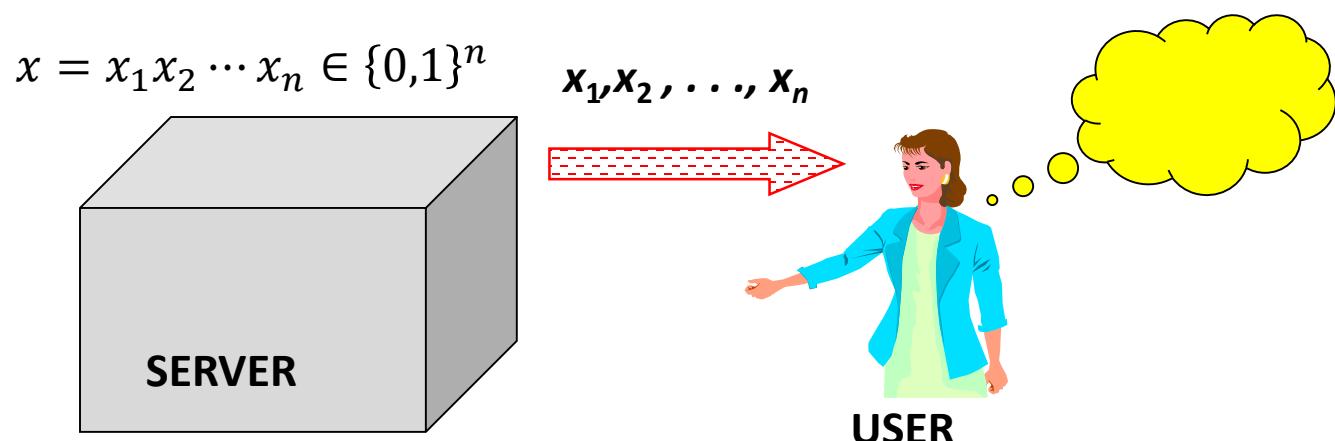


- NO privacy!!!

PIR Evolving (2)

- 2) Trivial Private Protocol

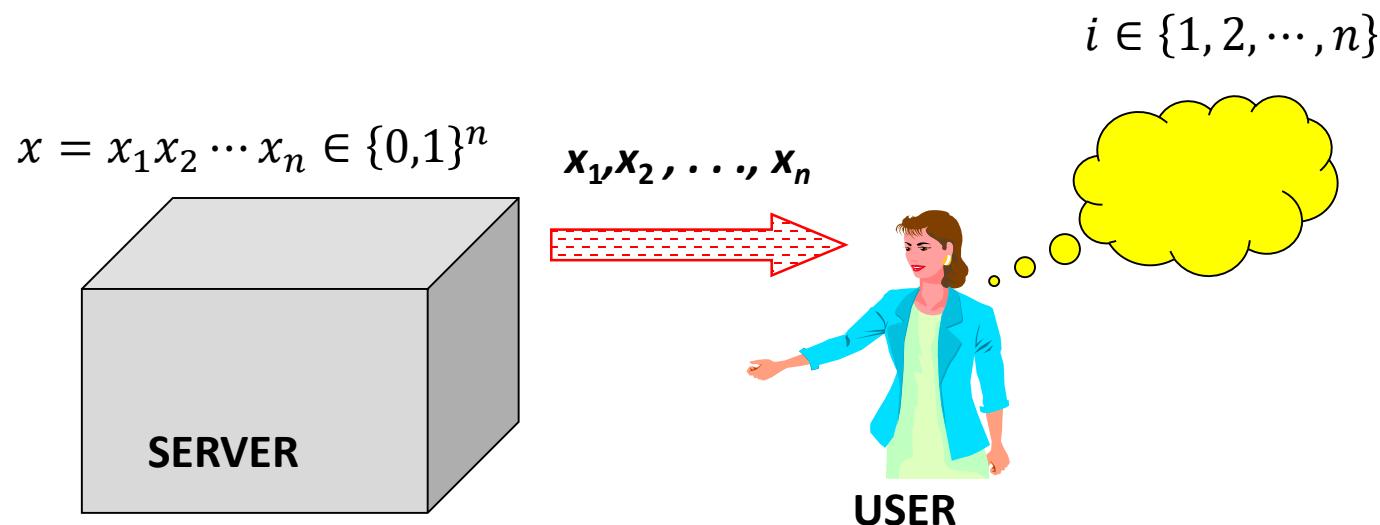
$$i \in \{1, 2, \dots, n\}$$



- Server sends entire database x to User.
- Information theoretic privacy.
- Communication: n
- Is this optimal?

Obstacle

- In any **1-server PIR** with information theoretic privacy the communication is at least n .



PIR Evolving (3)

- 3) Information-Theoretic PIR
 - Replicate database among k servers.
 - Unconditional privacy against t servers.
 - Default: $t=1$
- 4) Computational PIR
 - Computational privacy, based on cryptographic assumptions.

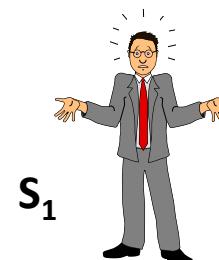
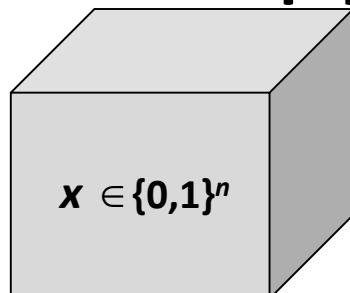
How to overcome the “obstacle”?

- User asks for additional random indices.
 - **Drawback:** reveals a lot of information
- Employ general crypto protocols to compute x_i privately.
 - **Drawback:** highly inefficient (polynomial in n).
- Anonymity (e.g., via Anonymizers).
 - **Note:** different concern: hides identity of user; not the fact that x_i is retrieved.

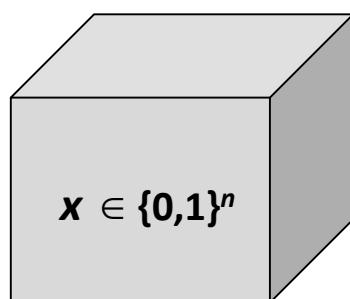
Known Comm. Upper Bounds

- Multiple servers, information-theoretic PIR:
 - 2 servers, comm. $n^{1/3}$
 - k servers, comm. $n^{1/\Omega(k)}$
 - $\log n$ servers, comm. $\text{Poly}(\log(n))$
- Single server, computational PIR:
 - Comm. $\text{Poly}(\log(n))$
 - Under appropriate computational assumptions

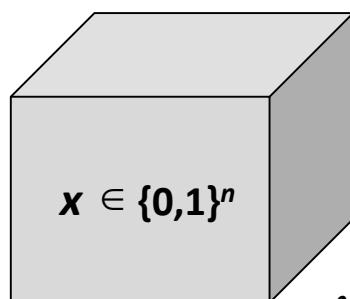
Approach I: k-Server PIR



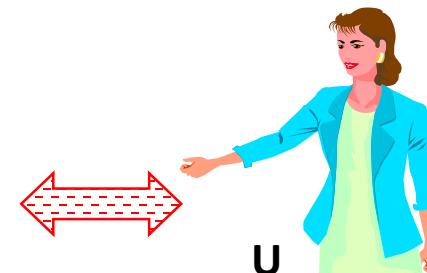
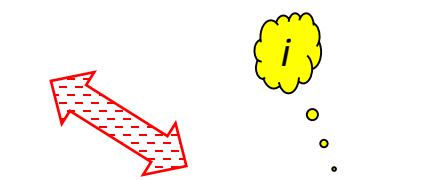
$$x = x_1 x_2 \cdots x_n \in \{0,1\}^n$$



$$x = x_1 x_2 \cdots x_n \in \{0,1\}^n$$



$$x = x_1 x_2 \cdots x_n \in \{0,1\}^n$$



Correctness: User obtains x_i

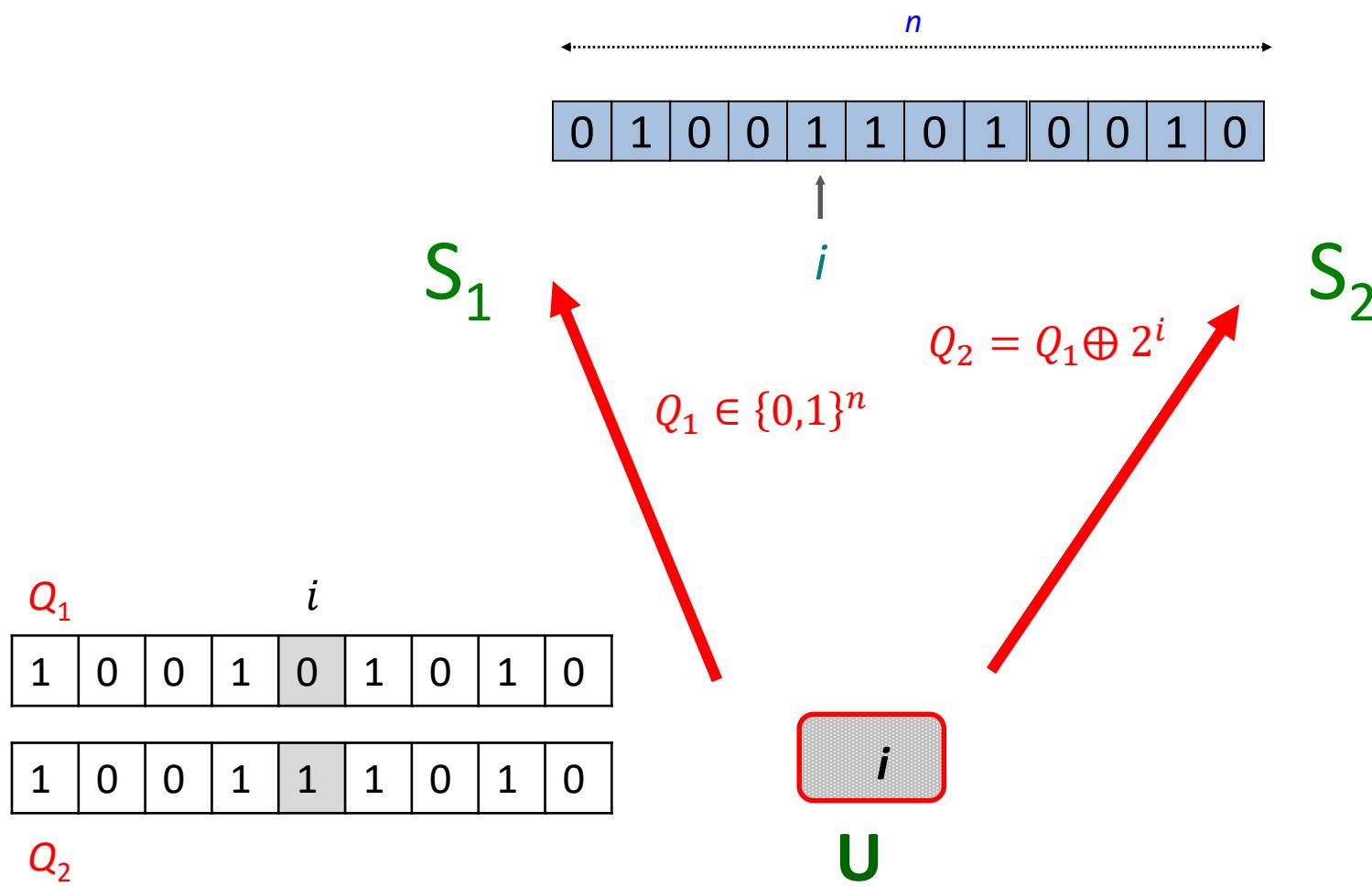
Privacy: No *single* server gets information about i

Information-Theoretic 2-Server PIR

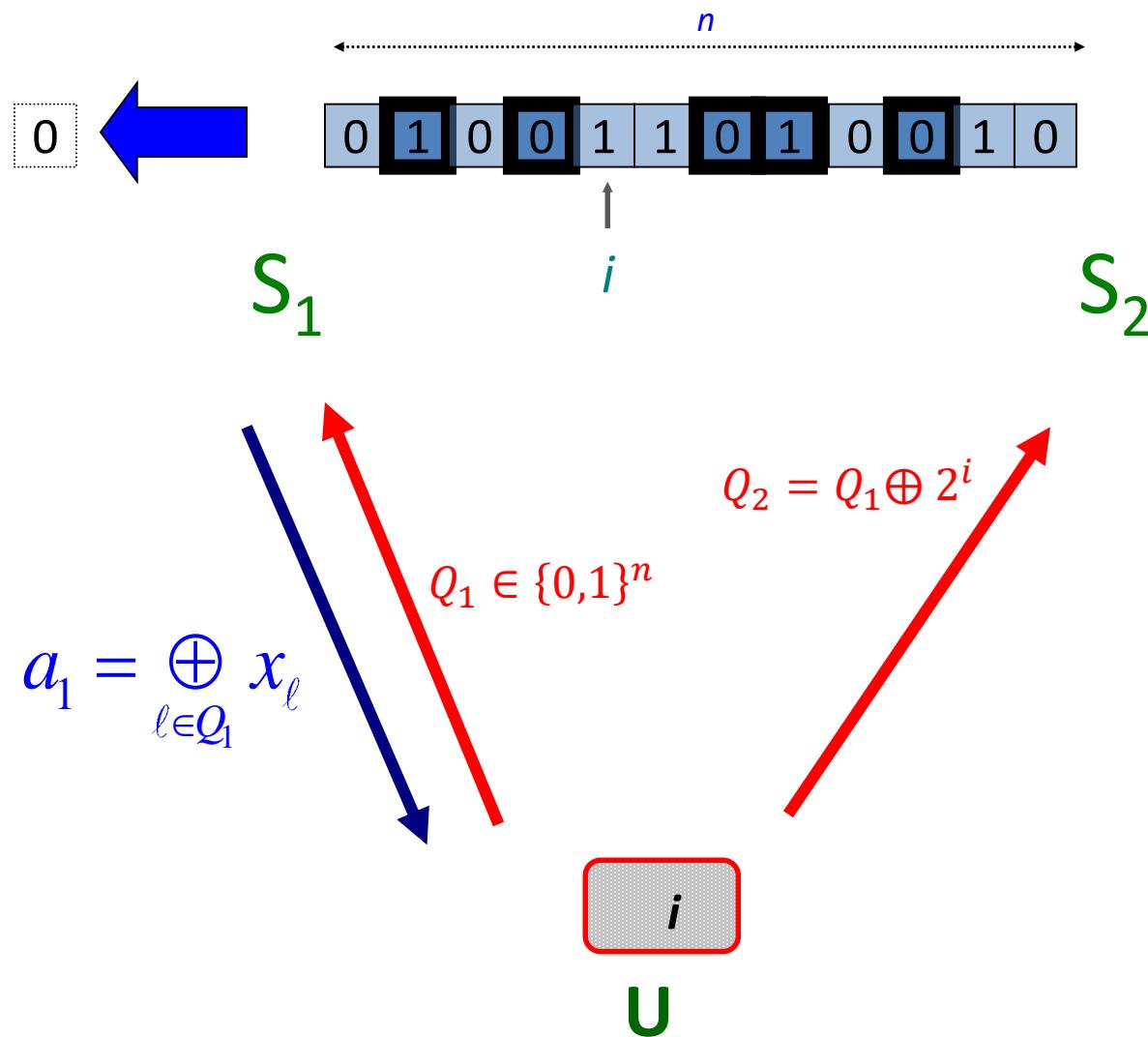
- Best Known Protocol: comm. $n^{1/3}$
- Open Question: Is this optimal?
- We first discuss two protocols
 - Protocol I:
 - n bit queries, 1 bit answers
 - Protocol II:
 - $n^{1/2}$ bit queries, $n^{1/2}$ bit answers

Protocol I: 2-server PIR

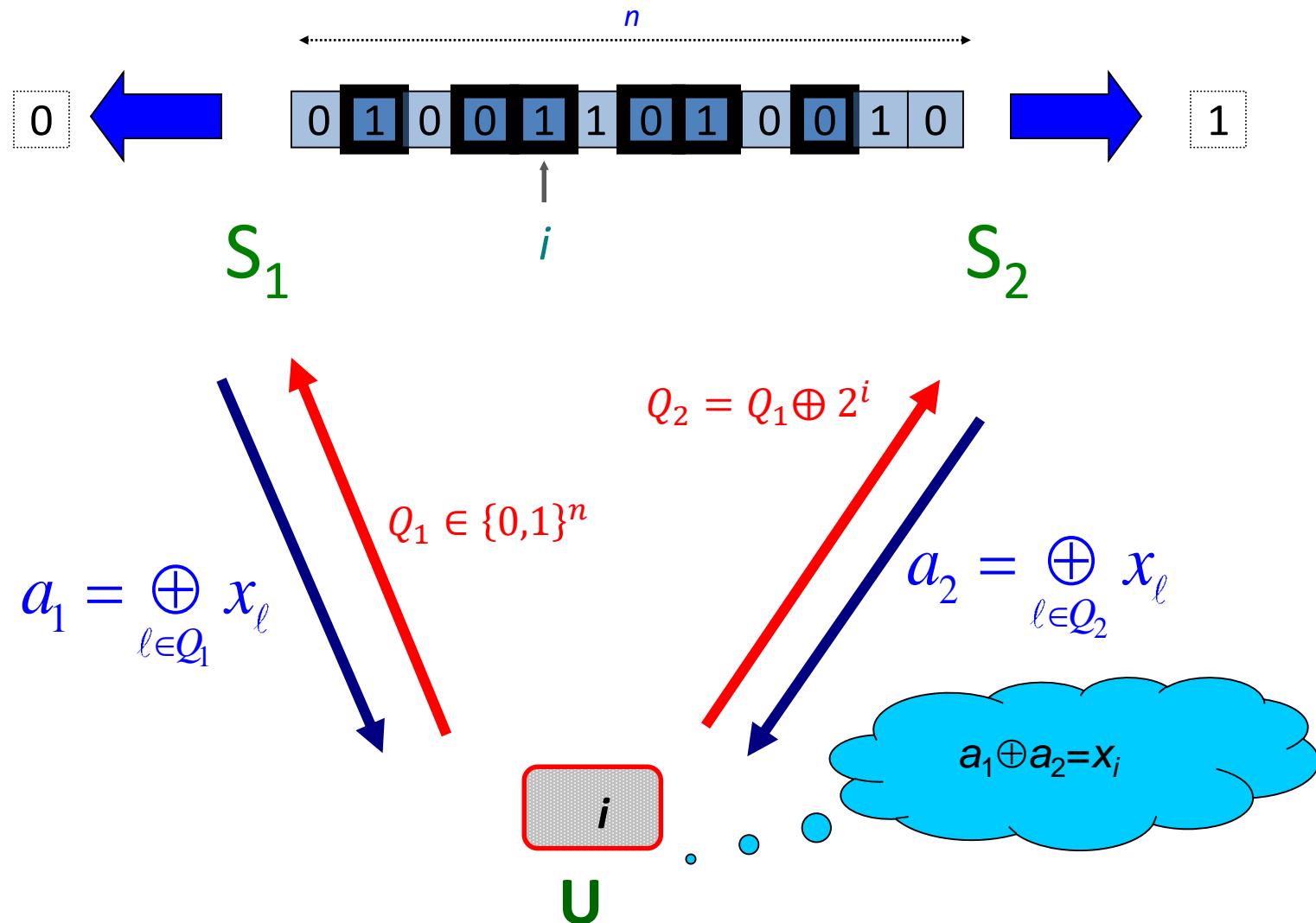
- n bit queries, 1 bit answers



Protocol I: 2-server PIR ...

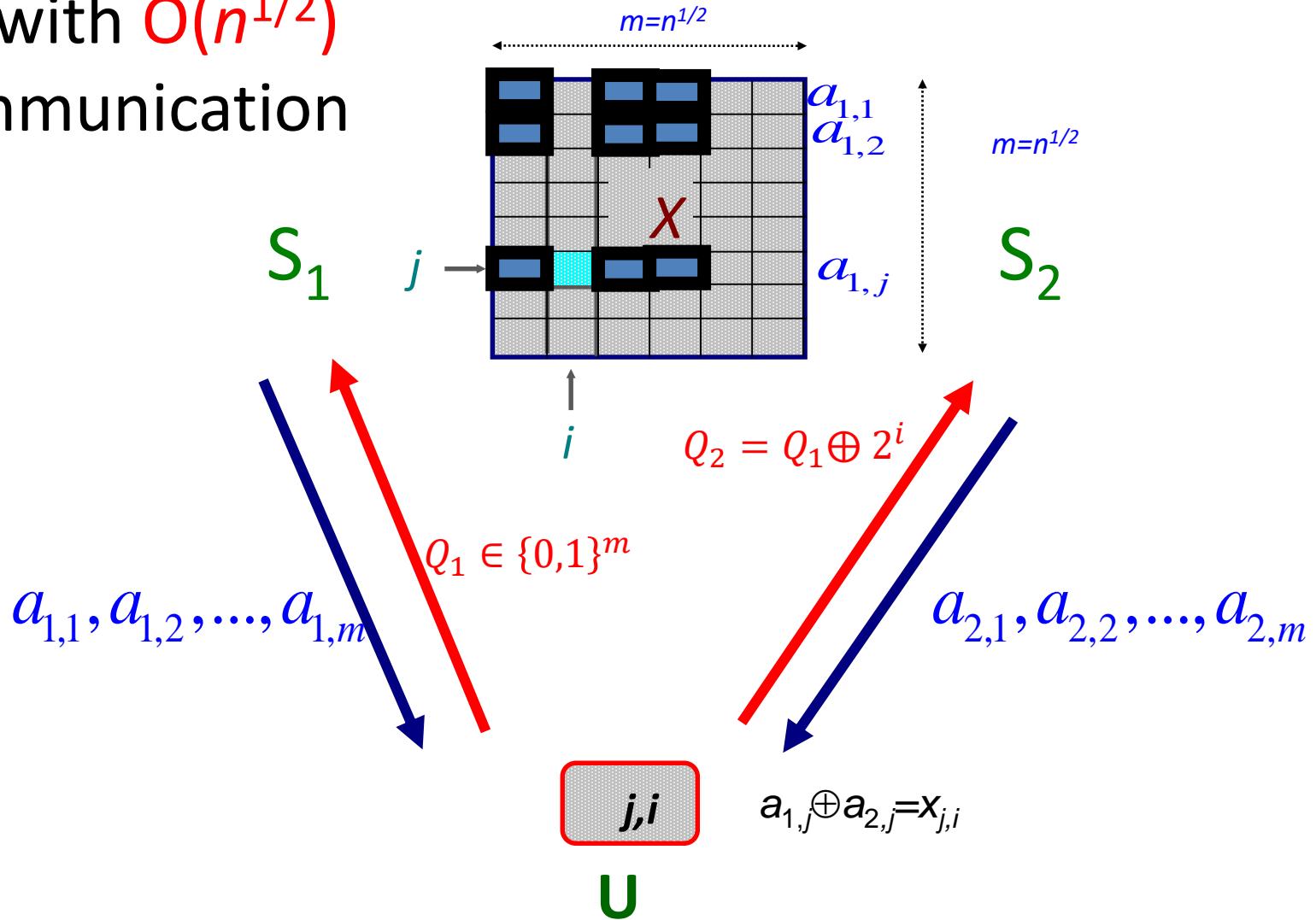


Protocol I: 2-server PIR ...



Protocol II: 2-server PIR

- PIR with $O(n^{1/2})$ Communication

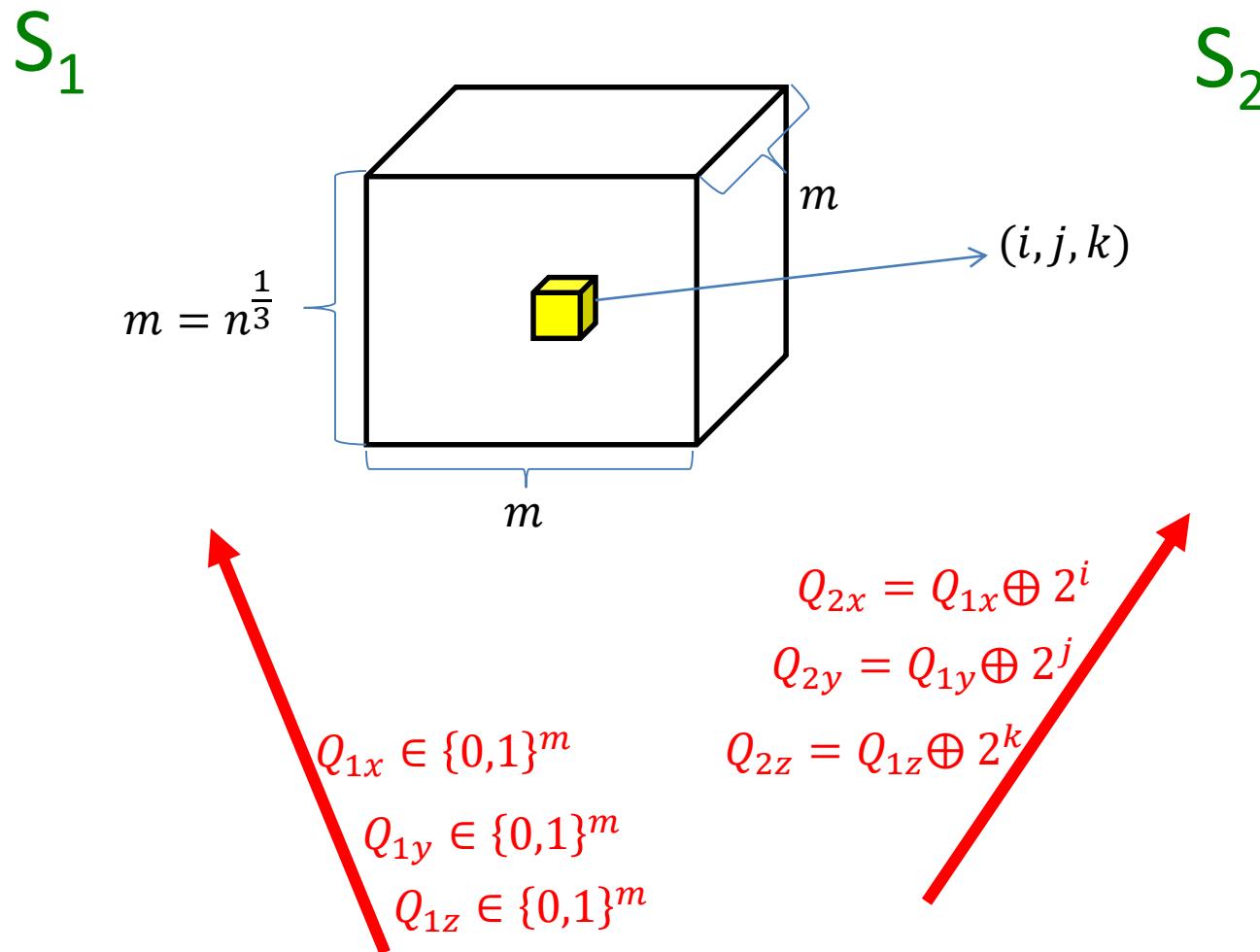


Question

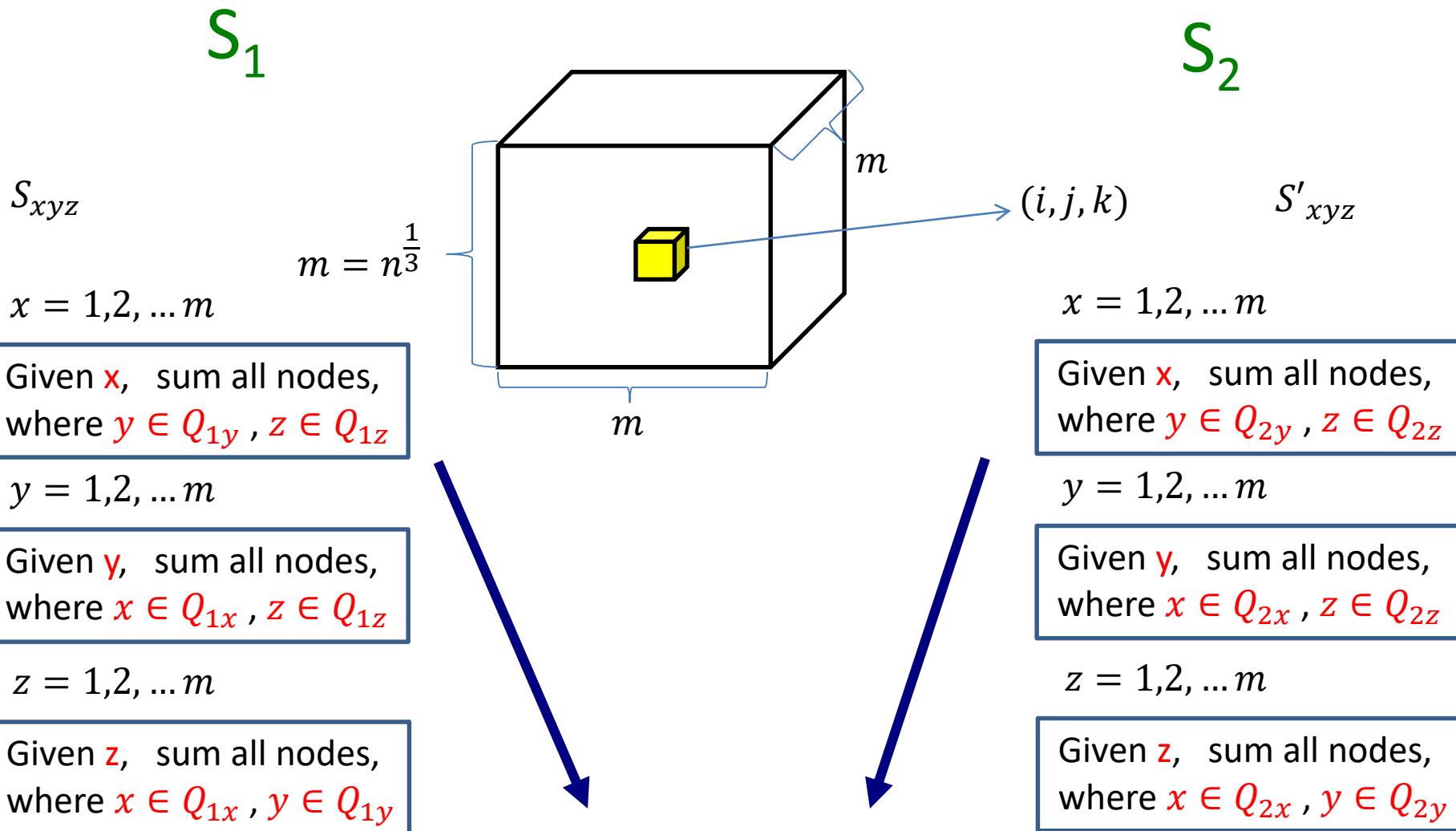
- How to design Information-Theoretic 2-Server PIR with comm. $n^{1/3}$



Information-Theoretic 2-Server PIR with comm. $n^{1/3}$



Information-Theoretic 2-Server PIR with comm. $n^{1/3}$ (2)



Information-Theoretic 2-Server PIR with comm. $n^{1/3}$ (3)

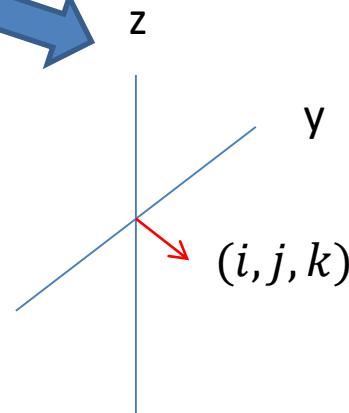
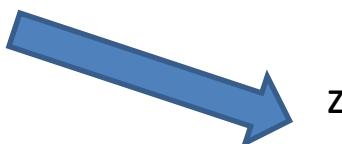
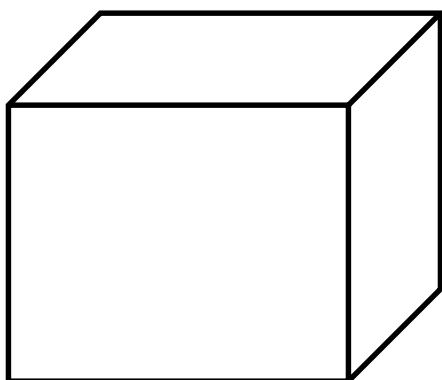
 S_{xyz} $X(i, j, k)$ S'_{xyz} $x = 1, 2, \dots m$

Given x , sum all nodes,
where $y \in Q_{1y}, z \in Q_{1z}$

 $x = 1, 2, \dots m$

Given x , sum all nodes,
where $y \in Q_{2y}, z \in Q_{2z}$

Choose $x = i$, compute $S_{iyz} \oplus S'_{iyz}$



Information-Theoretic 2-Server PIR with comm. $n^{1/3}$ (4)

$$S_{xyz} \quad (i, j, k) \quad S'_{xyz}$$

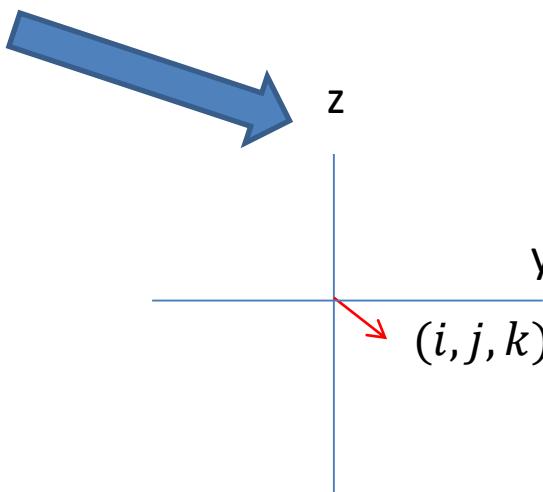
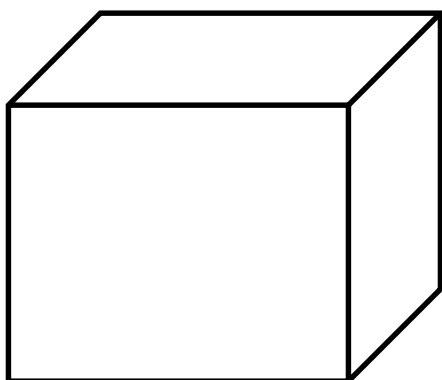
$y = 1, 2, \dots m$

Given y , sum all nodes,
where $x \in Q_{1x}, z \in Q_{1z}$

$y = 1, 2, \dots m$

Given y , sum all nodes,
where $x \in Q_{2x}, z \in Q_{2z}$

Choose $y = j$, compute $S_{xjz} \oplus S'_{xjz}$



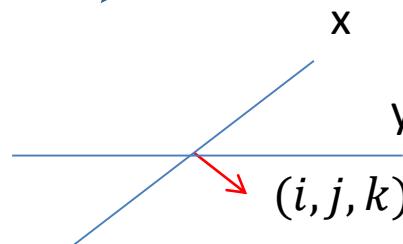
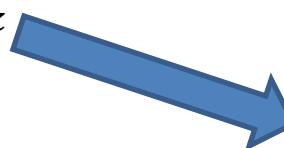
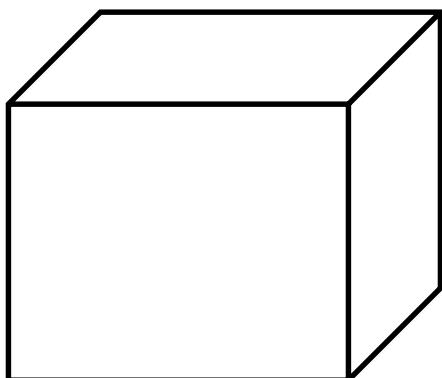
Information-Theoretic 2-Server PIR with comm. $n^{1/3}$ (5)

 S_{xyz} (i, j, k) S'_{xyz} $z = 1, 2, \dots m$ $z = 1, 2, \dots m$

Given z , sum all nodes,
where $x \in Q_{1x}, y \in Q_{1y}$

Given z , sum all nodes,
where $x \in Q_{2x}, y \in Q_{2y}$

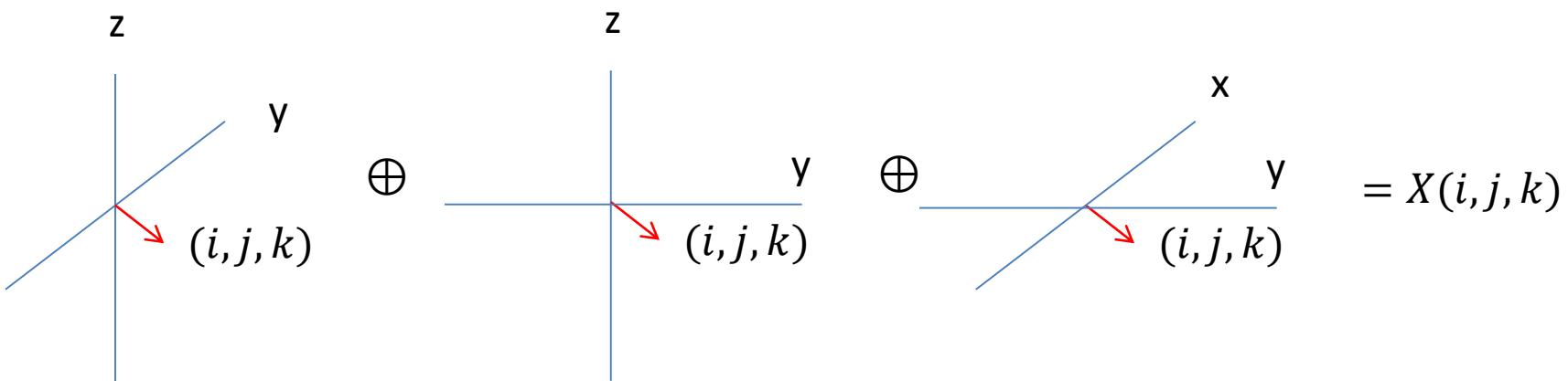
Choose $z = k$, compute $S_{xyk} \oplus S'_{xyk}$



Information-Theoretic 2-Server PIR with comm. $n^{1/3}$ (6)

- Compute

$$\begin{aligned} & S_{iyz} \oplus S'_{iyz} \oplus S_{xjz} \oplus S'_{xjz} \oplus S_{xyk} \oplus S'_{xyk} \\ & = X(i, j, k) \end{aligned}$$



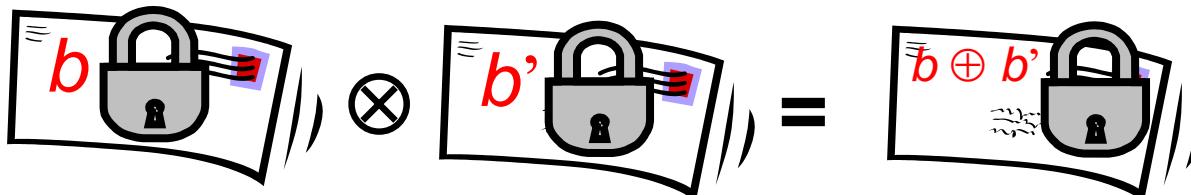
Information-Theoretic 2-Server PIR with comm. $n^{1/3}$ (7)

- Communication
 - User → Server
 - $3m + 3m = 6m = 6n^{\frac{1}{3}} = O(n^{\frac{1}{3}})$
 - Server → User
 - $3m + 3m = 6m = 6n^{\frac{1}{3}} = O(n^{\frac{1}{3}})$

Computational PIR with $O(n^{1/2})$ Comm.

(1 server)

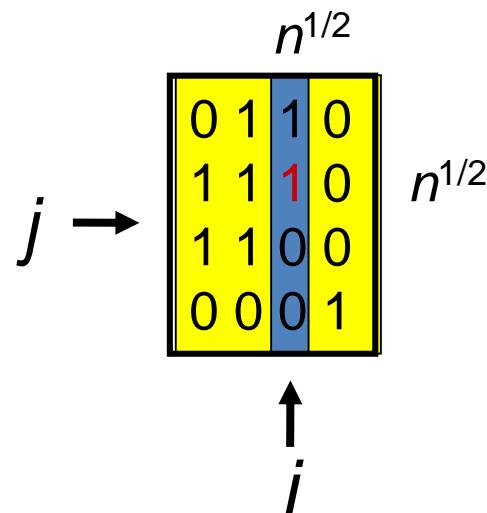
Tool: (randomized) homomorphic encryption



Example Paillier Encryption

- Key Generation
 - Input: two prime number p, q
 - Compute $n = pq, \lambda = lcm(p - 1, q - 1)$
 - Choose $g \in Z_{n^2}^*$ such that $\gcd(L(g^\lambda \bmod n^2), n) = 1$ with $L(u) = \frac{u-1}{n}$,
compute $\mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$
 - Output: public key $pk = (n, g)$ private key $sk = (\lambda, \mu)$
- Encrypt: Input $m \in Z_n$
 - Choose $r \in Z_n^*$, Compute $c = g^m \cdot r^n \bmod n^2$
- Decrypt: Input $c \in Z_{n^2}$
 - Compute $m = L(c^\lambda \bmod n^2) \cdot \mu \bmod n$

Computational PIR with $O(n^{1/2})$ Comm. (2)



- User sends $n^{\frac{1}{2}}$ encryptions
 $c_1 = E(0), c_2 = E(0), c_3 = E(1), c_4 = E(0)$
- For each row Server sends a “bit”

$$C_2 \times C_3 = E(0 + 1) = E(1)$$

$$C_1 \times C_2 \times C_3 = E(0 + 0 + 1) = E(1)$$

$$C_1 \times C_2 = E(0 + 0) = E(0)$$

$$C_4 = E(0)$$

- User recovers i -th column of x , $E(1) \Rightarrow 1$

Question

- How to design Computational PIR with $O(n^{1/2})$ Comm. (one server)
- Not a bit information, i.e., $x_i \in Z_N, N = pq$

x10	x9	x8	x7	x6	x5	x4	x3	x2	x1
-----	----	----	----	----	----	----	----	----	----



Question

- How to design Computational PIR with $O(n^{1/3})$ Comm. (one server)



Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 7: Oblivious Transfer Protocols

Lecturer: Rongxing LU

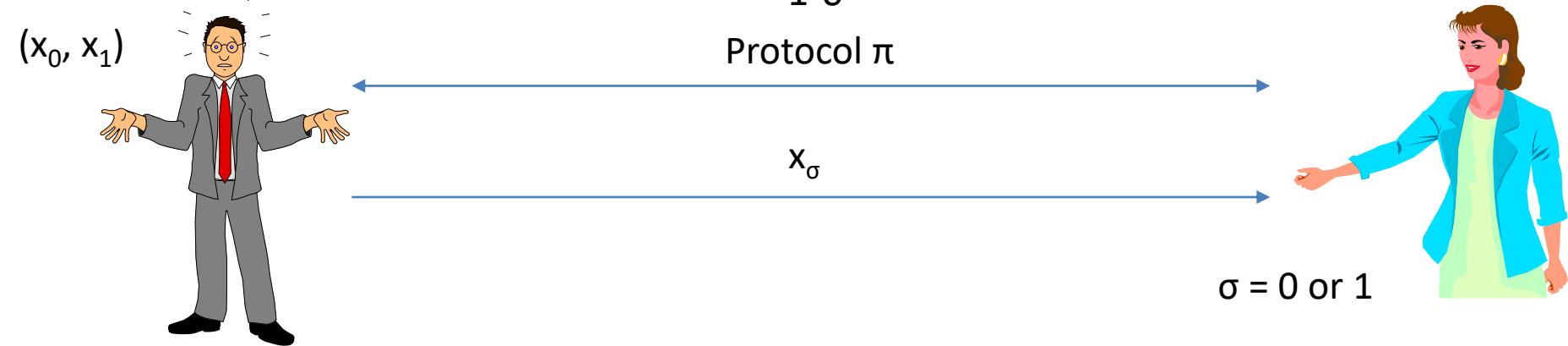
Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

Oblivious Transfer Protocols

- Protocol by which sender sends information to receiver, but remains oblivious as to what is received.
- 1-out-2 Oblivious transfer**
 - Bob does not know σ
 - Alice does not know $x_{1-\sigma}$

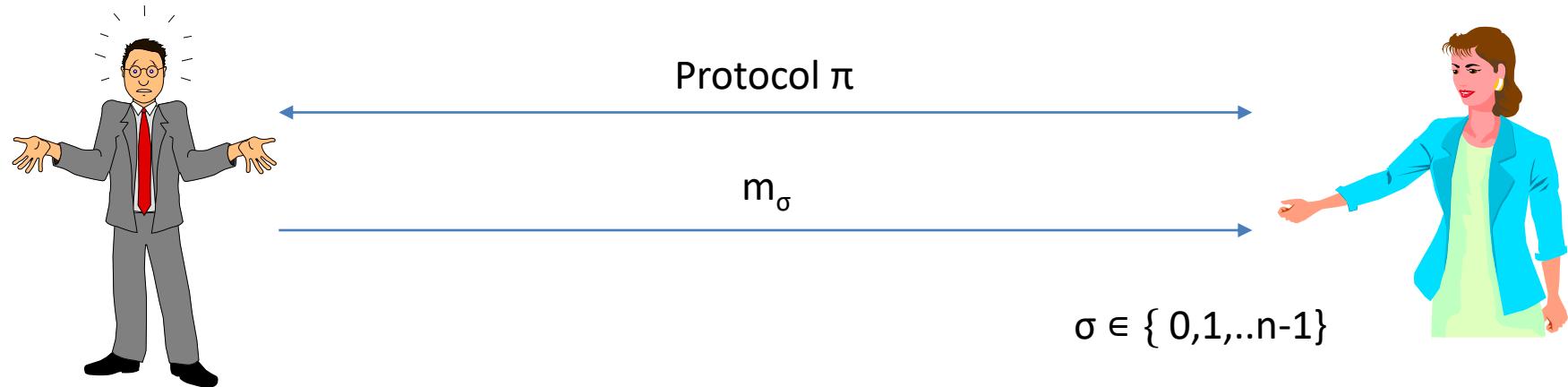


1-out-n Oblivious Transfer Protocols

- 1-out-n Oblivious transfer derived as natural generalization of 1-out-2 Oblivious transfer
 - Bob does not know σ
 - Alice can only know m_σ , not else

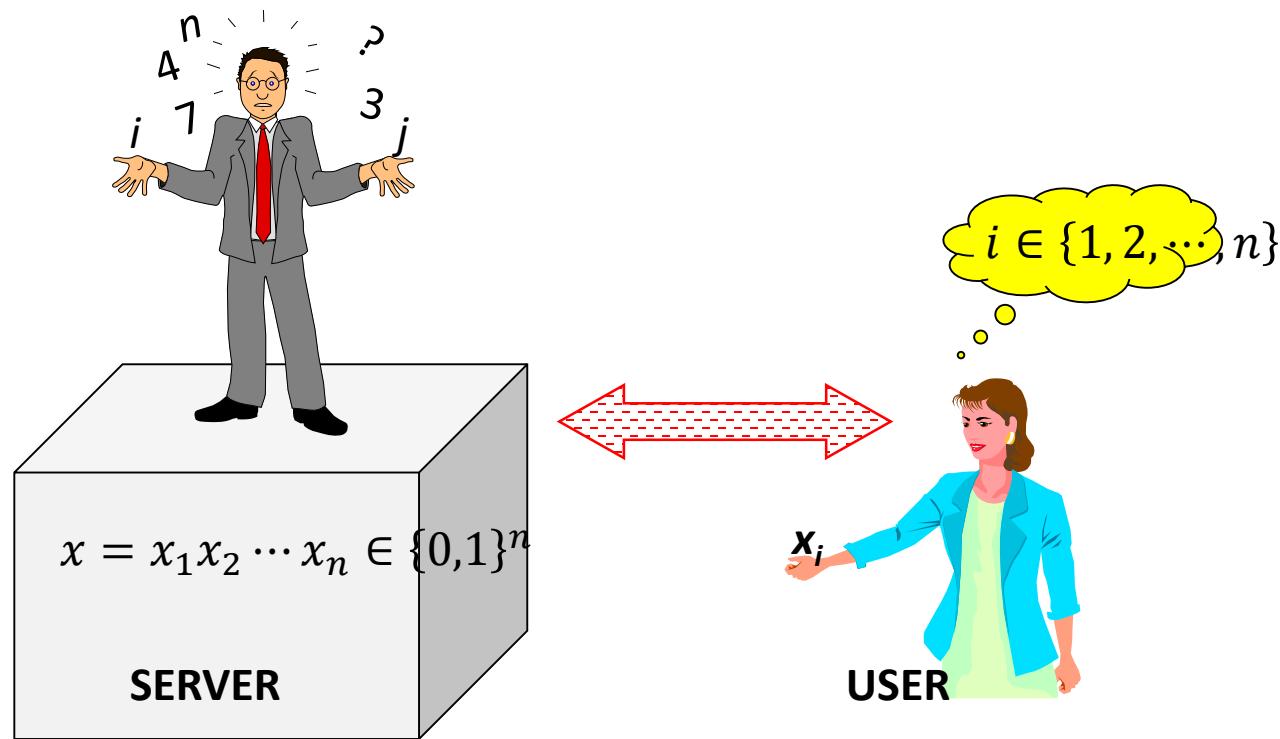
Private database

$(m_0, m_1 \dots m_{n-1})$



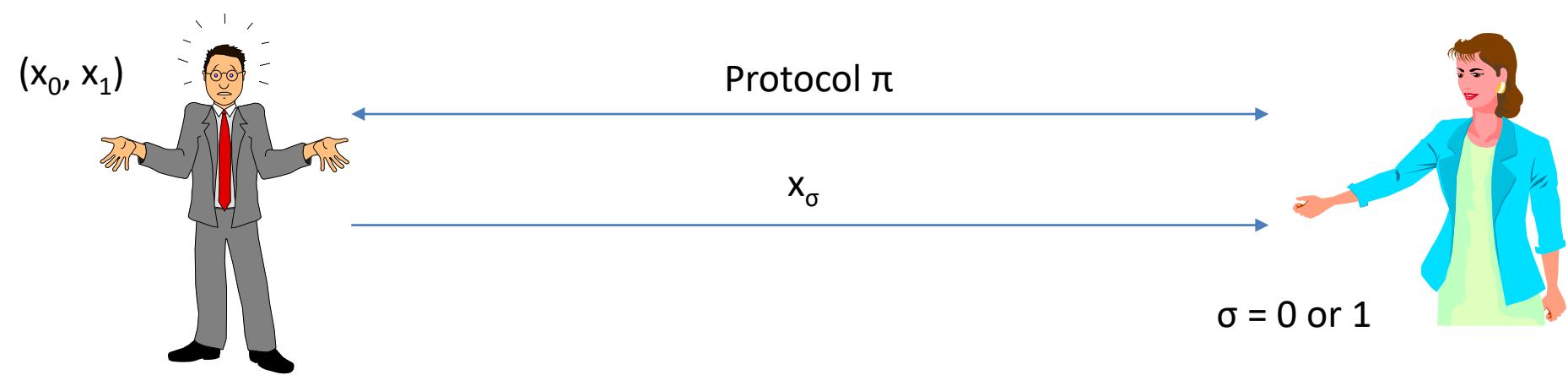
What is difference between OT (Oblivious Transfer) and PIR (Private Information Retrieval)

- Server: cannot know x_i in both OT and PIR
- User:
 - Can only know x_i in OT
 - Can know x_i and others in PIR



Why do we need to study OT?

- Motivation
 - Exchange of Secrets
 - Secure Multiparty Computation



A Simple 1-out-2 OT Protocol

honest-but-curious model

Input messages m_0, m_1

$\sigma \ k$ Choice bit σ , random k

RSA key pair d

$N, e \xrightarrow{} N, e$

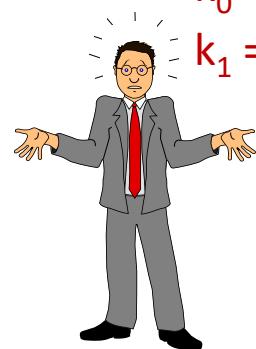
Random strings

$r_0, r_1 \xrightarrow{} r_0, r_1$

$$v \xleftarrow{} v = (r_\sigma + k^e) \bmod N$$

$$k_0 = (v - r_0)^d \bmod N$$

$$k_1 = (v - r_1)^d \bmod N$$



Sender (Bob)

$$\begin{aligned} m'_0 &= m_0 + k_0 \\ m'_1 &= m_1 + k_1 \end{aligned}$$

$$\begin{aligned} m'_0 \\ m'_1 \end{aligned}$$

$$m_\sigma = m'_\sigma - k$$



Receiver (Alice)

Involves exponentiations!

Protocol I: 1-out-2 OT based on Trapdoor Permutations (honest-but-curious model)

- Sender's input: $m_0, m_1 \in \{0,1\}^n$
- Receiver's input: $i \in \{0,1\}$
- Sender's first message
 - Select f from a trapdoor permutation family from $\{0,1\}^n$ to $\{0,1\}^n$.
- Receiver's message
 - Select x_{1-i} at random in $\{0,1\}^n$ and $x_i = f(w)$ where w is chosen randomly in $\{0,1\}^n$. Send x_0, x_1 to sender.
- Sender's message
 - let $w_i = f^{-1}(x_i)$, and send $w_i \oplus m_i$ for $i = 0,1$
- Receiver's output
 - If $x_1 = f(w)$, compute $w_1 \oplus m_1 \oplus w = m_1$

Protocol II: 1-out-2 OT based on Trapdoor Permutations (honest-but-curious model)

- Sender's input: $m_0, m_1 \in \{0,1\}^n$
- Receiver's input: $i \in \{0,1\}$
- Sender's first message
 - Select f from a trapdoor permutation family from $\{0,1\}^n$ to $\{0,1\}^n$.
 - Select $r_0, r_1 \in \{0,1\}^n$ and send r_0, r_1 to receiver
- Receiver's message
 - Select $i \in \{0,1\}$, a random $w \in \{0,1\}^n$
 - Send $x_i = r_i \oplus f(w)$ to sender.
- Sender's message
 - Recover $R_0 = f^{-1}(r_0 \oplus x_i)$ and $R_1 = f^{-1}(r_1 \oplus x_i)$, compute and send $R_0 \oplus m_0$ and $R_1 \oplus m_1$ to the receiver
- Receiver's output
 - If $i = 1$, m_1 can be correctly output $R_1 \oplus m_1 \oplus w = w \oplus m_1 \oplus w = m_1$

Which one is better?

Protocol I and Protocol II

- Communication
 - Protocol I: $3n$ comm. overheads
 - Protocol II: $5n$ comm. overheads
- Security
 - Protocol II is better, why?

Question

- How to design a Paillier Encryption based 1-out-2 OT protocol
- Consider Paillier encryption– pk public key, sk private key
 - Message $m \in Z_n$, $c = E_{pk}(m)$
 - Ciphertext $c = E_{pk}(m)$, $m = D_{sk}(c)$
 - Homomorphic properties



 $m_0, m_1 \in Z_n$

Solution

 pk, sk $r_0, r_1 \in Z_n$ $E_{pk}(r_0), E_{pk}(r_1)$ $\sigma \in \{0,1\}$ $w \in Z_n$ $c = E_{pk}(r_\sigma) + E_{pk}(w)$

recover $C_0 = Dec(c - E_{pk}(r_0))$ and $C_1 = Dec(c - E_{pk}(r_1))$

$$c_0 = C_0 + m_0 \bmod n$$

$$c_1 = C_1 + m_1 \bmod n$$

 c_0, c_1

If $\sigma = 1$

Compute $c_1 - w \bmod n = m_1$

DDH (Decisional Diffie-Hellman) based 1-out-2 OT Protocol

- q is a secure prime number
- Define $f(x) = s \cdot x + t \text{ mod } q$ is a function over Z_q . For $a \neq a' \in Z_q$, we have $\langle f(a), f(a') \rangle$ is distributed according to the uniform distribution over $Z_q \times Z_q$
- Discrete Logarithm problem: given g^a , cannot compute a
- DDH: g^a, g^b, g^c , decide $ab =? = c \text{ mod } q$

1-out-2 OT Protocol based on DDH



$$m_0, m_1 \in \{0,1\}^l$$

$$r, s \in Z_q$$

$$w = (g^a)^r g^s$$

$$\pi_0 = (g^{c_0})^r (g^b)^s$$

$$z_0 = \pi_0 \oplus m_0$$

$$\pi_1 = (g^{c_1})^r (g^b)^s$$

$$z_1 = \pi_1 \oplus m_1$$

$$p = 2q + 1 \quad \sigma \in \{0,1\}$$

$$g^q = 1 \bmod p \quad a, b \in Z_q$$

$$c_\sigma = a \cdot b \bmod q$$

$$c_{1-\sigma} \in Z_q$$

$$(p, q, g, g^a, g^b, g^{c_\sigma}, g^{c_{1-\sigma}})$$



w, z₀, z₁

$$\begin{aligned}\pi_\sigma &= w^b \\ m_\sigma &= \pi_\sigma \oplus z_\sigma\end{aligned}$$

Receiver's security

- Assume DDH is true. Then the receiver's message when $i = 0$ is **computationally indistinguishable from** the receiver's message when $i = 1$.

$$(p, q, g, g^a, g^b, g^{c_\sigma}, g^{c_{1-\sigma}})$$

$$\begin{aligned} c_\sigma &= a \cdot b \bmod q \\ c_{1-\sigma} &\in Z_q \end{aligned}$$

Sender's security

- If $c_i = a \cdot b$ for some $i \in \{0,1\}$, then π_{1-i} is a uniform element in Z_p , even conditioned on all the rest of the information the sender provides.
- This does indeed imply that m_{1-i} is completely hidden from the receiver.

Sender's security proof

- The main part of the receiver's message consists of a four-tuple $(g^a, g^b, g^{c_0}, g^{c_1})$. Since the sender verifies that $g^{c_0} \neq g^{c_1}$, we know that $c_0 \neq c_1$. Therefore, $c_{1-i} \neq a \cdot b \bmod q$.
- The sender selects random r and s and then computes $w = g^{ar+s} = g^{f(a)}$ where $f(x) = rx + s$. The number π_i is equal to $g^{abr+bs} = g^{bf(a)}$. The number π_{1-i} is equal to $g^{c_{1-i}r+bs} = g^{b(r \cdot \frac{c_{1-i}}{b} + s)} = g^{b \cdot f(\frac{c_{1-i}}{b})}$.
- Now the only parts of the sender's message that depend on s and r are the values $w = g^{f(a)}$ and $z_i = g^{bf(a)} \oplus x_i$. Since these depend only on $f(a)$, even conditioned on these values, the value $f(a')$ for $a' = \frac{c_{1-i}}{b}$ is completely random.

1-out-n OT Protocol

 $x_1, x_2, \dots, x_n \in \{0,1\}^l$ $i \in \{1,2,\dots,n\}$ 

Choose $k_0 \in 0^l$

For $j \in \{1,2,\dots,n\}$

 $k_j \in \{0,1\}^l$

$$\begin{cases} k_0 \oplus \dots \oplus k_{j-1} \oplus x_j \\ k_j \end{cases}$$

1-out-2 OT to get either x_j or k_j

The receiver can learn k_j for all $j \neq i$ and $k_0 \oplus \dots \oplus k_{i-1} \oplus x_i$, so she can recover x_i .

Protocol I (1-out-2): $3l$ comm. Overheads \rightarrow (1-out-n): $3l \cdot n$ comm. Overheads $O(n)$
Protocol II (1-out-2) : $5l$ comm. overheads \rightarrow (1-out-n): $5l \cdot n$ comm. Overheads $O(n)$

1-out-n OT Protocol based on DDH

$$x_1, x_2, \dots, x_n \in \{0,1\}^l$$

$$p = 2q + 1 \quad \sigma \in \{1, 2, \dots, n\}$$



$$g^q = 1 \bmod p \quad a, b \in Z_q$$

$$\begin{aligned} c_\sigma &= a \cdot b \bmod q, \\ c_j &\in_R Z_q \text{ for } j \neq \sigma \end{aligned}$$



$$(p, q, g, g^a, g^b, g^{c_1}, \dots, g^{c_n})$$

$$r, s \in Z_q$$

$$w = (g^a)^r g^s$$

for $i = 1, 2, \dots, n$

$$\pi_i = (g^{c_i})^r (g^b)^s$$

$$z_i = \pi_i \oplus x_i$$

$$W, Z_1, Z_2, \dots, Z_n$$

$$\pi_\sigma = w^b$$

$$x_\sigma = \pi_\sigma \oplus z_\sigma$$

Comm. Overheads $O(n)$

Question

- How to use the Paillier encryption to design 1-out-n OT with $O(n)$ Comm.



 $m_1, m_2, \dots, m_n \in Z_N$

Solution

 pk

pk, sk

$r_1, r_2, \dots, r_n \in Z_N$

$E_{pk}(r_0), E_{pk}(r_1), \dots, E_{pk}(r_n)$

$\sigma \in \{1, 2, \dots, n\}$

$w \in Z_N$

c

$c = E_{pk}(r_\sigma) + E_{pk}(w)$

recover $C_i = Decrypt(c - E_{pk}(r_i))$

for $i = 1, 2, \dots, n$

$c_i = C_i + m_i \bmod N$

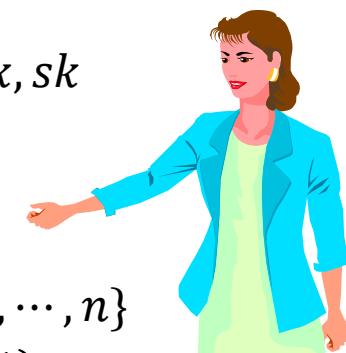
c_1, c_2, \dots, c_n

If $\sigma = 1$

Compute $c_1 - w \bmod N = m_1$

 $m_1, m_2, \dots, m_n \in Z_N$

Solution

 pk, sk

$\sigma \in \{1, 2, \dots, n\}$
 $c_\sigma = E(1);$

$for i = 1, 2, \dots, n \text{ and } i \neq \sigma$
 $c_i = E(0);$

$c = \prod_{i=1}^n c_i^{m_i}$

c_1, c_2, \dots, c_n

c

Recover c to get

$0 \cdot m_1 + 0 \cdot m_2 + \dots + 1 \cdot m_\sigma + 0 \cdot m_{\sigma+1} + \dots + 0 \cdot m_n = m_\sigma$

Question

- How to use the Paillier encryption to design 1-out-n OT with $O(\sqrt{n})$ Comm.



Question

- How to design a strong 1-out-n OT with $O(\sqrt{n})$ Comm.



Question

- How to design a even stronger 1-out-n OT with $O(\sqrt{n})$ Comm.



Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 8: Zero knowledge proof techniques

Lecturer: Rongxing LU

Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

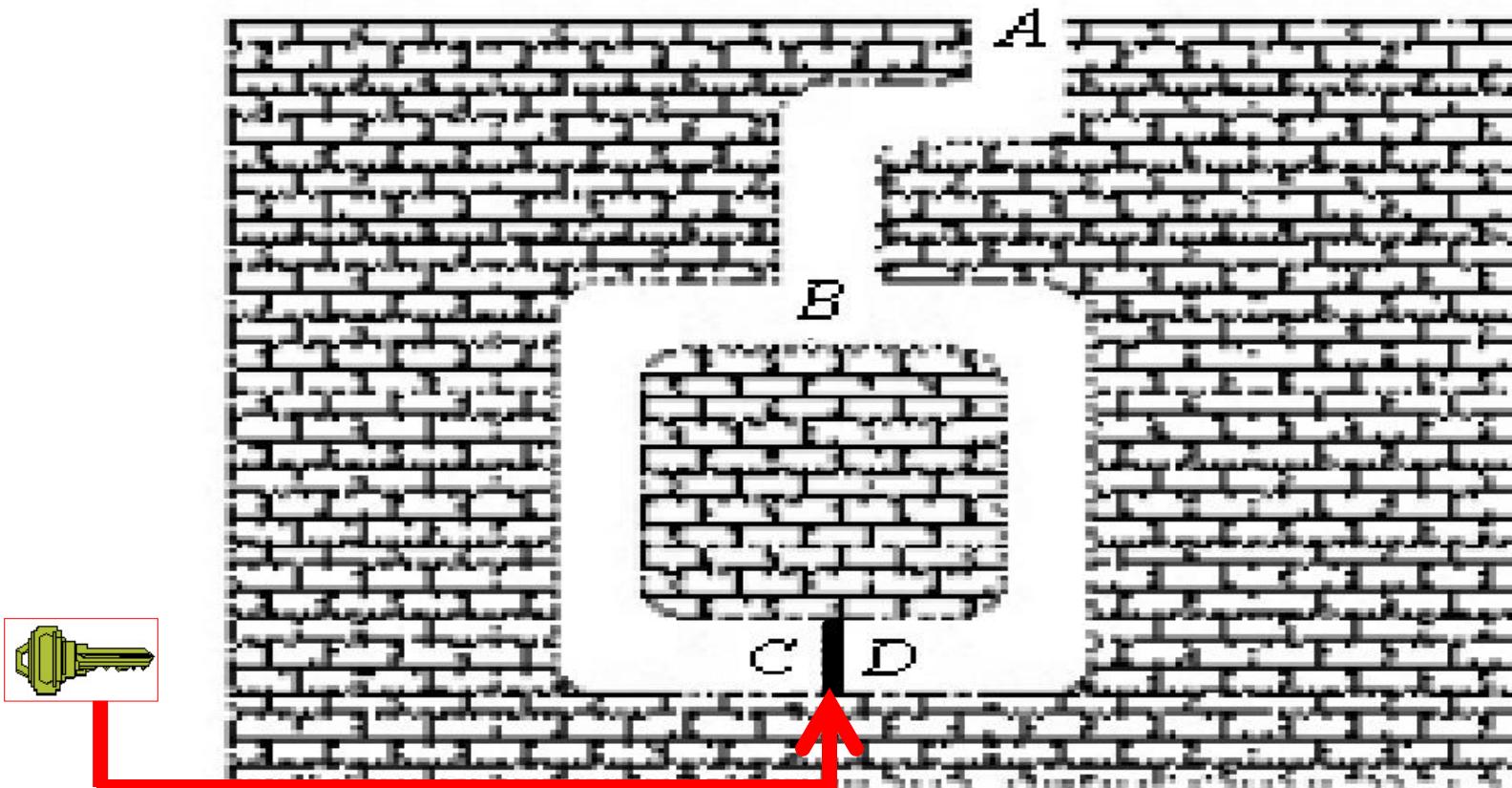
Zero-Knowledge Proof Identification Protocol

- 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,...)

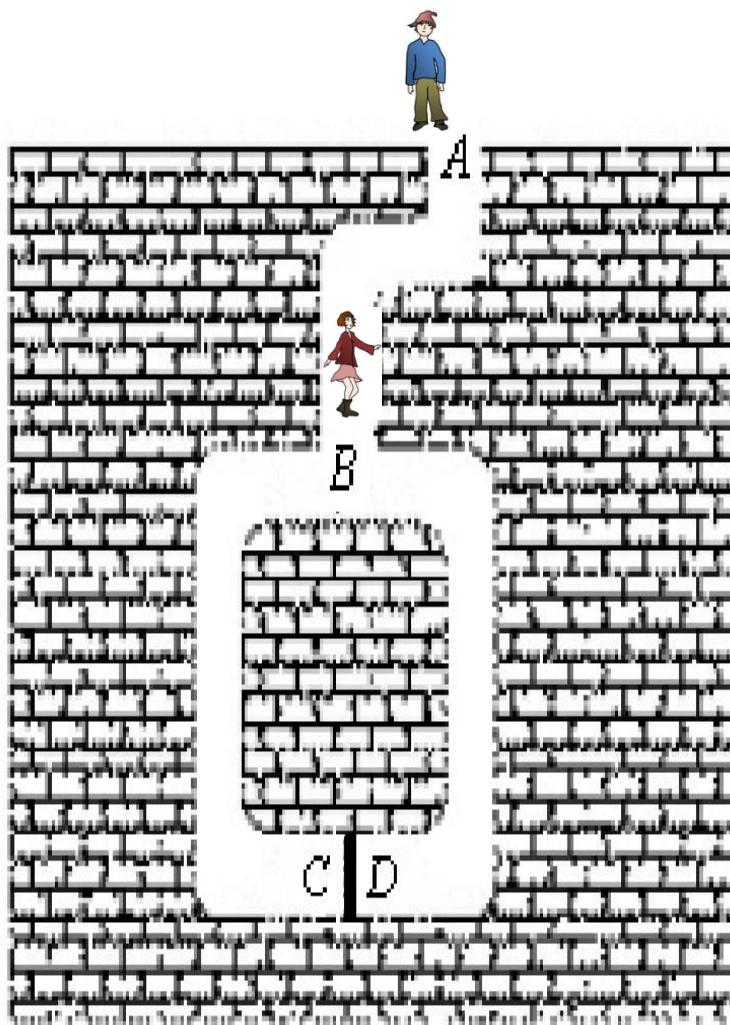


Zero-Knowledge Proof (Example)

- Alice knows a secret word opening the door in cave. How can she convince Bob about it without revealing this secret word?



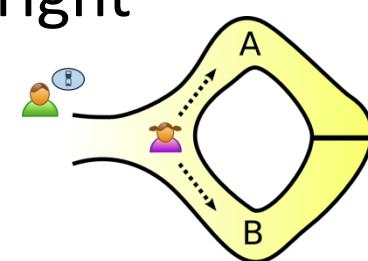
Zero-Knowledge Proof (Example)



1. Bob stands at Location A.
2. Alice walks into the cave, either location C or location D.
3. After Alice stops at either C or D, Bob walks at Location B.
4. Bob cries out, asks Alice to Either walk out from left side Or walk out from right side
5. Alice replies ok, and walks out from the right side.
6. Repeat (1)-(5) steps with n times. $\left(\frac{1}{2}\right)^n$

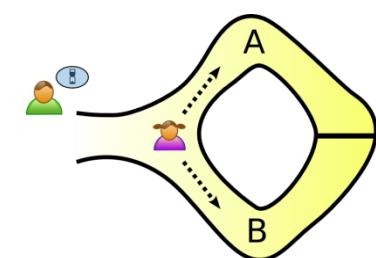
Zero-Knowledge Proof (Example)

- **Proof.**
- (1) If Alice has the secret word, Alice can always go out from the right side, no matter she stops at C or D, because she can open the door with the secret word.
- (2) If Alice does not have the secret word, if she stops at the right side, she can come out without keys. But the probability of (stopping the right side) is $\frac{1}{2}$. Once the game is repeated with n times, the probability is very small, i.e., $\left(\frac{1}{2}\right)^n$. When n=10, the probability is $1/1024 = 0.0009765$
- Therefore, if Alice can always go out from the right side, she should have the secret word.



Properties of Interactive Proof

- **Completeness:** if a statement is true, it can be proven.
- **Soundness:** if a statement is false, it cannot be proven.
- **Zero-knowledgeness:** any verifier does not learn anything except that a statement is true.

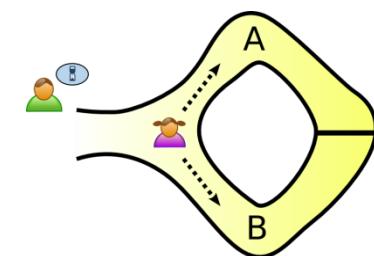


Knowledge needs to be captured by **VIEW**

- The **View** is a probability distribution over all P , $V^*(x)$, i.e., over all possible conversations that P and V^* might have about x .

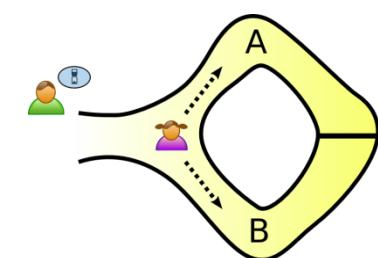
$$VIEW_{P,V^*(x)} = \langle \text{msg from } P, \text{randomness of } V^* \rangle$$

- Given a probabilistic verifier, $VIEW_{P,V^*(x)}$ encapsulates its whole view because the entire set of messages with the prover can be reconstructed.



Zero-knowledgeness -- **VIEW**

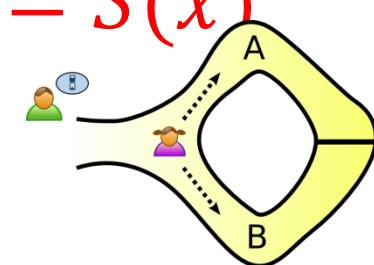
- For achieve ZK, the **VIEW** should be the same before and after communication with prover,
- i.e., the view will only consist of messages the **verifier** could have simulated themselves.



Formal Notion of Zero-knowledgeness

- For Zero-knowledgeness, we expect $\text{VIEW}_{P,V^*}(x) = S(x)$ where $S(x)$ is the view of a simulator that is capable of producing conversations selected from the same distribution as those of $P, V^*(x)$ and, as with the verifier, is bound on probabilistic polynomial time.

$$\forall V_{ppt}^* \exists S_{ppt} \forall x \in L \ s.t. \text{VIEW}_{P,V^*}(x) = S(x)$$



Example (Isomorphic Graph)

Prove $(G, H) \in ISO$

Prover

Knows $\emptyset(G) = H$

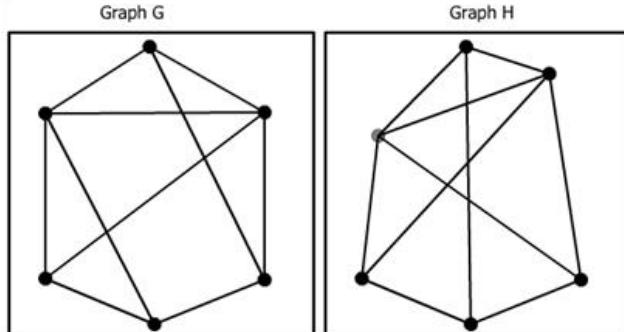
Pick random permutation π

$C = \pi(G) \rightarrow$

$\leftarrow "H"$

If “G” is received, set $\alpha = \pi$

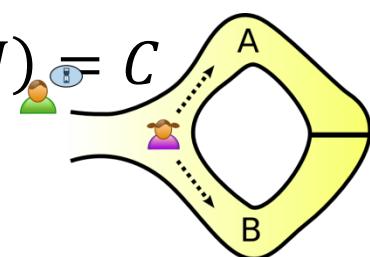
If “H” is received, set $\alpha = (\pi \cdot \phi^{-1})$



$\alpha \rightarrow$

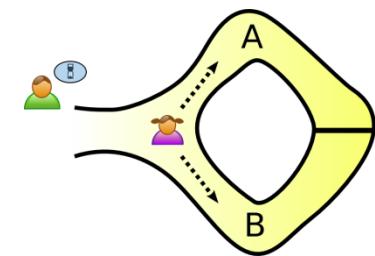
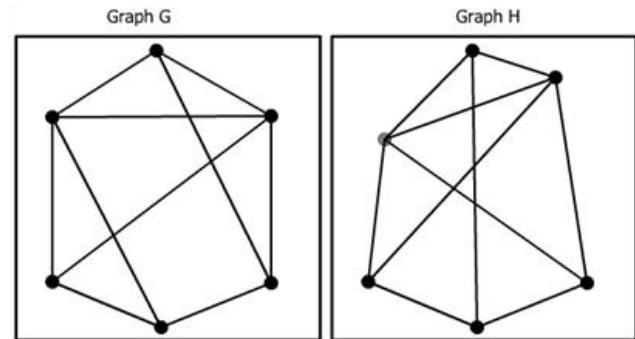
Pick G or H at hand, say H , show
 H is iso to C

Check that $\alpha(H) = C$

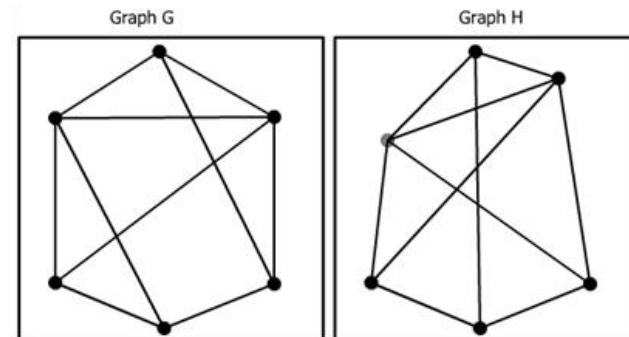


Analysis

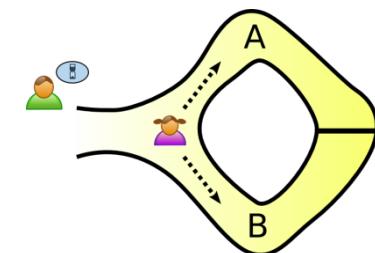
- Completeness
 - The verifier will be convinced that the graphs are isomorphic
- Soundness
 - Support G and H are not isomorphic
 - Prover cannot generate a C isomorphic to both G and H
 - Prover has $\frac{1}{2}$ probability to choose the same graph as the verifier
 - Probability of successful cheating is $\frac{1}{2^k}$ after k attempts.



Analysis (2)

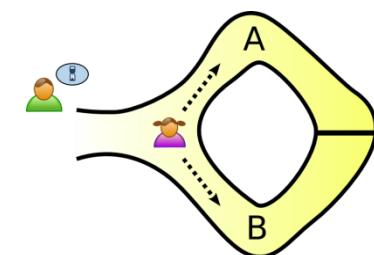


- Show this proof meets our zero-knowledgeness property, i.e., showing $\text{VIEW}_{P,V^*}(x) = S(x)$
1. We start by constructing the simulator:
 - a) Pick either graph G or H at random. (say G)
 - b) Choose a random permutation α
 - c) Compute $C = \alpha(G)$
 - d) Now, choose a random tape r for V^* and run it on input C
 - e) If V^* responses with “G”, return α and record (C, α) as the messages from the “prover” and the random tape r as V^* ‘s randomness
 - f) If V^* responses with “H”, abort, don’t record anything and start over at the beginning.



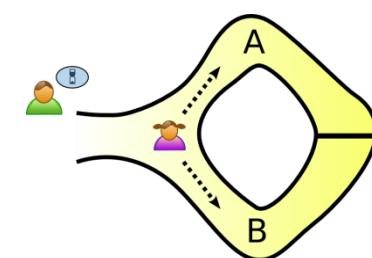
Fiat-Shamir ZK Identification Scheme

- Zero knowledge proofs can be used to cryptographically **identify parties**.
- Each party has a secret key and a public key.
- The prover i)convinces the verifier that he knows his secret key, ii)without revealing any information on his secret key that iii) the verifier could not know otherwise (except that the proven claim holds).



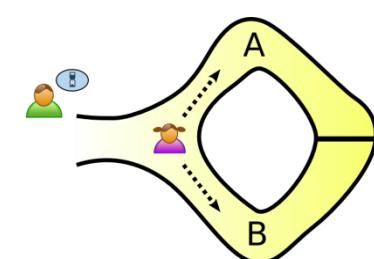
Fiat-Shamir ZK Identification Scheme (cont.)

- In the Fiat-Shamir scheme, the prover has an RSA modulo $n = pq$ whose factorization is secret. The factors themselves are not used in the protocol.
- Unlike in RSA, a center can generate a universal n , used by everyone, as long as nobody knows the factorization.
- The center itself should forget the factorization just after he computes n .



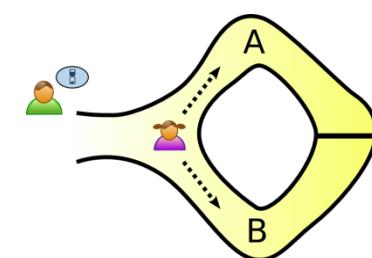
Fiat-Shamir ZK Identification Scheme (cont.)

- The Secret Key: The prover chooses a random value $1 < S < n$ (to be served as the secret key) ($\gcd(S, n) = 1$) and keeps it secret.
- The Public Key: The prover computes $I = S^2 \bmod n$, and publishes the pair I and n as the public key.
- The purpose of the protocol: The prover has to convince the verifier that he knows the secret key S corresponding to the public key (I, n) , (i.e., to prove that he knows a modular square root of I modulo n), without revealing S .



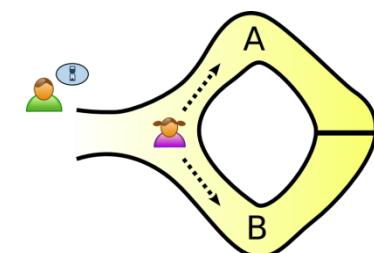
Fiat-Shamir ZK Identification Scheme (cont.)

- 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 .
 - $I = S^2 \bmod n$



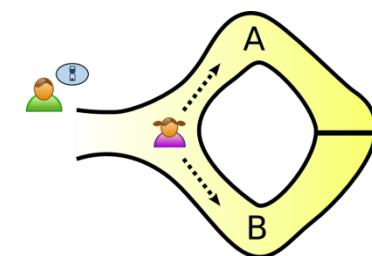
Fiat-Shamir ZK Identification Scheme (cont.)

1. The prover chooses a random value $1 < R < n$, and computes $X = R^2 \bmod n$.
2. The prover sends X to the verifier.
3. The verifier requests from the prover one of the following requests at random:
 - R , or
 - $RS \bmod n$.
4. The prover sends the requested information to the verifier.



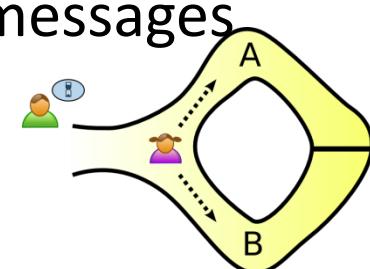
Fiat-Shamir ZK Identification Scheme (cont.)

5. The verifier verifies that he received the correct answer by checking whether:
 - $R^2 \equiv? X \pmod{n}$
 - Or $(RS)^2 \equiv? IX \pmod{n}$
6. If the verification fails, the verifier concludes that the prover does not know S , and thus he is not the claimed party.
7. This protocol is repeated t (usually 20, 30, or $\log n$) times, and if in all of them the verification succeeds, the verifier concludes that the prover is the claimed party.



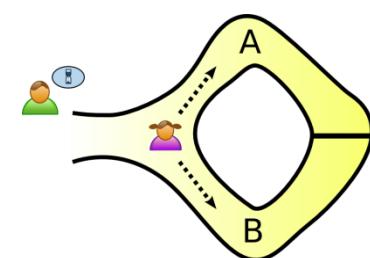
The Protocol does not Reveal Information

- We show that no information is revealed on S from the protocol:
- Clearly, when the prover sends X or R , he does not reveal any information on S .
- When the prover sends $RS \bmod n$:
 - $RS \bmod n$ is random, since R is random and $\gcd(S, n) = 1$.
 - If somebody can compute some information on S from I, n, X and $RS \bmod n$, he can also compute the same information on S from I and n , since he can choose $T = R'S \bmod n$ at random, and compute $X' = T^2 \cdot I^{-1} \bmod n$, from which he can compute the information on S .
- Thus, the verifier, and anybody else, cannot gain any information on S using the protocol, or from the messages transmitted in the protocol.



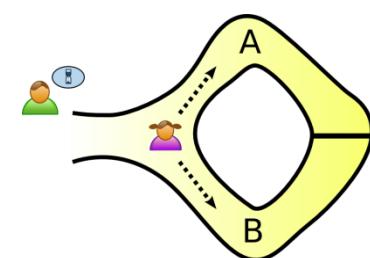
Security

- Clearly, if the prover knows S , the verifier is convinced in his identity. If the prover does not know S , he can either
 - know R , but not $RS \bmod n$, as he is choosing R , but cannot multiply it by the unknown value S , or
 - choose $RS \bmod n$, and thus can answer the second question $RS \bmod n$, but in this case he cannot answer the first question R , since he needs to divide by the unknown value S .



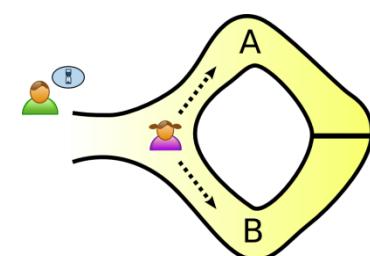
Security (cont.)

- In any case, he cannot answer both questions, since then he can compute S as the ratio between the two answers. But it is assumed that computing S is difficult, actually the difficulty is equivalent to that of factoring n .
- Since the prover does not know in advance (when he chooses R or $RS \bmod n$) which question the verifier will ask, he cannot choose the required choice. He can succeed in guessing the verifier's question with probability $1/2$ for each question, and thus the verifier can catch him in half of the times, and fails to catch him half of the times. The protocol is repeated t times, and thus the probability that the verifier fails to catch the prover in all the times is only 2^{-t} , which is exponentially reducing with t .



Security (cont.)

- In particular, for $t = 20$, the prover succeeds to cheat less than once in a million trials, and for $t = 30$, the prover succeeds to cheat less than once in a billion trials. Verifiers wishing a smaller probability of error, can use larger t 's.
- The verifier cannot use the information he received in the protocol to convince others that he is the original prover, since he cannot answer both questions R and $RS \bmod n$ for any R . If he could, he would know S .

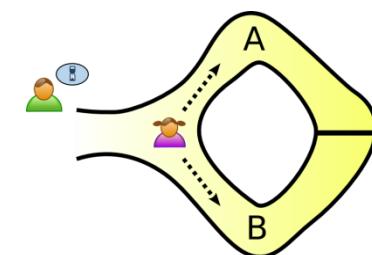


A Simulator for the Fiat-Shamir Scheme

- 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 $I \text{ modulo } N$. The output of the simulator is a forged transcript of a proof. A transcript for the problem is of the form:

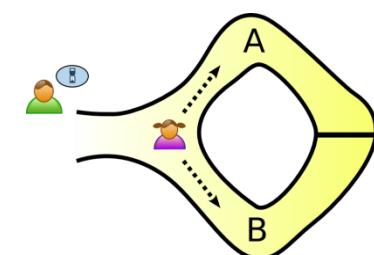
$$(I, N)(X_1, i_1, M_1) \cdots (X_n, i_n, M_n)$$

- where M_i is either the square root of X_i (in case $i_i = 1$), or the square root of IX_i (in case $i_i = 2$).



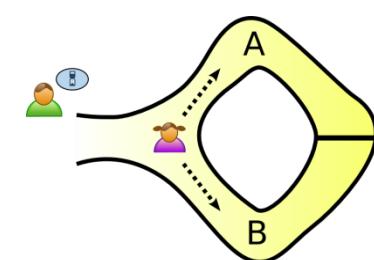
A Simulator for the Fiat-Shamir Scheme (cont.)

1. $T = (I, N)$
2. Do the following till n triples are found:
 - a) Let j be the round index $\{ 1, \dots, n \}$
 - b) Choose i_j to be 1 or 2 at random
 - c) Choose a random number $1 < R_j < N$
 - d) Compute $U_j = R_j^2$ if $i_j = 1$, and $U_j = R_j^2 \cdot I^{-1}$ if $i_j = 2$.
 - e) Call the original V with input U_j and obtain a challenge i_j'
 - f) If $i_j = i_j'$, concatenate the triple (U_j, i_j, R_j) to T
 - g) Otherwise, reset V 's state, and repeat this round with new random choices



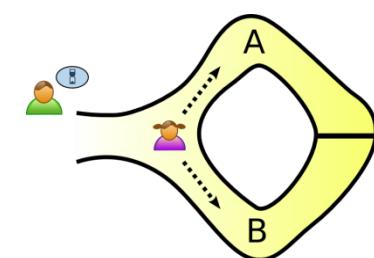
A Simulator for the Fiat-Shamir Scheme (cont.)

- The correctness of the simulator is derived from the following facts:
 - In each round the simulator has probability $1/2$ to guess the correct bit as the verifier. Therefore, we expect to find a valid triple every two trials. This yields a polynomial bound on the expected running time of the simulator.
 - In each round the relation between R_j and U_j can be verified correctly
 - The simulator does not need to know the root of I . Even if the chosen bit is $\textcolor{red}{2}$, still the equation $U_j \cdot I = R_j^2$ holds. Moreover, it is not detected even if I is a quadratic non-residue.
 - The distribution of the transcripts is the same as the distribution of real transcripts, since the random bit distribution is the same.



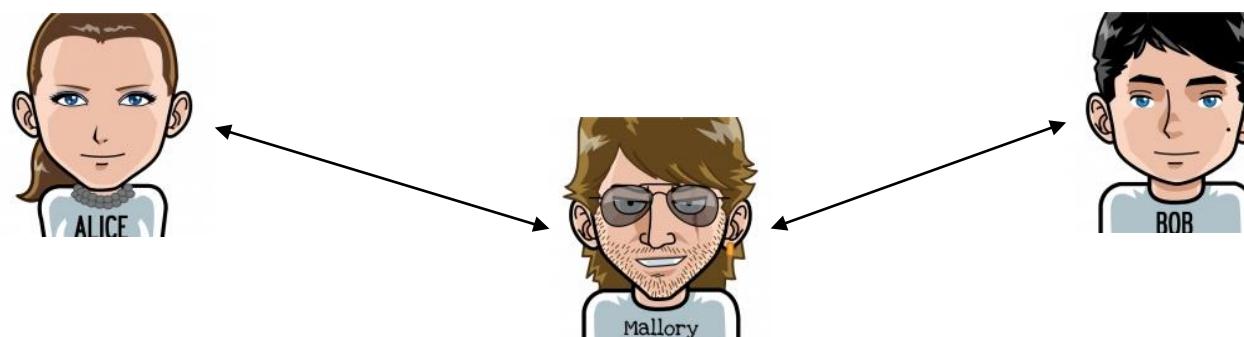
Parallel Fiat-Shamir

- We can apply all the rounds of Fiat-Shamir in parallel, instead of sending them sequentially. This modification makes the protocol more efficient.
- Assume we have a ZK system, for which the honest prover can always respond to V's challenges, whereas a dishonest prover can fool V with probability $1/2$ in every round.
- After n rounds the dishonest prover can fool V with probability 2^{-n} .
- Can this protocol be executed in parallel and still remain ZK ?
- **Partial Answer:** We cannot use the simulator from the original protocol, because this simulator has probability of 2^{-n} to succeed. Thus, **we get exponential expected running time**.
- In the case of Parallel Fiat-Shamir: Parallel Fiat-Shamir is not ZK.

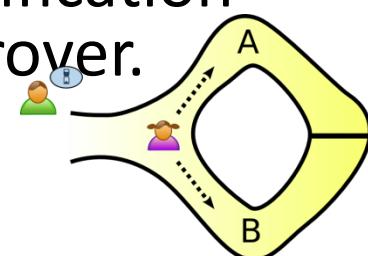


An Active Attack

- The only attack that some party can do is to actively use the protocol with both the prover and the verifier, to convince the verifier he is the prover, asking the prover to do the real work:

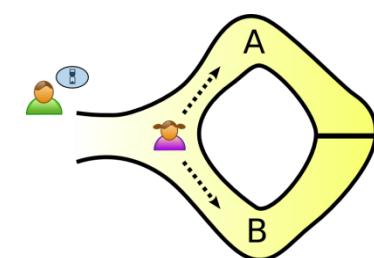


- In this attack, the attacker sends all the verifier's questions to the real prover, and all the answers of the prover are sent to the verifier. When the identification ends, the attacker can act as if he is the real prover.



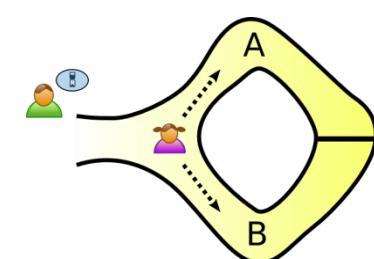
ZK Proofs of Knowledge

- The Fiat-Shamir protocol convinces the verifier that the prover knows the square root of I , without revealing any information on S . However, the verifier gets one bit of information: he learns that I is a quadratic residue.



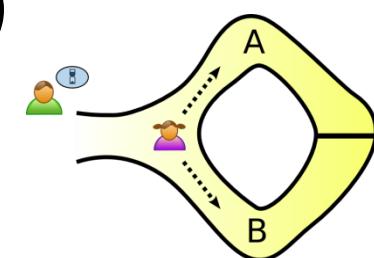
ZK Proofs of Knowledge (cont.)

- The following scheme does not even reveal whether l is a quadratic residue or not — it reveals only that the prover knows whether it is a quadratic residue or not:
- The moduli $n = pq$ is chosen such that both p and q are of the form $4m + 3$ (i.e., n is a Blum integer). Such moduli have the property that -1 is a quadratic non-residue whose Jacobi symbol modulo n is $+1$ (since -1 is a quadratic non-residue modulo p nor q). Thus, it is difficult to distinguish which of two numbers: a quadratic residue and its negation is the quadratic residue.



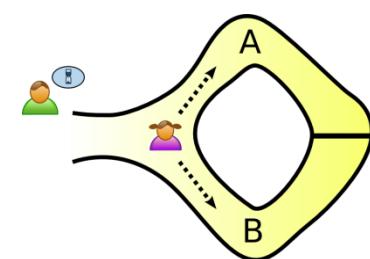
ZK Proofs of Knowledge (cont.)

- **The Secret Key:** The prover chooses k random values S_1, S_2, \dots, S_k , where $1 < S_i < n$, and keeps them secret.
- **The Public Key:** The prover computes $I_i = \pm \frac{1}{S_i^2} \bmod n$, where the **sign** is chosen randomly and independently, and publishes I_1, I_2, \dots, I_k and n as the public key.
- In this protocol, k secrets are proved in parallel, resulting with a smaller probability of cheating in each iteration. (However, k should be kept constant to keep it ZK, due to the details of the definition of ZK)



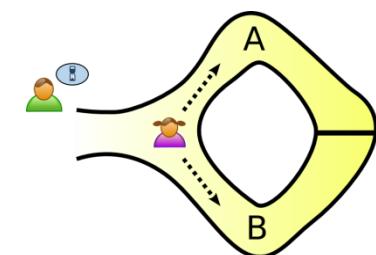
ZK Proofs of Knowledge (cont.)

- The Identification Protocol:
 1. The prover chooses a random value $1 < R < n$, and computes $X = \pm R^2 \bmod n$.
 2. The prover sends X to the verifier.
 3. The verifier sends a random boolean vector E_1, E_2, \dots, E_k .
 4. The prover sends $Y = R \cdot \prod_{E_j=1} S_j \bmod n$
 5. The verifier verifies that $X = \pm Y^2 \cdot \prod_{E_j=1} I_j \bmod n$



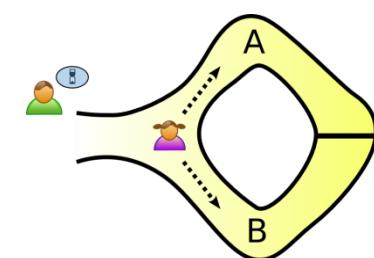
ZK Proofs of Knowledge (cont.)

- 6. If the verification fails, the verifier concludes that the prover does not know X , and thus he is not the claimed party.
- 7. 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 this protocol, the cheating probability is 2^{-k} for each iteration, and thus after t iterations the cheating probability is 2^{-kt} .



ZK Proofs of Knowledge (cont.)

- In this protocol, the values I_i , X and Y can be any numbers with Jacobi symbol +1.
- This is a zero knowledge protocol, since if the prover can answer two distinct questions, for two distinct values of the boolean vector E_1, E_2, \dots, E_k , he can compute the square root of a product of a subset of the I 's.



Question

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



Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 9: Private matching protocols in mobile social networks

Lecturer: Rongxing LU

Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

What is Mobile Social Networking

- Social Network Services:
 - Software for building online social networks for communities (virtual) of people who share interests and activities or who are interested in exploring the interests and activities of others
- Mobile Social Network Services
 - Use of social network services on mobile devices
 - Provide the proximity services



What's difference between Mobile social networking and the web-based social networking

- **Mobile social networking** is social networking where individuals with similar interests converse and connect with one another through their mobile phone and/or tablet. Much like web-based social networking, mobile social networking occurs in virtual communities.



What will we discuss in this topic?

- Privacy issues in mobile social networking
 - Secret handshake
 - Privacy-Preserving Friend Match
 - Other privacy enhancing techniques can be applied in mobile social networking
 - Private Equality Test
 - Private Set Inclusion
 - Private Set Intersection



What is the secret handshake?

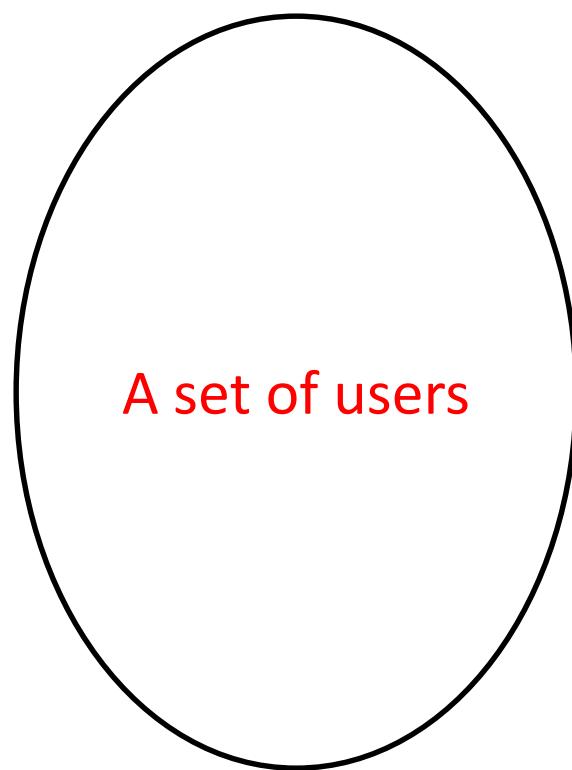
- Consider a CIA agent who wants to authenticate herself to a server, but does not want to reveal her CIA credentials unless the server is a genuine CIA outlet. Consider also that the CIA server does not want to reveal its CIA credentials to anyone but CIA agents – not even to other CIA servers.



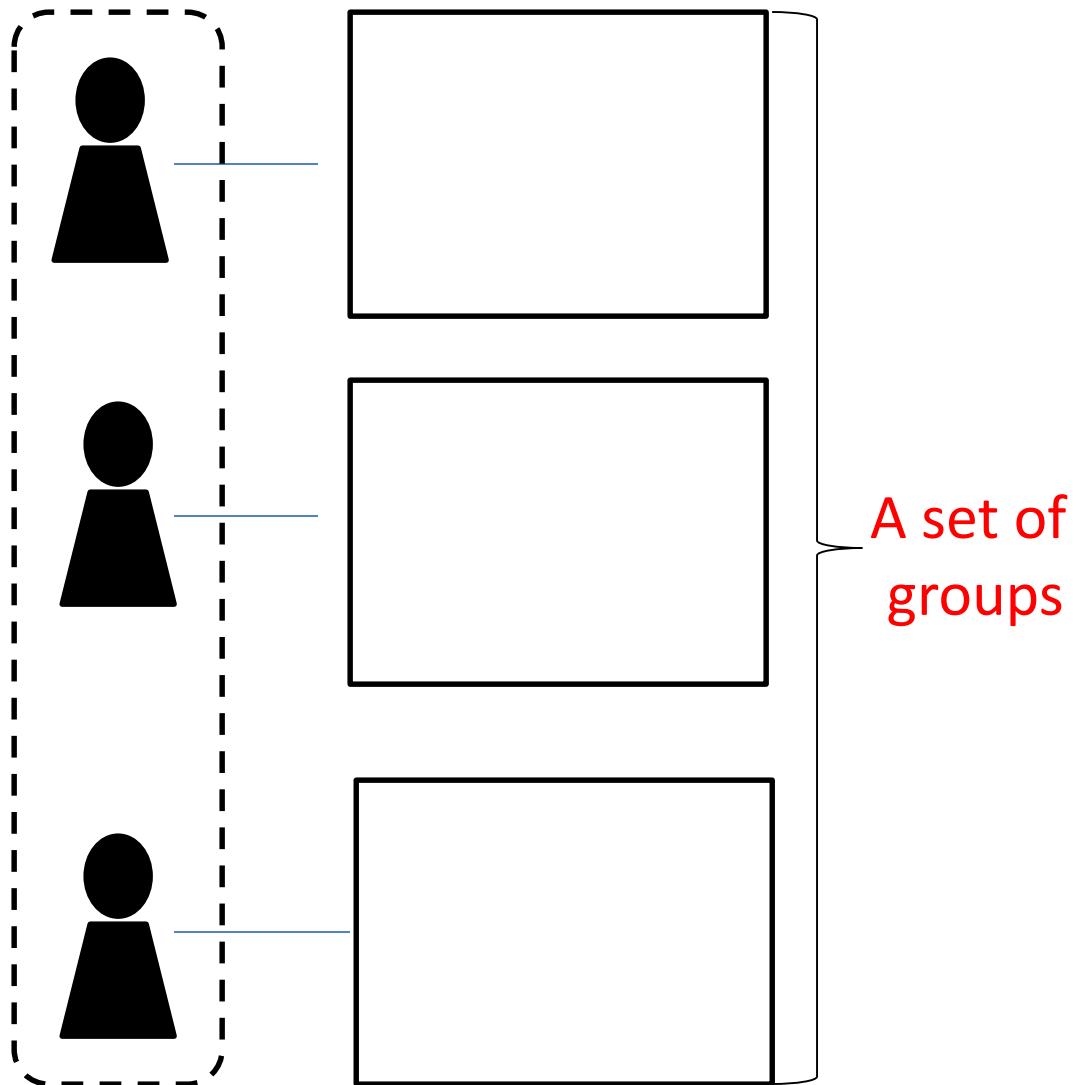
Bilinear Group

- Prime p
- Groups G, GT of order p
- g generates G
- Bilinear map $e: G \times G \rightarrow GT$
 - $e(g, g)$ generates G_T
 - $e(ga, gb) = e(g, g)^{ab}$
- Efficiently computable group operations,
group membership, bilinear map, etc.

Bilinear Pairing based Secret Handshake?



A set of
Administrators



Bilinear Pairing based Secret Handshake? (2)

- For one group **Go**, the administrator first chooses a random $s \in Z_p$ as the group **Go**'s secret
- When a user **U** enrolls the group **Go**, the administrator first generates a list of random “pseudonyms” $id_{U1}, \dots, id_{Ut} \in \{0,1\}^*$ for **U**, where t is chosen to be larger than the number of handshakes **U** will execute before receiving new user secrets
- Since only the administrator and the user itself know the identity **U**, only the administrator and the user can link any id_{Ui} back to **U**.
- The administrator then calculates a corresponding list of secret points $priv_{U1}, \dots, priv_{Ut}$ as $priv_{Ui} = H1(id_{Ui})^s$

Bilinear Pairing based Secret Handshake? (3)

- Let A and B be two users who wish to conduct a secret handshake. A pulls from his user secret an unused pseudonym $id_A \in \{id_{Ai}, \dots, id_{At}\}$, together with the corresponding secret point $priv_A$. B likewise pulls id_B and $priv_B$.
- First, A sends his pseudonym, along with a random nonce n_A , to B.
- B replies with her pseudonym, a nonce n_B of her choosing, and a value V_0 .
- A verifies that $V_0 = H_2(e(priv_A, H1(id_B)) \parallel id_A \parallel id_B \parallel n_A \parallel n_B \parallel 0)$ and replies with V_1

Bilinear Pairing based Secret Handshake? (4)

- B verifies that $V_1 = H_2(e(priv_B, H1(id_A)) \parallel id_A \parallel id_B \parallel n_A \parallel n_B \parallel 1)$
- If both verifications succeed, then A and B can create a shared secret S for future communication.
 - A calculates the shared secret like this: $S = H_2(e(priv_A, H1(id_B)) \parallel id_A \parallel id_B \parallel n_A \parallel n_B \parallel 2)$
 - B calculates the same shared secret S as: $S = H_2(e(priv_B, H1(id_A)) \parallel id_A \parallel id_B \parallel n_A \parallel n_B \parallel 2)$
- Note that, if A and B are affiliated in the same group Go, both verification will succeed. Otherwise, both will fail.

Bilinear Pairing based Secret Handshake? (5)

- **TraceUser.** Given a transcript of a handshake between user **A** and **B**, the administrator can easily recover the pseudonyms \mathbf{id}_A and \mathbf{id}_B and look up which users these pseudonyms had been issued to.
- **RemoveUser.** To remove a user **U** from the group **Go**, the administrator looks up the user secret $(\mathbf{id}_{U_1}, \dots, \mathbf{id}_{U_t}, \mathbf{priv}_{U_1}, \dots, \mathbf{priv}_{U_t})$ it has issued to **U** and alerts every other user to abort any handshake should they find themselves performing the handshake with a user using any pseudonym $\mathbf{id}_U \in \{\mathbf{id}_{U_i}, \dots, \mathbf{id}_{U_t}\}$.

Privacy-Preserving Friend Match

- Two users in mobile social network, when they meet on the road, they want to launch some talk if they can become friends:
 - “become friends” means they have at t items of common interests.
 - For example, if they both like football, basketball, and badminton, i.e., $\geq t = 3$, they can make friends, and then they know each other’s interests by talking.
 - Otherwise, if $< t = 3$, they cannot make friends and know nothing about other’s interests.

Privacy-Preserving Friend Match

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
----	----	----	----	----	----	----	----	----	-----

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
1	0	1	1	0	0	1	1	0	1

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0	0	1	1	0	1	1	0	1	1



If set $t = 5$, Alice and Bob can not make friends.

If set $t = 4$, Alice and Bob can make friend, because they have the common interests on (A3, A4, A7, A10)

Question

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



Paillier –based Privacy-Preserving Friend Match



ALICE

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
----	----	----	----	----	----	----	----	----	-----



BOB

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
1	0	1	1	0	0	1	1	0	1

a

$$t = 4$$

b

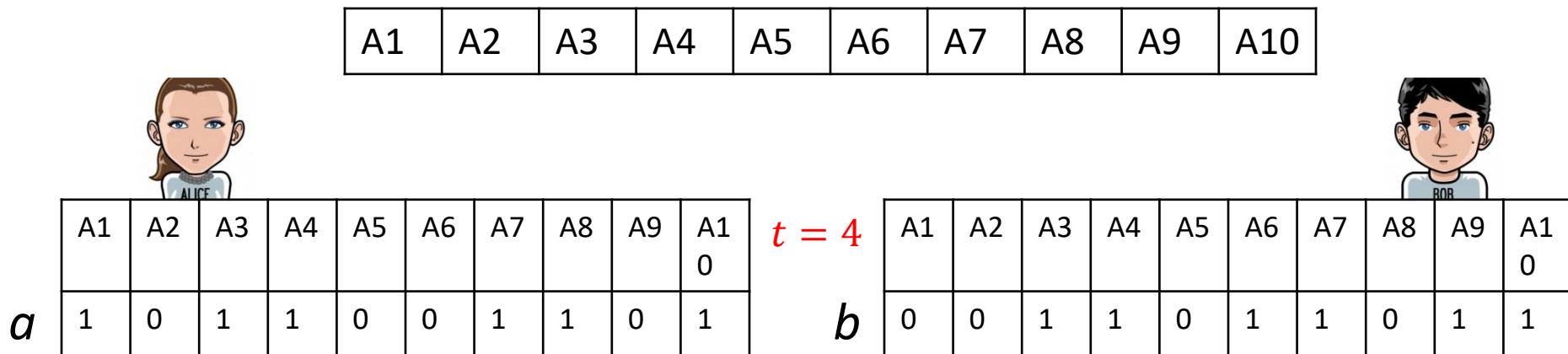
A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0	0	1	1	0	1	1	0	1	1

Paillier $PK = (n, g)$ $\xrightarrow{\hspace{10em}}$ $PK = (n, g)$
 $SK = (\mu, \lambda)$

$E(1), E(0), E(1), E(1), E(0), \xrightarrow{\hspace{10em}} E(1)^0 + E(0)^0 + E(1)^1 + E(1)^1 + E(0)^0 + E(0)^1 + E(1)^1 + E(1)^0 + E(0)^1 + E(1)^1$
 $E(0), E(1), E(1), E(0), E(1)$
 \longleftrightarrow
 $E(1 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 + 1 \cdot 1 + 0 \cdot 0 + 0 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1)$

Recover 4 from $E(4)$, as $4 \geq t$, they continue to make friends.

Paillier –based Privacy-Preserving Friend Match



Paillier $PK = (n, g)$ $\xrightarrow{\hspace{10cm}}$ $PK = (n, g)$
 $SK = (\mu, \lambda)$

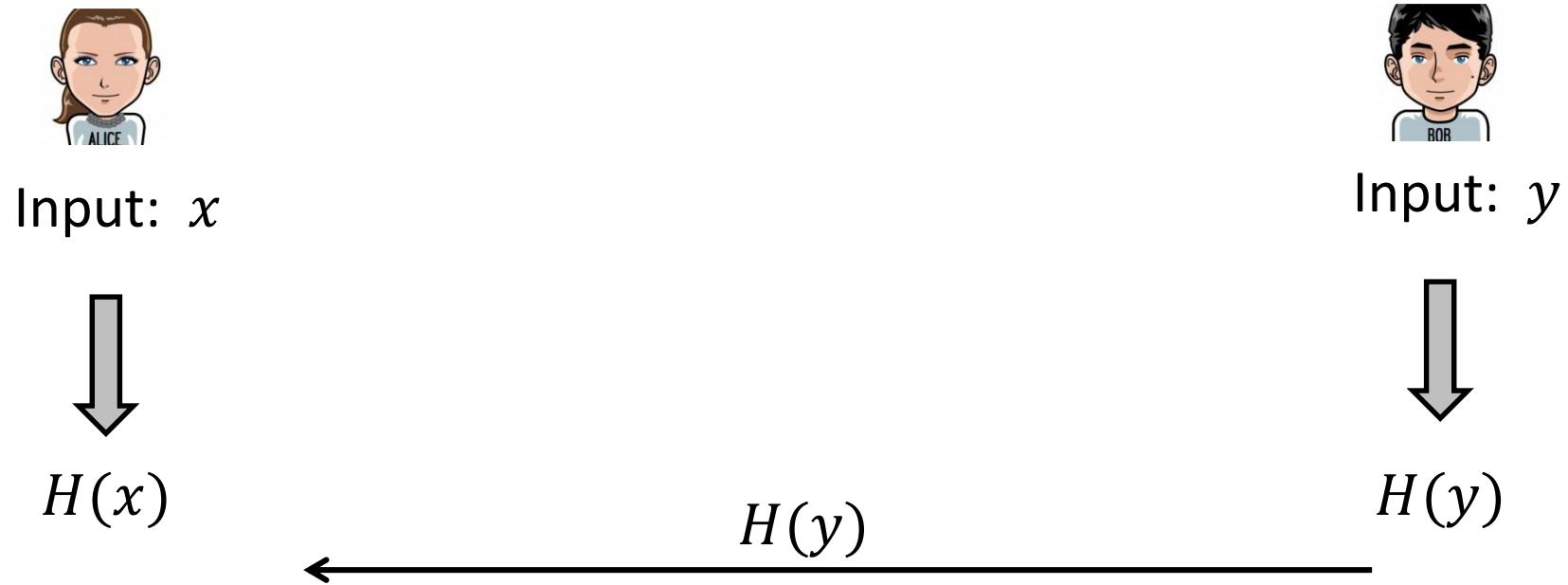
For $i = 1, 2, \dots, N$ $\xrightarrow{\hspace{10cm}}$ all c_i for $i = 1, 2, \dots, N$
 $c_i = E(a_i)$

$$D = \prod_{i=1}^N c_i^{b_i}$$

$\xleftarrow{\hspace{10cm}}$

Decrypt D , if the result $\geq t$, make friends; and stop otherwise.

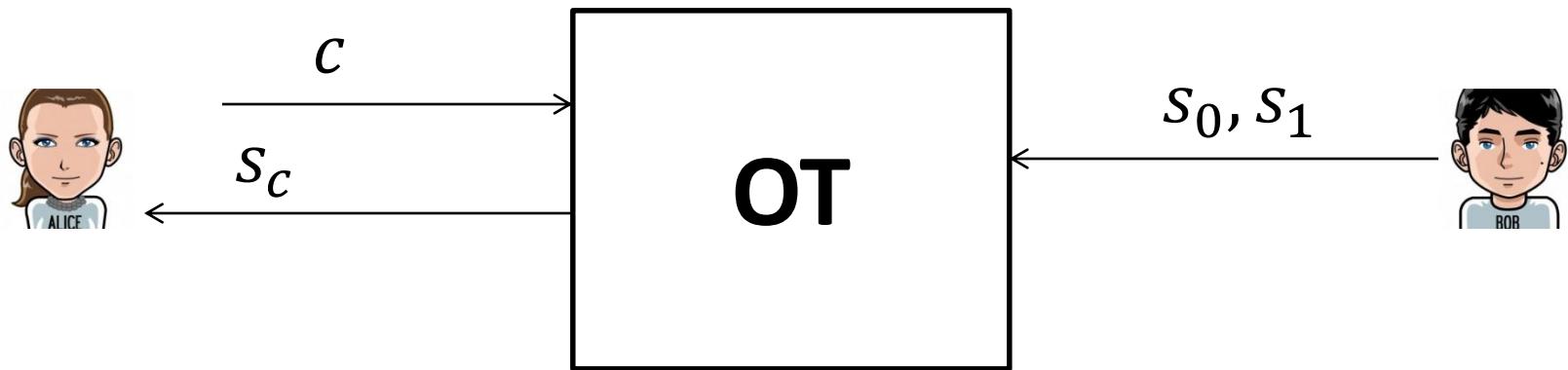
Private Equality Test



Check $H(x) =? = H(y)$

- **Pro:** fast, little communication
- **Con:** can leak privacy of Bob's inputs, **why?**

Review of Oblivious Transfer (OT)



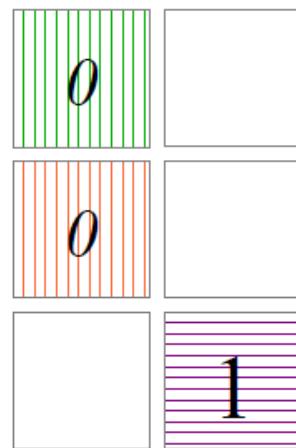
Input: Bob holds two strings (s_0, s_1) , Alice holds a choice bit c

Output: Alice receives s_c but learns nothing about s_{1-c} , Bob learns nothing about c

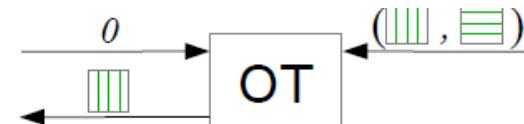
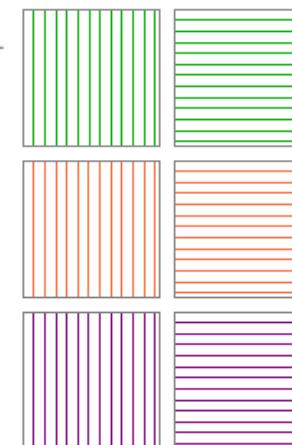
OT-based Private Equality Test



Input: x $x = 001$



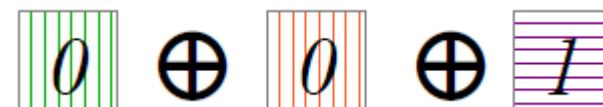
$y = 011$ Input: y



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

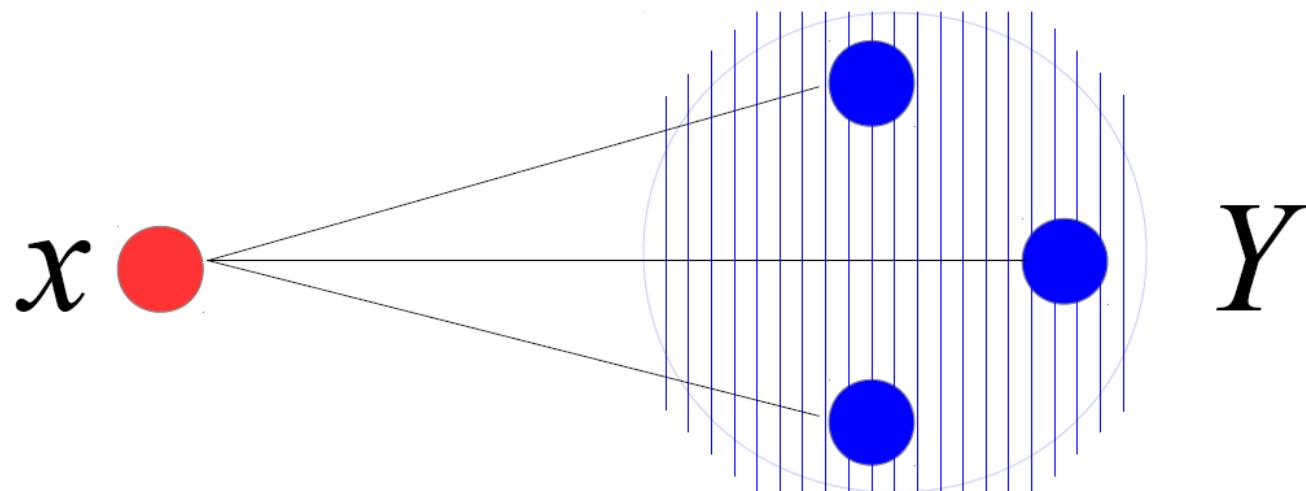


Alice computes $0 \oplus 0 \oplus I$ and compares



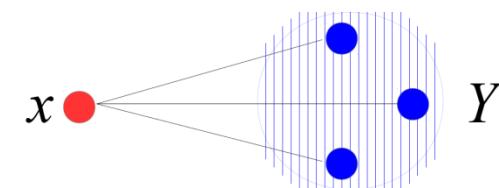
OT-based Private Set Inclusion

- **Input:** Alice has x , Bob has $Y = \{y_1, \dots, y_n\}$.
- **Output:** $x \in? Y$



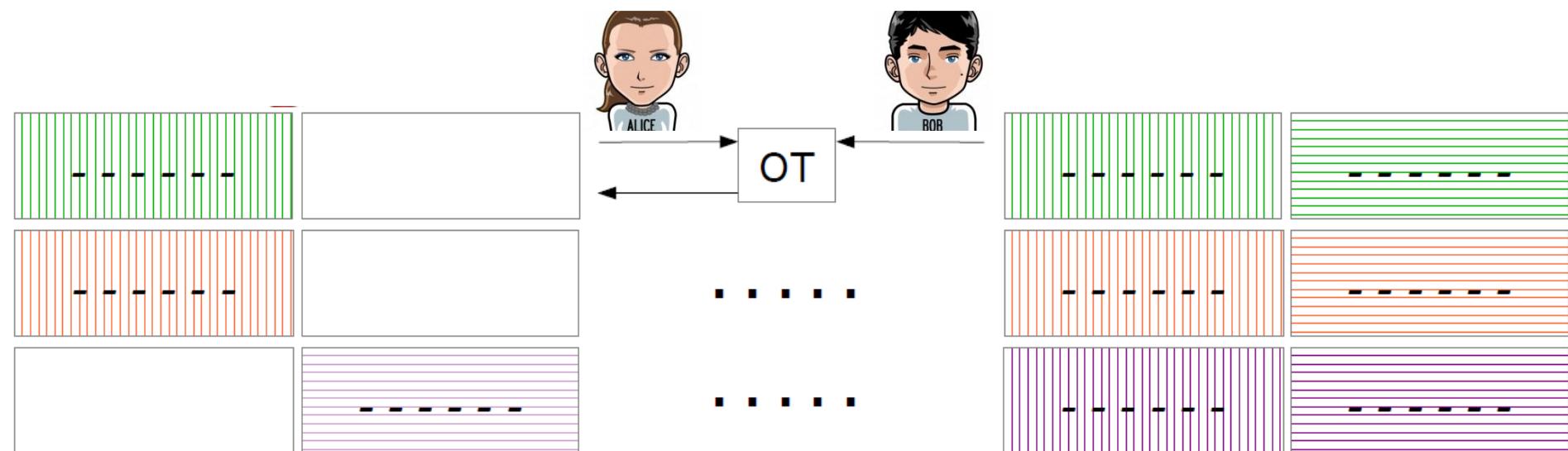
OT-based Private Set Inclusion (2)

- **Input:** Alice has x , Bob has $Y = \{y_1, \dots, y_n\}$.
- **Output:** $x \in? Y$



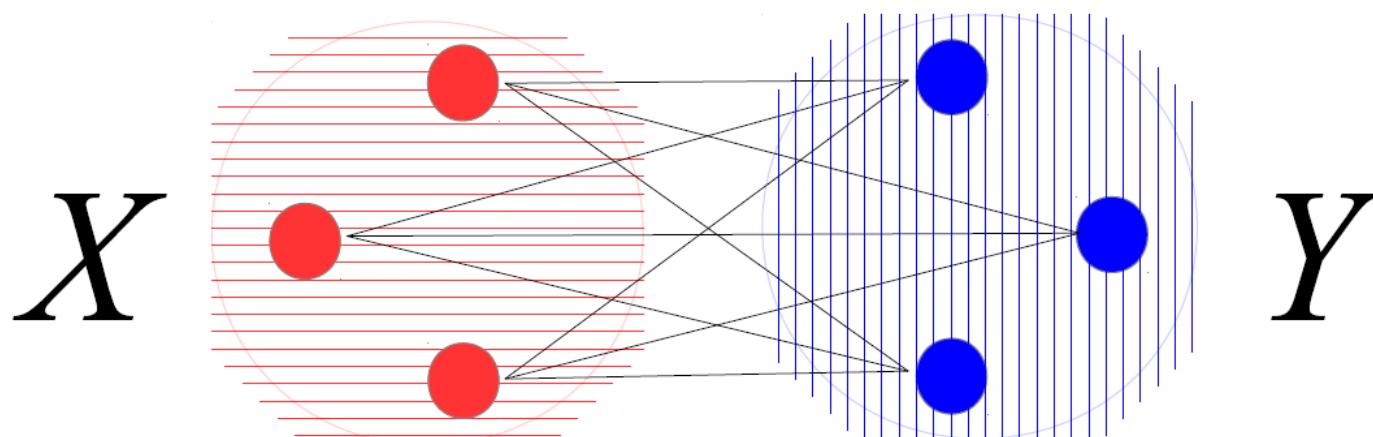
Run n Private Equality Tests in parallel

- Alice's OT choices for all (y_1, \dots, y_n) are the same
- Send $n\lambda$ bits from Bob to Alice



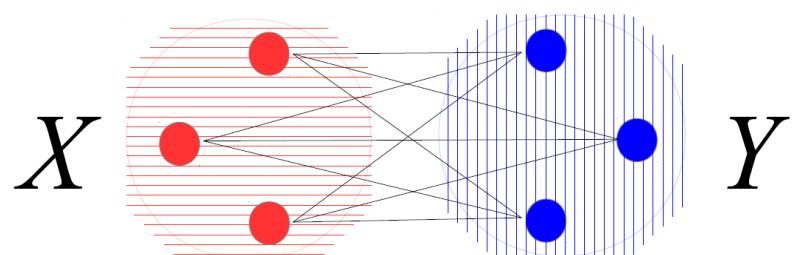
OT-based Private Set Intersection

- **Input:** Alice has $X = \{x_1, \dots, x_n\}$, Bob has $Y = \{y_1, \dots, y_n\}$.
- **Output:** $X \cap Y$



OT-based Private Set Intersection (2)

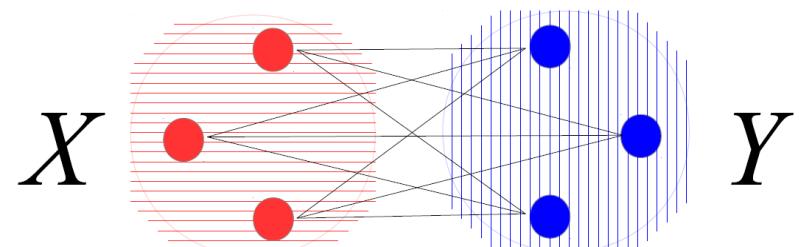
- **Input:** Alice has $X = \{x_1, \dots, x_n\}$, Bob has $Y = \{y_1, \dots, y_n\}$.
- **Output:** $X \cap Y$



- Run n Private Set Inclusions in parallel
 - Requires n^2 comparisons, hence not an option
 - Is it possible to find an efficient solution?
 - Rely on hashing

OT-based Private Set Intersection with Hashing

- **Input:** Alice has $X = \{x_1, \dots, x_n\}$, Bob has $Y = \{y_1, \dots, y_n\}$.



- **Output:** $X \cap Y$

- Hash elements to bins to reduce comparisons

- **Example:**

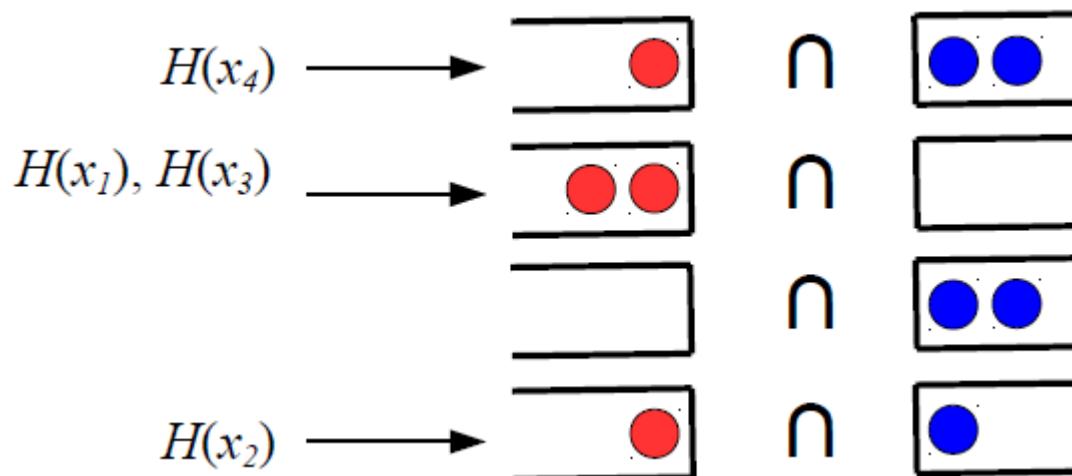
- Alice holds $X = \{x_1, \dots, x_4\}$, Bob holds $Y = \{y_1, \dots, y_4\}$

- Reduces comparisons from n^2 to $O(n \log n)$

- Why?

OT-based Private Set Intersection with Hashing (2)

- Hash elements to bins to reduce comparisons
- **Example:**
 - Alice holds $X = \{x_1, \dots, x_4\}$, Bob holds $Y = \{y_1, \dots, y_4\}$
 - Reduces comparisons from n^2 to $O(n \log n)$
 - Why?



Why?

- When we hash n elements into m bins (B_1, B_2, \dots, B_m), averagely, each bin has $\frac{n}{m}$ items.
- Alice just needs to compare the elements in the same B_i .
- Then, for each bin, the comparisons are $\left(\frac{n}{m}\right)^2$
- As there are m bins, the total comparisons are $\frac{n^2}{m}$
- If we set $\frac{n}{m} = \log n$, i.e., $m = \frac{n}{\log n}$, we have the total comparisons are $\frac{n^2}{m} = n \cdot \log n = O(n \cdot \log n)$

Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 10: Secure data sharing in cloud computing

Lecturer: Rongxing LU

Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

What is Cloud Computing?

- **Cloud Computing** is a general term used to describe a new class of network based computing that takes place over the Internet,
 - basically a step up 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
- These platforms hide the complexity and details of the underlying infrastructure from users and applications by providing very simple graphical interface or API (Applications Programming Interface).



Cloud Computing

- In addition, the platform provides on demand services, that are always on, anywhere, anytime and any place.
- Pay for use and as needed, elastic
 - scale up and down in capacity and functionalities
- The hardware and software services are available to
 - general public, enterprises, corporations and businesses markets

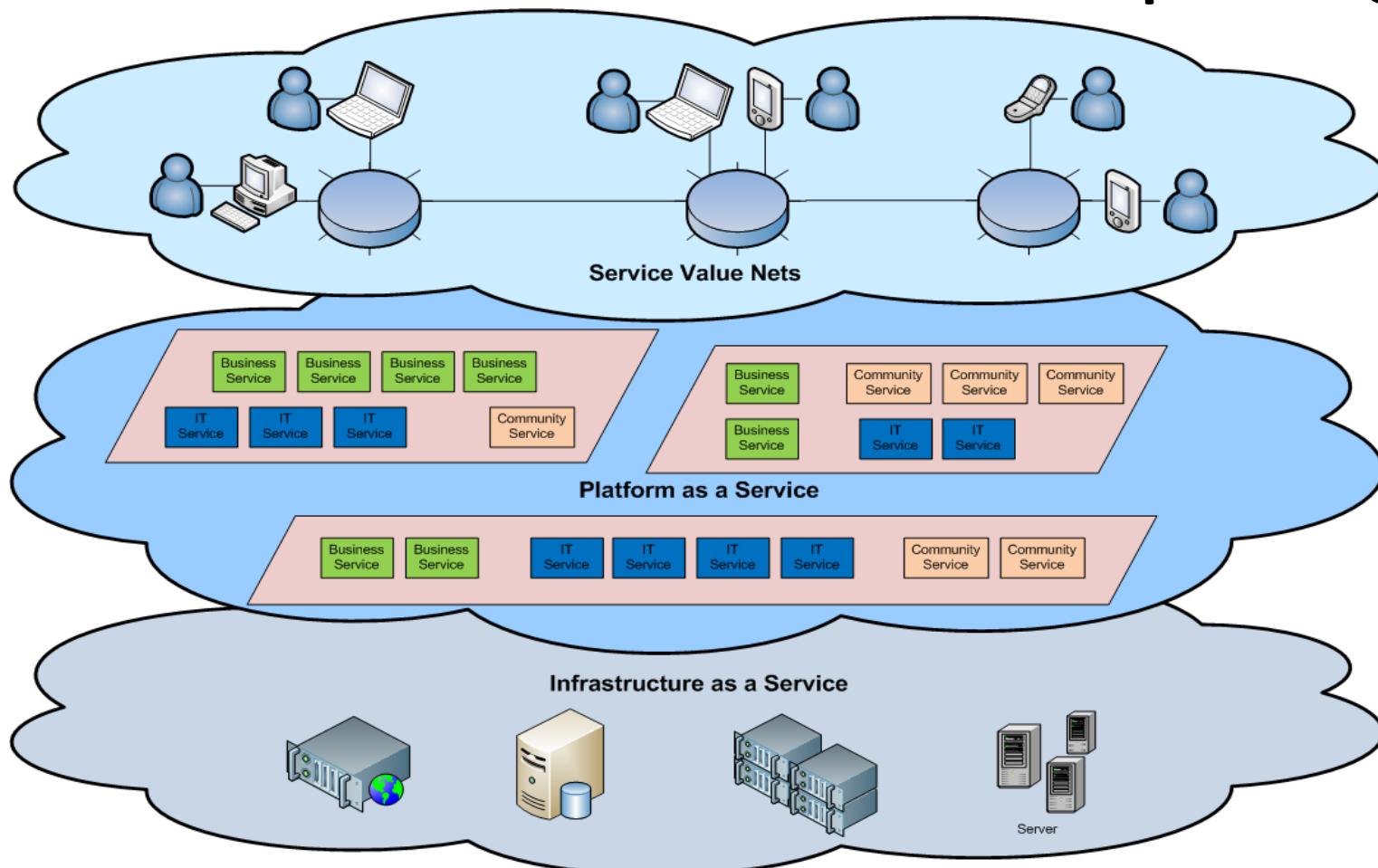


Cloud Computing is an Umbrella

- Cloud computing is an umbrella term used to refer to Internet based development and services
- A number of characteristics define cloud data, applications services and infrastructure:
 - **Remotely hosted:** Services or data are hosted on remote infrastructure.
 - **Ubiquitous:** Services or data are available from anywhere.
 - **Commodified:** The result is a utility computing model similar to traditional that of traditional utilities, like gas and electricity - you pay for what you would want!



Architecture of Cloud Computing



- Shared pool of configurable computing resources
- On-demand network access
- Provisioned by the Service Provider

Characteristics of Cloud Computing

Common Characteristics:

Massive Scale

Resilient Computing

Homogeneity

Geographic Distribution

Virtualization

Service Orientation

Low Cost Software

Advanced Security

Essential Characteristics of Cloud Computing

- ***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 service's 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.*** 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).

Cloud Service Models

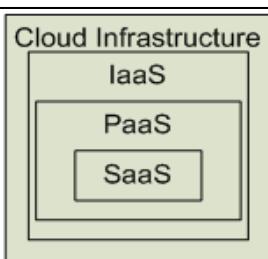
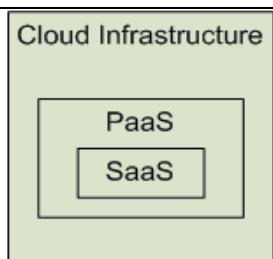
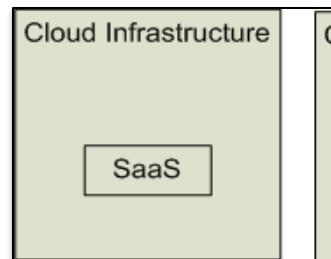
Software as a Service (SaaS)

Platform as a Service (PaaS)

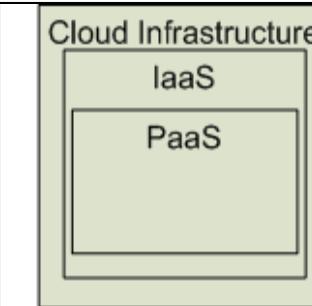
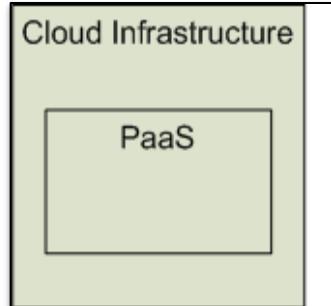
Infrastructure as a Service (IaaS)

SalesForce CRM

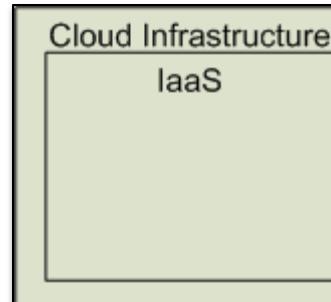
LotusLive



Software as a Service (SaaS)
Providers
Applications



Platform as a Service (PaaS)
Deploy customer created Applications



Infrastructure as a Service (IaaS)
Rent Processing, storage, N/W capacity & computing resources



Different Cloud Computing Layers

Application Service (SaaS)	MS Live/ExchangeLabs, IBM, Google Apps; Salesforce.com Quicken Online, Zoho, Cisco
Application Platform	Google App Engine, Mosso, Force.com, Engine Yard, Facebook, Heroku, AWS
Server Platform	3Tera, EC2, SliceHost, GoGrid, RightScale, Linode
Storage Platform	Amazon S3, Dell, Apple, ...

Cloud Computing Service Layers

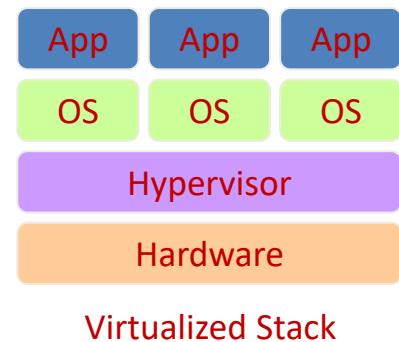
Services	Description
Services	Services – Complete business services such as PayPal, OpenID, OAuth, Google Maps, Alexa
Application	Application – Cloud based software that eliminates the need for local installation such as Google Apps, Microsoft Online
Development	Development – Software development platforms used to build custom cloud based applications (PAAS & SAAS) such as SalesForce
Platform	Platform – Cloud based platforms, typically provided using virtualization, such as Amazon ECC, Sun Grid
Storage	Storage – Data storage or cloud based NAS such as CTERA, iDisk, CloudNAS
Hosting	Hosting – Physical data centers such as those run by IBM, HP, NaviSite, etc.

Application Focused {

Infrastructure Focused {

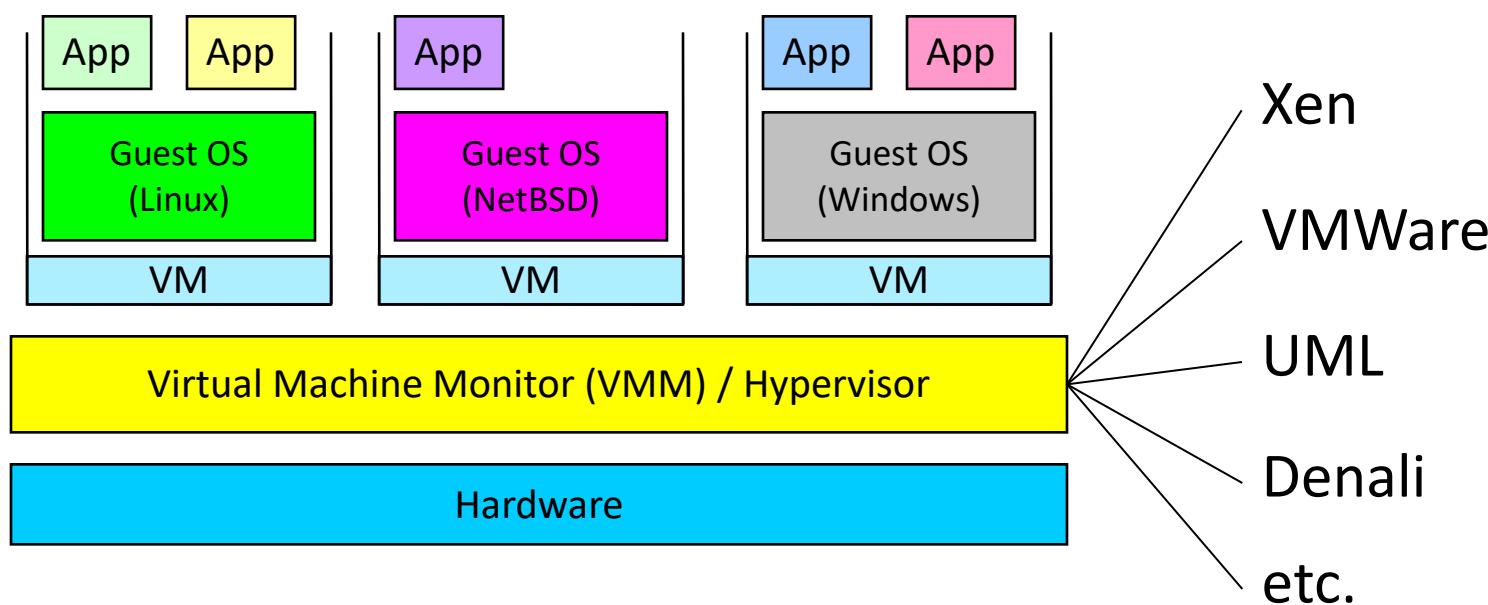
Virtualization

- Virtual workspaces:
 - An abstraction of an execution environment that can be made dynamically available to authorized clients by using well-defined protocols,
 - Resource quota (e.g. CPU, memory share),
 - Software configuration (e.g. O/S, provided services).
- Implement on Virtual Machines (VMs):
 - Abstraction of a physical host machine,
 - Hypervisor intercepts and emulates instructions from VMs, and allows management of VMs,
 - VMWare, Xen, etc.
- Provide infrastructure API:
 - Plug-ins to hardware/support structures



Virtual Machines

- VM technology allows multiple virtual machines to run on a single physical machine.



Performance: Para-virtualization (e.g. Xen) is very close to raw physical performance!

Virtualization in General

- Advantages of virtual machines:
 - Run operating systems where the physical hardware is unavailable,
 - Easier to create new machines, backup machines, etc.,
 - Software testing using “clean” installs of operating systems and software,
 - Emulate more machines than are physically available,
 - Timeshare lightly loaded systems on one host,
 - Debug problems (suspend and resume the problem machine),
 - Easy migration of virtual machines (shutdown needed or not).
 - Run legacy systems!

Some Commercial Cloud Offerings



Amazon Elastic Compute Cloud (Amazon EC2) - Beta



3tera™ info@3tera.com (949) 306-0050

CAREERS | SU APPLOGIC | UTILITY COMPUTING | TECHNOLOGY | PARTNERS | GRID UNIVERSITY | COMPANY |

Cloud Computing

Overview

Cloudware - Cloud Computing Without Compromise



MOSSO®
the hosting cloud



VERIO

An NTT Communications Company

Cloud Storage

- Several large Web companies are now exploiting the fact that they have data storage capacity that can be hired out to others.
 - allows data stored remotely to be temporarily cached on desktop computers, mobile phones or other Internet-linked devices.
- Amazon's Elastic Compute Cloud (EC2) and Simple Storage Solution (S3) are well known examples



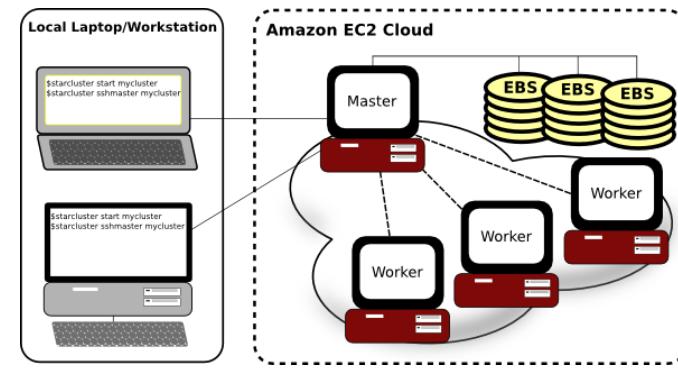
Amazon Simple Storage Service (S3)

- Unlimited Storage.
- Pay for what you use:
 - \$0.20 per GByte of data transferred,
 - \$0.15 per GByte-Month for storage used,
 - Second Life Update:
 - 1TBytes, 40,000 downloads in 24 hours - \$200,



Utility Computing – EC2

- Amazon Elastic Compute Cloud (EC2):
 - Elastic, marshal 1 to 100+ PCs via WS,
 - Machine Specs...,
 - Fairly cheap!
- Powered by Xen – a Virtual Machine:
 - Different from Vmware and VPC as uses “para-virtualization” where the guest OS is modified to use special hyper-calls:
 - Hardware contributions by Intel (VT-x/Vanderpool) and AMD (AMD-V).
 - Supports “Live Migration” of a virtual machine between hosts.
- Linux, Windows, OpenSolaris
- Management Console/AP



Opportunities and Challenges

- The use of the cloud provides a number of opportunities:
 - It enables services to be used without any understanding of their infrastructure.
 - Cloud computing works using economies of scale:
 - It potentially lowers the outlay expense for start up companies, as they would no longer need to buy their own software or servers.
 - Cost would be by on-demand pricing.
 - Vendors and Service providers claim costs by establishing an ongoing revenue stream.
 - Data and services are stored remotely but accessible from “anywhere”.

Opportunities and Challenges

- In parallel there has been backlash against cloud computing:
 - Use of cloud computing means dependence on others and that could possibly limit flexibility and innovation:
 - The others are likely become the bigger Internet companies like Google and IBM, who may monopolise the market.
 - Some argue that this use of supercomputers is a return to the time of mainframe computing that the PC was a reaction against.
 - Security could prove to be a big issue:
 - It is still unclear how safe out-sourced data is and when using these services ownership of data is not always clear.
 - There are also issues relating to policy and access:
 - If your data is stored abroad whose policy do you adhere to?
 - What happens if the remote server goes down?
 - How will you then access files?
 - There have been cases of users being locked out of accounts and losing access to data.

Advantages & Disadvantages of Cloud Computing

- **Advantages**

- Improved performance:
- Reduced software costs:
- Instant software updates:
- Improved document format compatibility.
- Unlimited storage capacity:
- Increased data reliability:
- Universal document access:
- Latest version availability:
- Easier group collaboration:
- Device independence.



- **Disadvantages**

- Requires a constant Internet connection:
- Does not work well with low-speed connections:
- Stored data might not be secure:
- Stored data can be lost:

Security and Privacy Challenges in Cloud Computing

- Most security problems stem from:
 - Loss of control
 - Lack of trust (mechanisms)
 - Multi-tenancy



Loss of Control 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 and privacy
 - Resource availability
 - Monitoring and repairing of services/resources



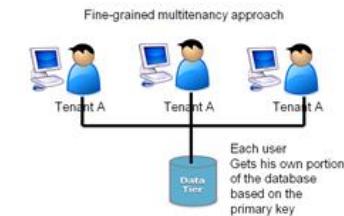
Lack of Trust in the Cloud

- Trusting a third party requires taking risks
- Defining trust and risk
 - Opposite sides of the same coin (J. Camp)
 - People only trust when it pays (Economist's view)
 - Need for trust arises only in risky situations
- Defunct third party management schemes
 - Hard to balance trust and risk
 - e.g. Key Escrow (Clipper chip)
 - Is the cloud headed toward the same path?



Multi-tenancy Issues in the Cloud

- Conflict between tenants' opposing goals
 - Tenants share a pool of resources and have opposing goals
- How does multi-tenancy deal with conflict of interest?
 - Can tenants get along together and 'play nicely' ?
 - If they can't, can we isolate them?
- How to provide separation between tenants?
- Cloud Computing brings new threats
 - Multiple independent users share the same physical infrastructure
 - Thus an attacker can legitimately be in the same physical machine as the target



Security Requirements

- Confidentiality
 - Fear of loss of control over data
 - Will the sensitive data stored on a cloud remain confidential?
 - Will cloud compromises leak confidential client data
 - Will the cloud provider itself be honest and won't peek into the data?
- Integrity
 - How do I know that the cloud provider is doing the computations correctly?
 - How do I ensure that the cloud provider really stored my data without tampering with it?



Security Requirements

- Availability

- Will critical systems go down at the client, if the provider is attacked in a Denial of Service attack?
- What happens if cloud provider goes out of business?
- Would cloud scale well-enough?
- Often-voiced concern
 - Although cloud providers argue their downtime compares well with cloud user's own data centers



Security Requirements

- Privacy issues raised via massive data mining
 - Cloud now stores data from a lot of clients, and can run data mining algorithms to get large amounts of information on clients
- Increased attack surface
 - Entity outside the organization now stores and computes data, and so
 - Attackers can now target the communication link between cloud provider and client
 - Cloud provider employees can be phished



Security Requirements

- Auditability and forensics (out of control of data)
 - Difficult to audit data held outside organization in a cloud
 - Forensics also made difficult since now clients don't maintain data locally
- Legal quagmire and transitive trust issues
 - Who is responsible for complying with regulations?
 - e.g., SOX, HIPAA, GLBA ?
 - If cloud provider subcontracts to third party clouds, will the data still be secure?



Cloud Computing is a security nightmare

- Security is one of the most difficult task to implement in cloud computing.
 - Different forms of attacks in the application side and in the hardware components
- Attacks with catastrophic effects only needs one security flaw

(<http://www.exforsys.com/tutorials/cloud-computing/cloud-computing-security.html>)



Threat Model

- A threat model helps in analyzing a security problem, design mitigation strategies, and evaluate solutions
- Steps:
 - Identify attackers, assets, threats and other components
 - Rank the threats
 - Choose mitigation strategies
 - Build solutions based on the strategies
- Basic components
 - Attacker modeling
 - Choose what attacker to consider
 - insider vs. outsider?
 - single vs. collaborator?
 - Attacker motivation and capabilities
 - Attacker goals
 - Vulnerabilities / threats

Threat Modeling
REVIEW WITH MANAGEMENT



What is the issue?

- The core issue here is the levels of trust
 - Many cloud computing providers trust their customers
 - Each customer is physically commingling its data with data from anybody else using the cloud while logically and virtually you have your own space
 - The way that the cloud provider implements security is typically focused on the fact that those outside of their cloud are evil, and those inside are good.
- But what if those inside are also evil?



Attacker Capability: Outside attacker

- What?
 - Listen to network traffic (passive)
 - Insert malicious traffic (active)
 - Probe cloud structure (active)
 - Launch DoS
- Goal?
 - Intrusion
 - Network analysis
 - Man in the middle
 - Cartography



Attacker Capability: Malicious Insiders

- At client
 - Learn passwords/authentication information
 - Gain control of the VMs
- At cloud provider
 - Log client communication
 - Can read unencrypted data
 - Can possibly peek into VMs, or make copies of VMs
 - Can monitor network communication, application patterns
 - Why?
 - Gain information about client data
 - Gain information on client behavior
 - Sell the information or use itself

Challenges for the attacker

- How to find out where the target is located?
- How to be co-located with the target in the same (physical) machine?
- How to gather information about the target?



Infrastructure Security in Cloud Computing

- Infrastructure Security
 - Network Level
 - Host Level
 - Application Level



Network Level

- Ensuring confidentiality and integrity of your organization's data-in-transit to and from your public cloud provider
- Ensuring proper access control (authentication, authorization, and auditing) to whatever resources you are using at your public cloud provider
- Ensuring availability of the Internet-facing resources in a public cloud that are being used by your organization, or have been assigned to your organization by your public cloud providers

The Host Level

- SaaS/PaaS
 - Both the PaaS and SaaS platforms abstract and hide the host OS from end users
 - Host security responsibilities are transferred to the CSP (Cloud Service Provider)
 - You do not have to worry about protecting hosts
 - However, as a customer, you still own the risk of managing information hosted in the cloud services.

Local Host Security

- Are local host machines part of the cloud infrastructure?
 - Outside the security perimeter
 - While cloud consumers worry about the security on the cloud provider's site, they may easily forget to harden their own machines
- The lack of security of local devices can
 - Provide a way for malicious services on the cloud to attack local networks through these terminal devices
 - Compromise the cloud and its resources for other users
- With mobile devices, the threat may be even stronger
 - Users misplace or have the device stolen from them
 - Security mechanisms on handheld gadgets are often times insufficient compared to say, a desktop computer
 - Provides a potential attacker an easy avenue into a cloud system.
 - If a user relies mainly on a mobile device to access cloud data, the threat also increased as mobile devices malfunction or are lost
- Devices that access the cloud should have
 - Strong authentication mechanisms
 - Tamper-resistant mechanisms
 - Strong isolation between applications
 - Methods to trust the OS
 - Cryptographic functionality when traffic confidentiality is required



Data Security and Storage in Cloud Computing

- Several aspects of data security, including:
 - Data-in-transit
 - Confidentiality + integrity using secured protocol
 - Confidentiality with non-secured protocol and encryption
 - Data-at-rest
 - Generally, not encrypted , since data is commingled with other users' data
 - Encryption if it is not associated with applications?
 - But how about indexing and searching?
 - Then homomorphic encryption vs. predicate encryption?
 - Processing of data, including multitenancy
 - For any application to process data, not encrypted

Data Security and Storage in Cloud Computing

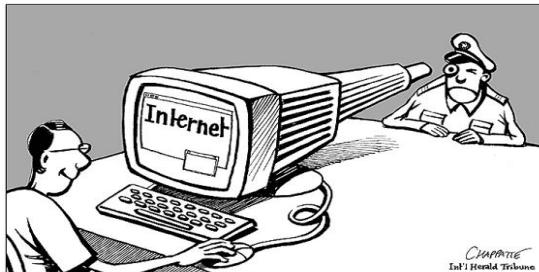
- Data remanence
 - Inadvertent disclosure of sensitive information is possible
- Data security mitigation?
 - Do not place any sensitive data in a public cloud
 - Encrypted data is placed into the cloud?
- Provider data and its security: storage
 - To the extent that quantities of data from many companies are centralized, this collection can become an attractive target for criminals
 - Moreover, the physical security of the data center and the trustworthiness of system administrators take on new importance.

Identity and Access Management (IAM)

- Organization's trust boundary will become dynamic and will move beyond the control and will extend into the service provider domain.
- Managing access for diverse user populations (employees, contractors, partners, etc.)
- Increased demand for authentication
 - personal, financial, medical data will now be hosted in the cloud
 - S/W applications hosted in the cloud requires access control
- Need for higher-assurance authentication
 - authentication in the cloud may mean authentication outside F/W
 - Limits of password authentication
- Need for authentication from mobile devices

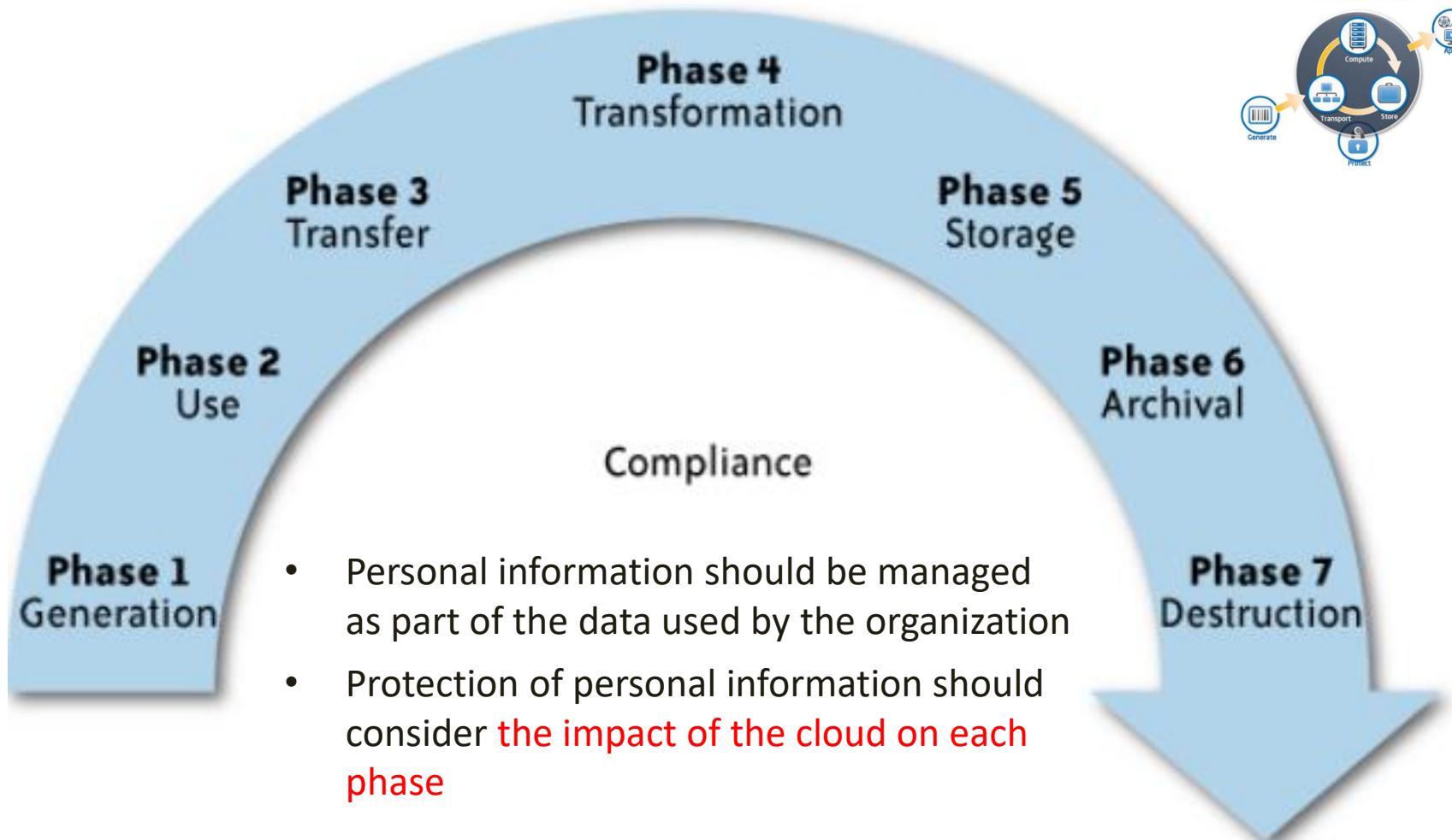
Privacy in Cloud Computing

- The concept of privacy varies widely among (and sometimes within) countries, cultures, and jurisdictions.
- It is shaped by public expectations and legal interpretations; as such, a concise definition is elusive if not impossible.
- Privacy rights or obligations are related to the collection, use, disclosure, storage, and destruction of personal data (or Personally Identifiable Information—PII).
- At the end of the day, privacy is about the accountability of organizations to data subjects, as well as the transparency to an organization's practice around personal information.



Privacy?

What is the data life cycle?



What Are the Key Privacy Concerns?

- Storage
- Retention
- Destruction
- Auditing, monitoring and risk management

- Privacy breaches
- Who is responsible for protecting privacy?



Storage

- Is it commingled with information from other organizations that use the same CSP?
- 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
- Whether the cloud provider itself has any right to see and access customer data?
- Some services today track user behaviour for a range of purposes, from sending targeted advertising to improving services



Retention



- How long is personal information (that is transferred to the cloud) retained?
- Which retention policy governs the data?
- Does the organization own the data, or the CSP?
- Who enforces the retention policy in the cloud, and how are exceptions to this policy (such as litigation holds) managed?

Destruction



- How does the cloud provider destroy PII at the end of the retention period?
- How do organizations ensure that their PII is destroyed by the CSP at the right point and is not available to other cloud users?
- Cloud storage providers usually replicate the data across multiple systems and sites—increased availability is one of the benefits they provide.
 - How do you know that the CSP didn't retain additional copies?
 - Did the CSP really destroy the data, or just make it inaccessible to the organization?
 - Is the CSP keeping the information longer than necessary so that it can mine the data for its own use?

Auditing, monitoring and risk management

- How can organizations monitor their CSP and provide assurance to relevant stakeholders that privacy requirements are met when their PII is in the cloud?
- Are they regularly audited?
- What happens in the event of an incident?
- 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



Privacy breaches

- How do you know that a breach has occurred?
- How do you ensure that the CSP notifies you when a breach occurs?
- Who is responsible for managing the breach notification process (and costs associated with the process)?
- If contracts include liability for breaches resulting from negligence of the CSP?
 - How is the contract enforced?
 - How is it determined who is at fault?



Who is responsible for protecting privacy?

- Data breaches have a cascading effect
- Full reliance on a third party to protect personal data?
- In-depth understanding of responsible data stewardship
- Organizations can transfer liability, but not accountability
- Risk assessment and mitigation throughout the data life cycle is critical.
- Many new risks and unknowns
 - The overall complexity of privacy protection in the cloud represents a bigger challenge.



Responsibility

A Solution: Privacy-Preserving Noisy Keyword Search in cloud Computing

- Privacy-Preserving Noisy Keyword Search in Cloud Computing
 - Xiaoqiong Pang, Bo Yang, and Qiong Huang
 - ICICS 2012



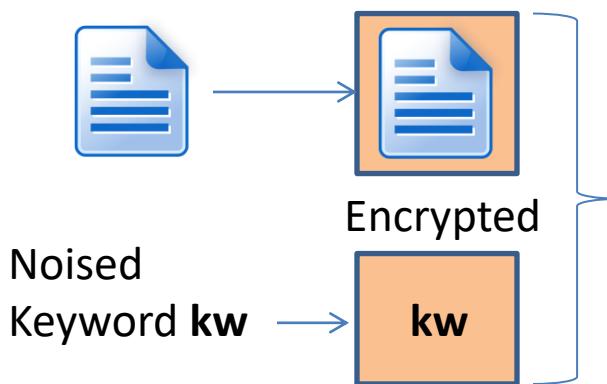
Motivation



- Consider a cloud data system: a user with either limited resources or limited expertise wishes to outsource its private data to an untrusted cloud server in a private manner, while maintaining the ability to retrieve the stored data.
- The informal security requirement for this system claims that the cloud server should not learn any useful information about the data that it stores for the user and the queried words.
- A feasible solution to preserve privacy is to design a searchable encryption scheme such that:
 - the records are stored in disguised/encrypted form.
 - the key employed to encrypt the data is kept secret from the cloud server.
 - the records can be searched for securely and efficiently
- Why privacy-preserving noisy-based search ?
 - For example, biometric data are noisy, even two readings of the same biometric source are rarely identical. Therefore, exact-match search over biometric data does not work.
 - privacy-preserving noisy-keyword-based search on remote encrypted data in a fault-tolerant manner

Scenario

Document-storage

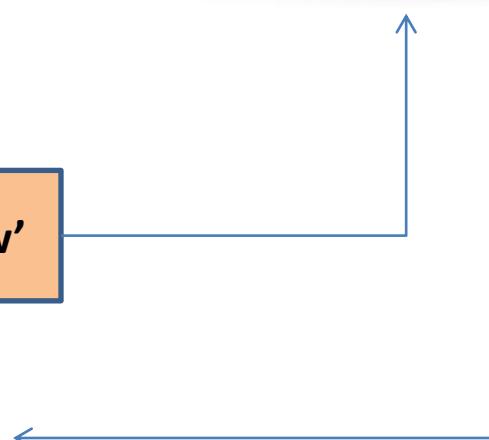


Noised
Keyword **kw**



Search

Keyword **kw'**



Return **encrypted documents** whose keywords **kw** are close to the **kw'**

Hamming Distance

- Different noisy data, such as different biometric information, has different error patterns. In concrete construction, we consider

$$\sum_{k=1}^n (w_{ik} - x_k)^2$$

- as metric to measure the **closeness** between binary strings

$$x = x_1 \cdots x_n$$

$$w_i = w_{i1} \cdots w_{in}$$

fuzzy extractor

- A fuzzy extractor is a pair of efficient randomized procedures (Gen, Rep) such that the following hold:
 - Given $w \in M$,
 - EXT. Gen outputs an extracted string $R \in \{0,1\}^l$ and a helper string $P \in \{0,1\}^*$.
 - EXT. Rep takes as input an element $w \in M$ and a string $P \in \{0,1\}^*$.
 - Correctness:
 - If $\text{dis}(w, w') \leq t$, a threshold and $(R, P) \leftarrow \text{EXT. Gen}(w)$, then $\text{EXT. Rep}(w', P) \rightarrow P$.
 - The string R is nearly uniform even given P

Key Generation

- Given the security parameter k , sample a random number $K_1 \in \{0,1\}^k$
- Choose a random invertible matrix $Q \in M_{n+3,n+3}(F)$, where $M_{n+3,n+3}(F)$ is a predetermined finite integral matrix group consisting of invertible $(n + 3) \times (n + 3)$ matrices over field F
- Output the **secret key** $K = (K_1, Q)$

Document-Storage: Init

- Given the key $K = (K_1, Q)$ and a document set $D = \{D_1, D_2, \dots, D_N\}$ of N documents
- Scan $D = \{D_1, D_2, \dots, D_N\}$ and generate a set of noisy keywords $\delta(D)$ for D
- For all $w_i \in \delta(D)$, output $D(w_i)$, the set of identifiers of documents in D that are labeled with noisy keyword w_i .

Document-Storage: BuildIndex

- Given $K = (K_1, Q)$, $\delta(D)$, $\{D(w_i) | w_i \in \delta(D)\}$
- Randomly choose $R \in F$
- For $1 \leq i \leq |\delta(D)|$, Create index $I_{w_i} = (A_{w_i}, B_{w_i})$ for keyword $w_i = w_{i1} \cdots w_{in}$
 - For each keyword $w_i = w_{i1} \cdots w_{in} \in \delta(D)$, choose a random $R_i \in F$, compute $n+3$ size vector $W_i = (\sum_{k=1}^n w_{ik}^2 + R - R_i, w_{i1}, \dots, w_{in}, 1, R_i)$, then compute $A_{w_i} = Q \cdot W_i^T$
 - For $w_i = w_{i1} \cdots w_{in}$, use the Extract algorithm to output an extracted string k_i and a helper P_i , i.e., $EXT.Gen(w_i) \rightarrow k_i, P_i$
 - Compute $B_{w_i} = \pi_{K_1}(D(w_i) || P_i)$, where π is a pseudorandom permutation

Document-Storage: Data-Storage

- all documents in $D = \{D_1, D_2, \dots, D_N\}$
- For $j = 1, \dots, N$
- For $i = 1, \dots, |\delta(D)|$
- for each $D_j \in D$
- if D_j is associated with keywords w_i
- compute $c_j = AES_{k_i}(D_j)$
- end if
- end for
- end for
- End for

- All encrypted documents c_j are stored in cloud together with the index $I = \{I_{w_i} = (A_{w_i}, B_{w_i}) \mid w_i \in \delta(D)\}$

Search: User side

- For any query $x = x_1 \cdots x_n$, user
- Choose a random number $R_A \in F$
- Construct a vector
 $X = (1, -2x_1, \dots, -2x_n, R_A, 1)$
- Generate a trapdoor $t = XQ^{-1}$ under secret key Q
- Send $t = XQ^{-1}$ to the server.

Search: Cloud Server side

- For $i = 1, \dots, |\delta(D)|$ /* the number of keywords */
 - take out the index $I_{w_i} = (A_{w_i}, B_{w_i} = \pi_{K_1}(D(w_i) || P_i))$
 - compute $t \times A_{w_i} = XQ^{-1}QW_i^T = XW_i^T$
 - get $dis' = \min_{i=1}^{|\delta(D)|} XW_i^T$ and the corresponding index i
- return (dis', B_{w_i}) to the user
- End for

Search: User side (2)

- Compute

$$dis(x, w_i) = dis' + \sum_{k=1}^n x_k^2 - R - R_A$$

- If $dis(x, w_i) > threshold$,
- there is no matched record
- Else
- compute
$$\pi^{-1}(K_1, B_{w_i}) = \pi^{-1}(K_1, \pi_{K_1}(D(w_i) || P_i)) = D(w_i) || P_i$$
- send $D(w_i)$ to the cloud server.
- End if
- **Note that:** $D(w_i)$: the set of identifiers of documents in D that are labeled with noisy keyword w_i .

Search: Cloud Server side (2)

- The server returns encrypted documents C_{w_i} identified by $D(w_i)$

Search: User side (3)

- The server returns encrypted documents C_{w_i} identified by $D(w_i)$
- The user
 - computes $EXT.\ Rep(x, P_i)$ to reproduce the decryption key k_i
 - Recover C_{w_i} with k_i

Correctness

- For a query string $x = x_1 \cdots x_n$,
Let $X = (1, -2x_1, \dots, -2x_n, R_A, 1)$

- For each $w_i = w_{i1} \cdots w_{in}$
Let $W_i = (\sum_{k=1}^n w_{ik}^2 + R - R_i, w_{i1}, \dots, w_{in}, 1, R_i)$
- Thus,

$$\begin{aligned} XQ^{-1}QW_i^T &= XW_i^T = \left(\sum_{k=1}^n w_{ik}^2 + R - R_i \right) - 2(x_1w_{i1} + \dots + x_nw_{in}) + R_A + R_i \\ &= \sum_{k=1}^n (w_{ik} - x_k)^2 - \sum_{k=1}^n x_k^2 + R_A + R \end{aligned}$$

- Therefore,

$$\sum_{k=1}^n (w_{ik} - x_k)^2 = XW_i^T + \sum_{k=1}^n x_k^2 - R_A - R$$

- Since $\sum_{k=1}^n x_k^2 - R_A - R$ is a constant, the server can use XW_i^T to compute the closest match and return corresponding (dis', B_{w_i}) .

Thank
you



CS4413/6413: Foundations of Privacy --- Winter 2019

Topic 11: Privacy-preserving data aggregation in
smart grid

Lecturer: Rongxing LU

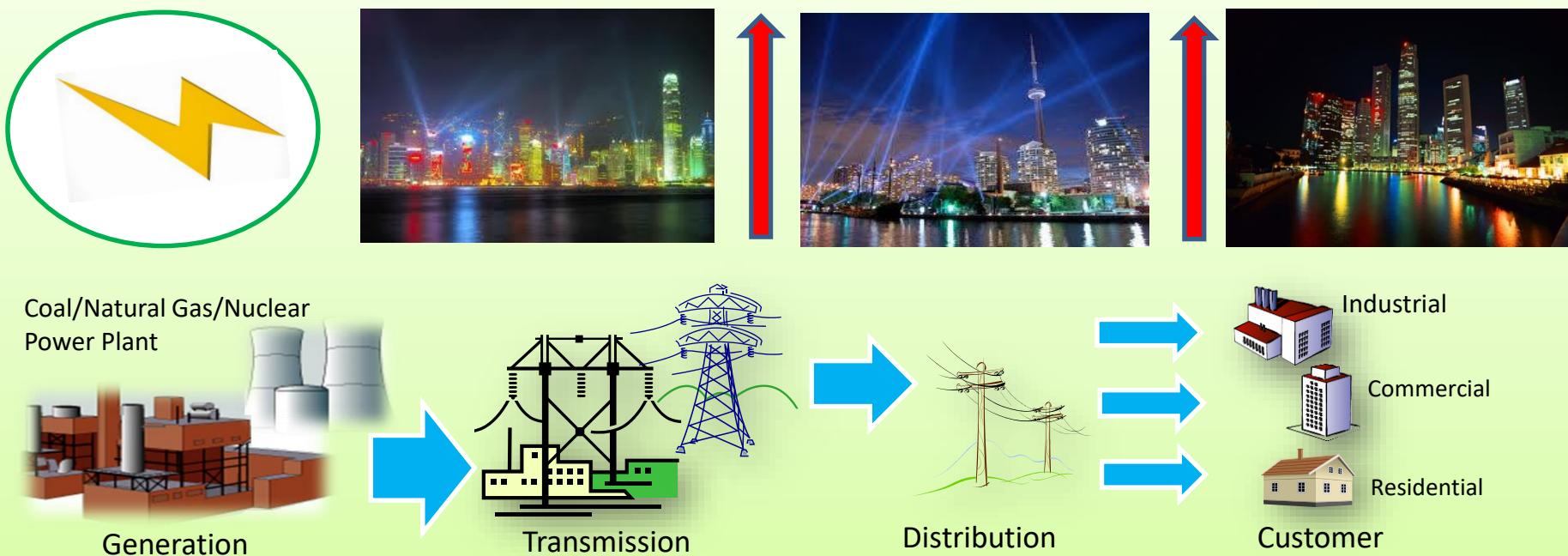
Email: RLU1@unb.ca Office: GE 114

Website: <http://www.cs.unb.ca/~rlu1/>

Faculty of Computer Science, University of New Brunswick

Current Stage of Electrical Grid

- One of the important engineering achievements of the 20th century [1].
- With affordable and reliable electric power,
 - our daily lives have been improved tremendously.
- However, some basic designs are little changed [2]
 - Typical characteristic: **one-way electricity flow**

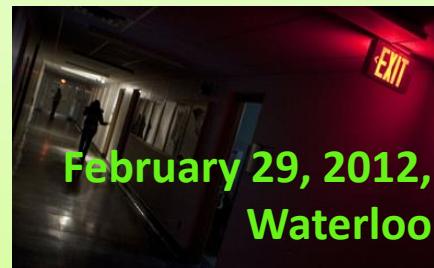
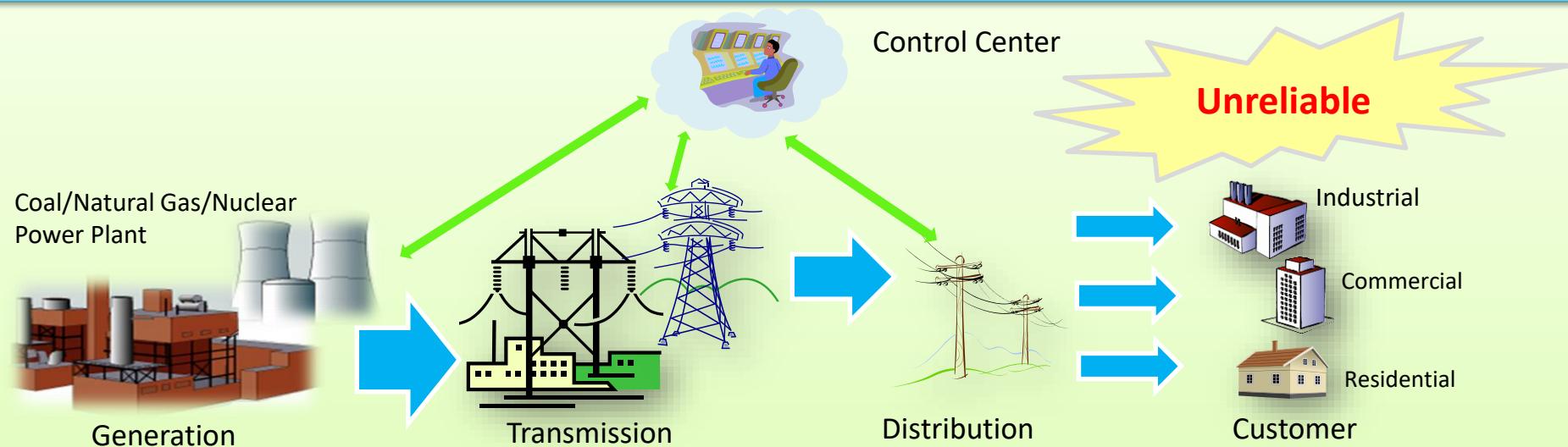


[1] W. A. Wulf, "Great Achievements and Grand Challenges," *The Bridge*, pp. 5–10, Fall/Winter, 2000.

[2] G.W. Arnold, "Challenges and Opportunities in Smart Grid: A Position Article," *Proceedings of the IEEE*, vol. 99, no. 6, pp. 922-927, 2011.

Challenges in Electrical Grid

- What's the problem in one-way flow?
 - Have no information about the cost of electricity, no idea on whether it is overloaded.
- Electrical grid becomes unreliable
- Even though info. flow shared between Control Center and Grid
 - Not enough! Eg. Blackout on Aug. 14, 2003, Feb. 29, 2012 , July 30, 2012



Modernization of Electrical Grid

- Modernization of the aging electrical grid -- a major national priority



- Est. The investment required in replacement/new generation is $\approx \$560$ billion by 2030 in US [1].
- In Canada, Ontario's energy ministry planned to provide $\$20$ billion in investment [2].
- According to the plan of State Grid Corporation of China (SGCC), the total investment will be $\text{RMB } 286.11$ billion during the "12th Five-Year Plan" period, with annual investment of $\text{RMB } 57.22$ billion on average [3].



[1] P. S. Fox-Penner et al., "Transforming America's power industry: The investment challenge preliminary findings," in Edison Foundation Conference, Apr. 21, 2008.

[2] <http://www.reuters.com/article/2011/07/07/us-energy-ontario-idUSTRE7665S320110707>

[3] http://www.businessreviewcanada.ca/press_releases/china-smart-grid-equipment-segmented-market-report-2010-2012

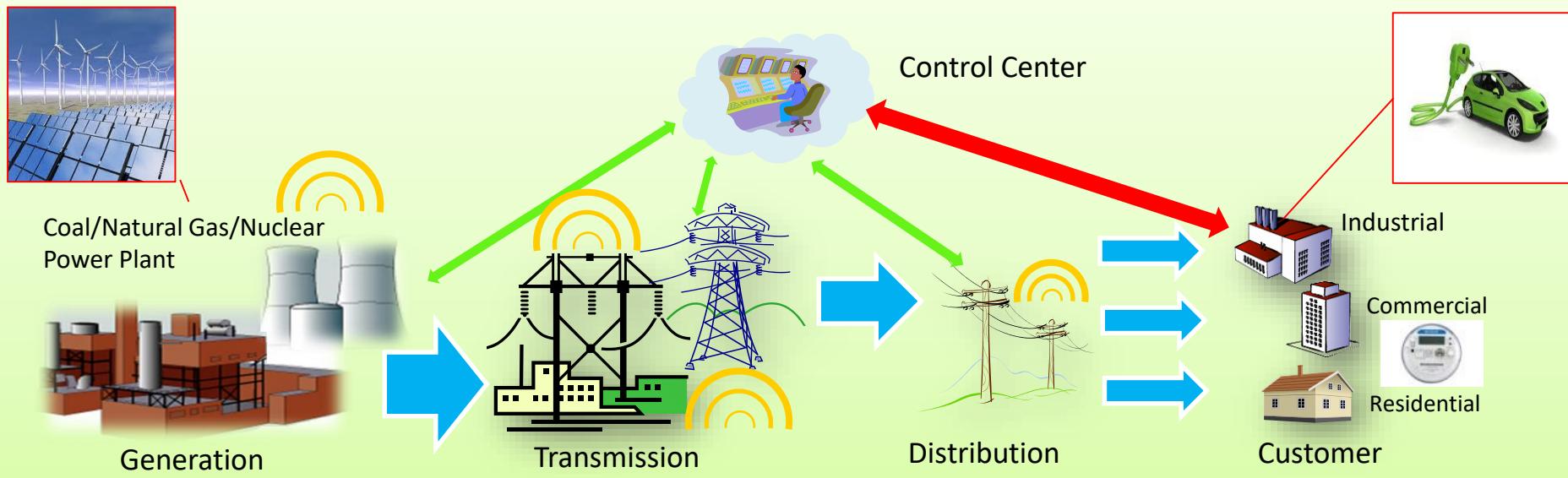
Goals of Electrical Grid Modernization

- To make the production and delivery of electricity more cost-effective
- To provide consumers with electronically available information
 - About the energy consumption and costs
- To help reduce the production of greenhouse gas emission
- To improve the reliability of service
- To prepare the grid to support a growing fleet of electric vehicles



The Concept of Smart Grid

- The essential concept of the smart grid is the integration of ... → power system
 - advanced information technology
 - digital communications, sensing
 - measurement and control technologies (**Smart Meters**)
- To make the Electrical Grid more **intelligent**
- Integrate **renewable sources, electric vehicles** to achieve above goals

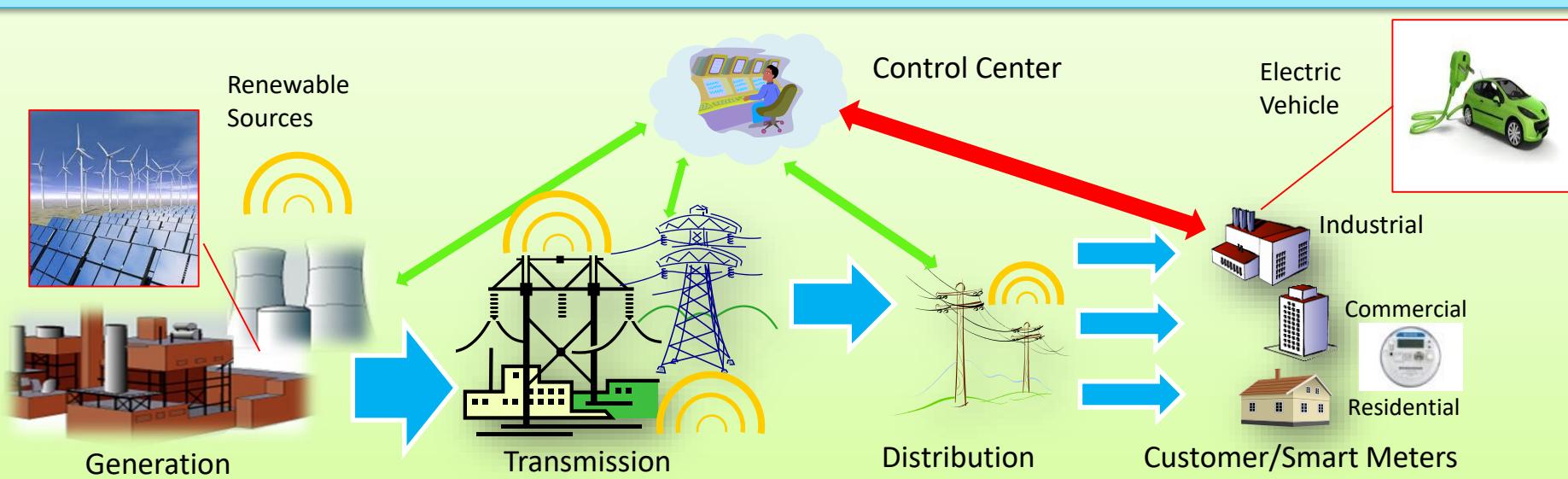
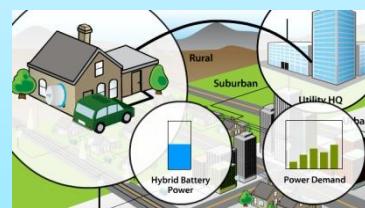


The Characteristics of Smart Grid

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

Research Issues in Smart Grid

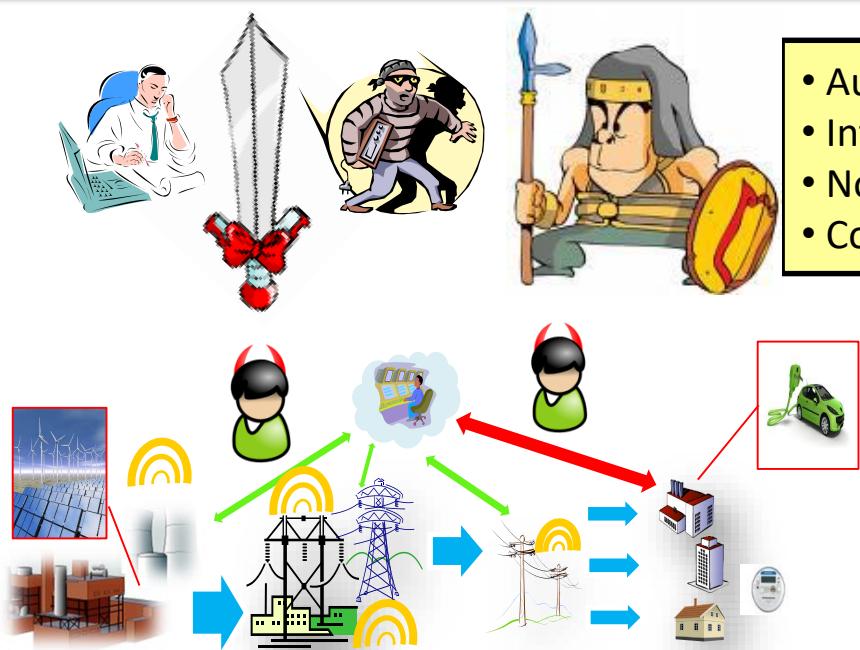
- Communication
 - Heterogeneous communication architectures
- Distributed Generation
 - Fixed (not so adaptive) electricity supply
 - Diversifying power generation options (i.e., renewable sources)
- Vehicle to Grid Systems
- Energy Storage
 - Battery Management



Security Challenge in Smart Grid

- Besides the above research issues, security is also challenging
- Information flow -> 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

Security



- Authentication
- Integrity
- Non-repudiation
- Confidentiality

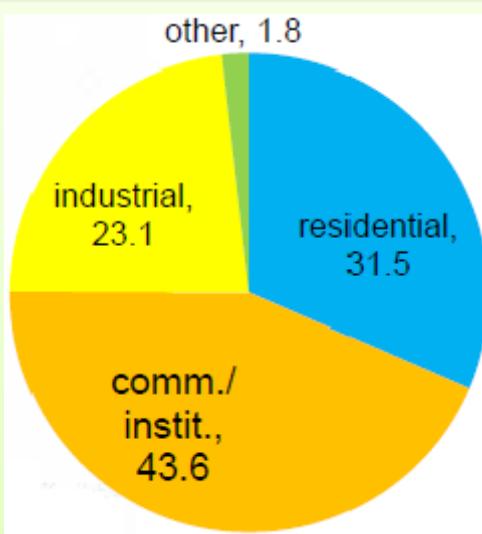


CIA Director calls Smart grid "Stupid" due to Security problems

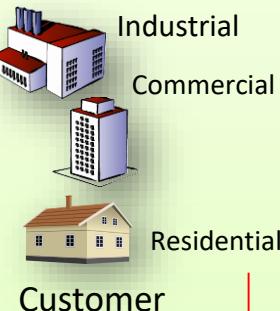
<http://www.youtube.com/watch?v=MAid1bS8t9U>

Privacy Challenge in Smart Grid

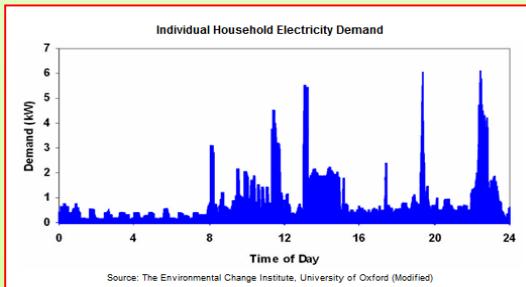
- Privacy is also challenging in Smart Grid
- Residential customers care more about their privacy
- Actually, privacy in smart grid has been paid great attention



**Electricity demand in
Ontario 2008**



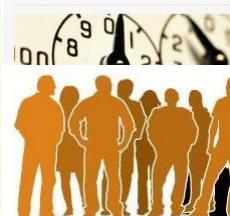
**Smart meter
Individual**



Smart meters could b US Energy Department in smart grid privacy warning
Smart meters could become a 'spy in the authorities to monitor households, adding

By Alastair Jamieson
Published: 10:30AM BST 11 Oct 2009

Comments



ABOUT MISSION

06/25/2010

**Smart meters - the latest e
the world in invasive tech!**

We've written before about the privacy implications of Smart Meters, both here and abroad. This story over at the Washington Post begins:

CAMBRIDGE, England - Wary homeowners could scupper the rollout of smart technologies meant to boost energy efficiency, without secure controls over data and access to appliances, executives said this week.

Great, yet another technological debate in which the UK is the pilot for the most intrusive equipment around. "Home meters" allow two-way wireless communication demand and charge more at peak times and even switch remotely.

The rollout of smart grid technologies into homes raises several data privacy issues lawmakers need to recognise and address, a new US Department of Energy report cautions.

The concerns over privacy are related to the collection and use of energy consumption data gathered from homes in which the technologies are going to be installed over the next several years, the department report noted.

"Consumer-specific energy usage data has enormous potential to enable utilities or other third-party service providers to help consumers significantly reduce energy consumption," the Department of Energy noted.

However, it said that "because such data can also disclose fairly detailed information about the behavior and activities of a particular household," controls needs to be implemented for ensuring the data is collected, used and shared in line with privacy expectations.

A smart grid basically uses digital technology to transmit, distribute and deliver power to consumers in a more reliable and efficient

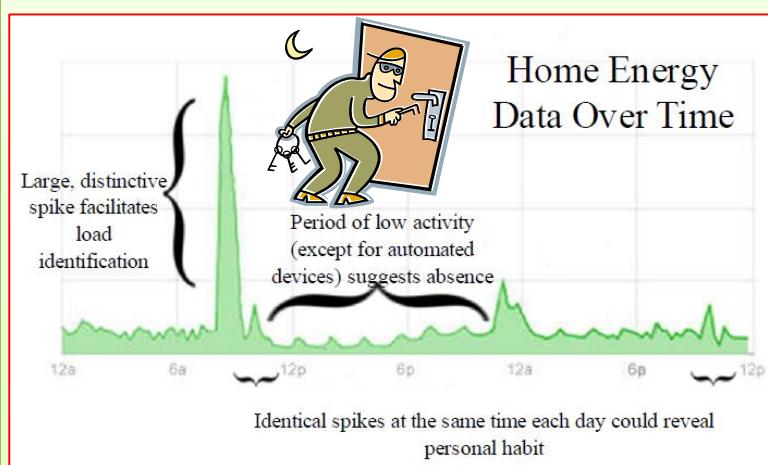
Privacy concerns scotch Smart Meters plan in Holland

Back at the beginning of December, I wrote a piece and spoke on the radio regarding Ed Miliband's announcement that the government would soon be rolling-out Smart Meters across the UK - and the danger that this posed to the sovereignty of our energy supply and the uncertainties surrounding the information that utility companies would now have immediate access to.

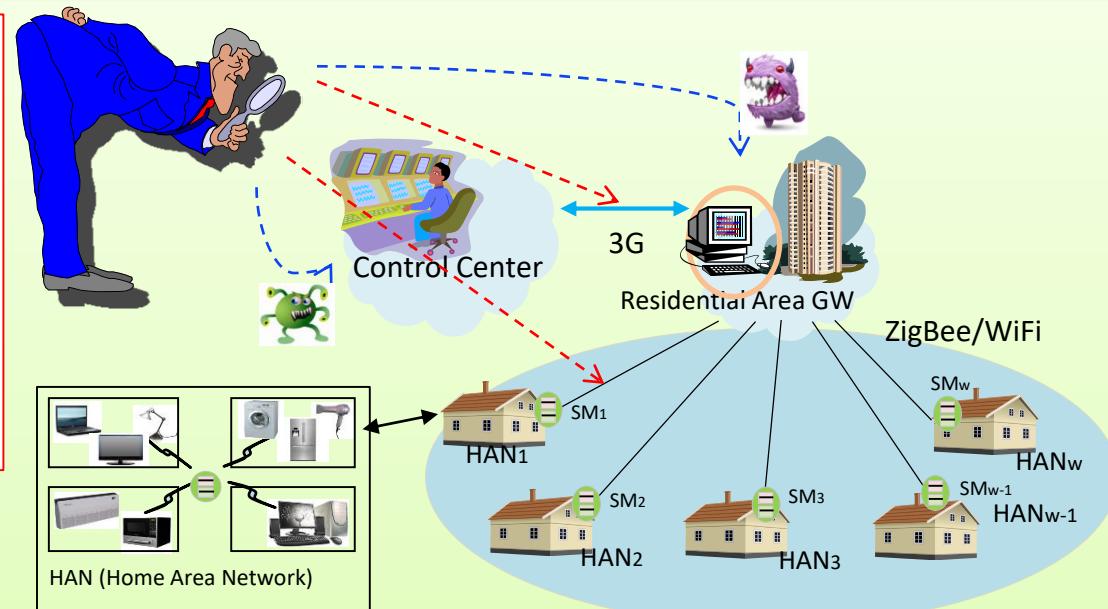


Privacy Enhanced Techniques for Smart Grid

- Privacy Enhancing Technique -- to address Security & Privacy challenges in Smart Grid -- in the following scenario with Privacy Threat Model
- To protect user's privacy, subsequently keep the property security



Fine-Grained Consumption Data



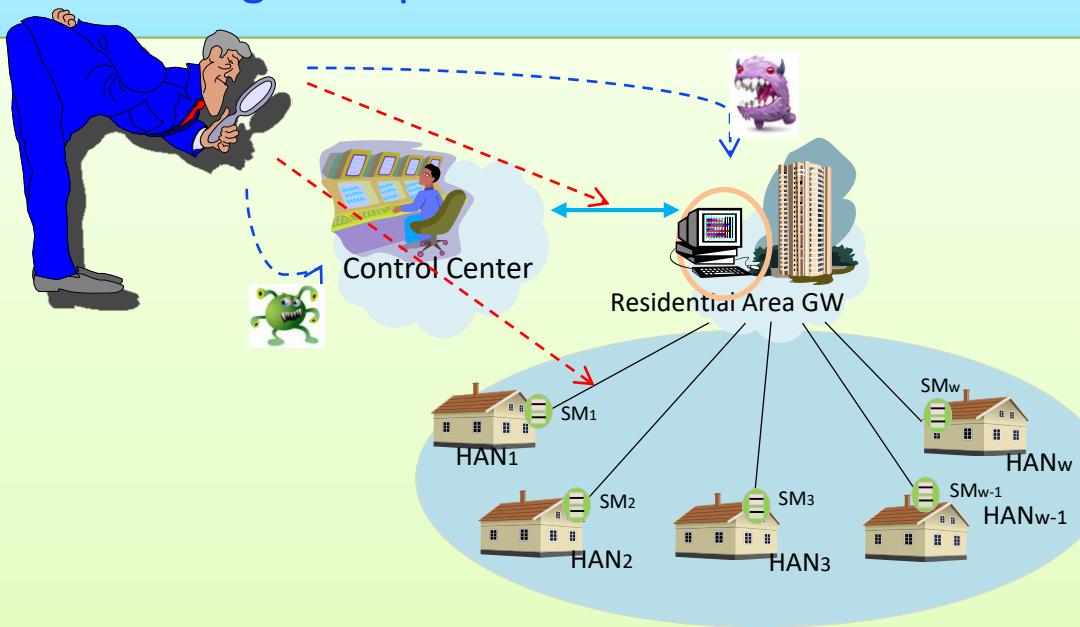
Personal Information can be inferred:

- When you are at home
- Which appliances you use
- When you eat
- Whether you arrive late to work

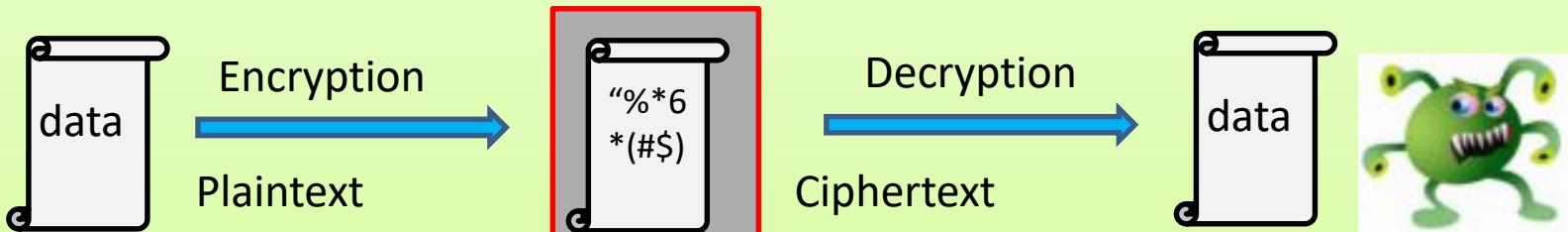


First Attempt: Encryption

- Encryption: to protect user's privacy
- Encryption is **not sufficient** to protect user's privacy in threat model
- Cause **huge computation cost** at control center on decryption



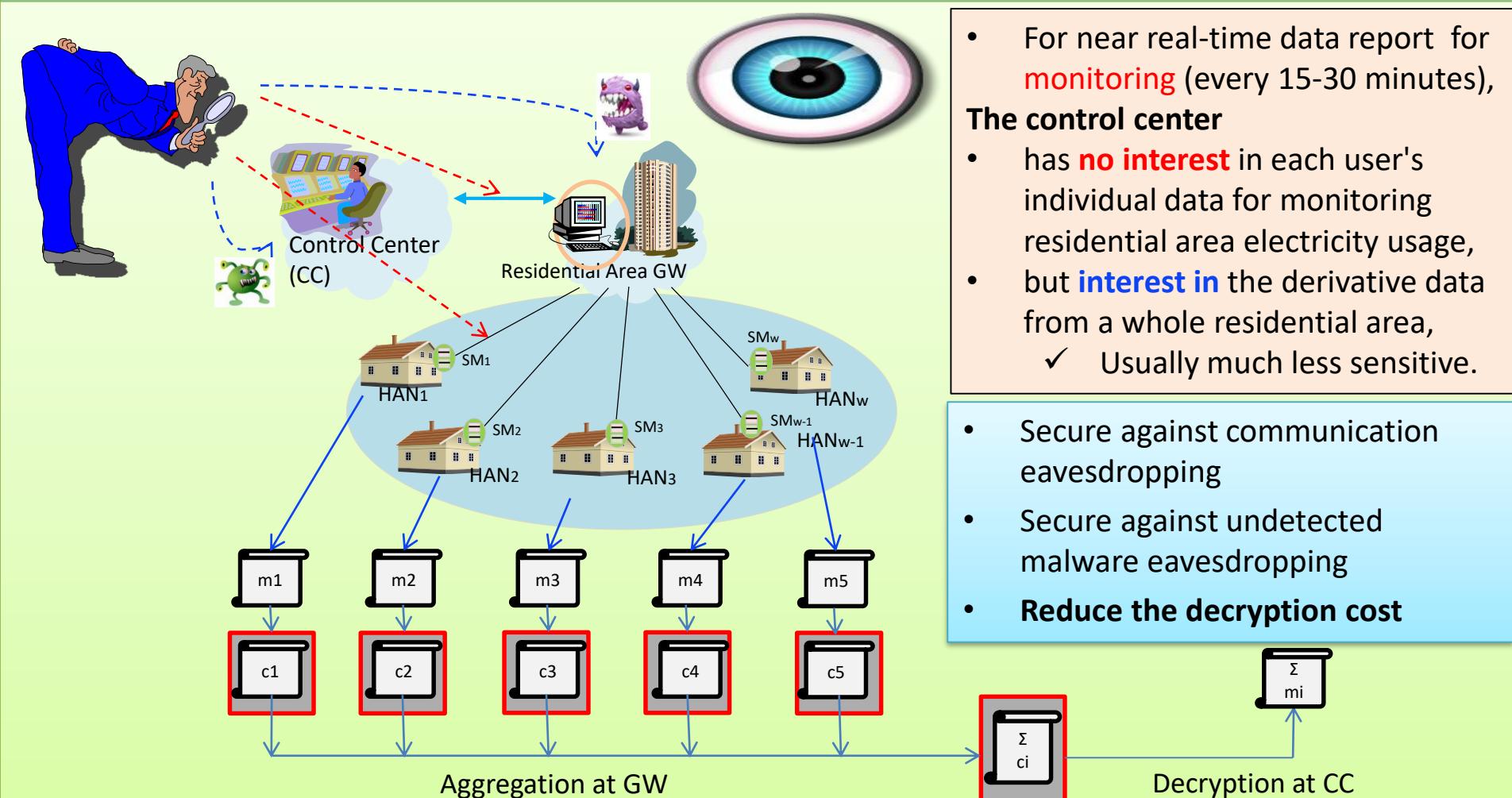
- In addition, the data report in every 15-30 minutes [1] → incur **huge decryption tasks** at the control center



[1] http://research.microsoft.com/en-us/projects/privacy_in_metering/

Second Attempt: Aggregation Encryption

- Aggregation Encryption: to protect user's privacy
- Aggregation Encryption is possible to protect user's privacy



Communication Cost in Aggregation Encryption

- Aggregation requires Homomorphic Encryption, i.e., Paillier Cryptosystem[1]

$$E(M_1) \circ E(M_2) \circ \cdots \circ E(M_w) = E(M_1 + M_2 + \cdots + M_w) \xrightarrow{Dec} \sum_{i=1}^w M_i$$

Key Generation :

$$p_1, q_1, n = p_1 q_1, \lambda = lcm(p_1 - 1, q_1 - 1)$$

$$L(u) = \frac{u-1}{n}, g \in \mathbb{Z}_{n^2}^*, \mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$$

$$pk = (n, g), sk = (\lambda, \mu)$$

Encryption :

$$m \in \mathbb{Z}_n, r \in \mathbb{Z}_n^*, c = E(m) = g^m \cdot r^n \bmod n^2$$

Decryption :

$$c \in \mathbb{Z}_{n^2}^*, m = D(c) = L(c^\lambda \bmod n^2) \cdot \mu \bmod n$$

Aggregation :

$$m_i \in \mathbb{Z}_{n^2}^*, r_i \in \mathbb{Z}_n^*, c_i = E(m_i) = g^{m_i} \cdot r_i^n \bmod n^2, i = 1, \dots, w$$

$$C = \prod_{i=1}^w c_i = \prod_{i=1}^w g^{m_i} \cdot r_i^n \bmod n^2 = g^{\sum_{i=1}^w m_i} \cdot \left(\prod_{i=1}^w r_i \right)^n \bmod n^2$$

A CLOSE LOOK

To achieve 1024-bit RSA security, $|n|=1024$

Message Space

$$m \in \mathbb{Z}_n \rightarrow \{0,1\}^{1024}$$

Ciphertext Space

$$C \in \mathbb{Z}_{n^2}^* \rightarrow \{0,1\}^{2048}$$



Communication Cost in Aggregation Encryption (2)

- Message Space $m \in Z_n \rightarrow \{0,1\}^{1024}$ Ciphertext Space $C \in Z_{n^2}^* \rightarrow \{0,1\}^{2048}$
- In smart grid scenario
 - every 15 minutes, each user's electricity consumption should be less than 100 kwh
 - assume there are 10,000 users in a residential area
 - →  the aggregated data value < 1,000,000
- → Actual Message Space is $\{0,1\}^{20}$ $\{0,1\}^{1024}$

VERY INEFFICIENT IN COMMUNICATION

Aggregated Message 1000000

Ciphertext of Aggregated Message

412023436986659543855531365332575948179811699844327982845455626433876445565248426198091234867
893641941202343698665954385553136533257594817981169984432798284545562643387644556524842619809
123486789364194120234369866595438555313653325759481798116998443279828454556264338764455652484
261980912348678936419412023436986659543855531123853120234369866595438555313653325759481798116
998443279828454556264338764455652484261980912348678936419412023436986659543855531365332575948
179811699844327982845455626433876445565248426198091234867893641941202343698665954385553136533
257594817981169984432798284545562643387644556524842619809123486789364194120234369866595438555
31123853

(around 620 decimal digits)

Even Worse in Smart Grid

- For fine-grained monitoring
 - multiple smart appliances => **multi-dimensional data**
 - for nearly-real time usage data collection
 - Example: Monitoring CO₂ emission in a residential area
 - different appliance produces different emission of CO₂, impacting on environment [1]
 - Communication resources are **terribly wasted**
- For example, for 2-dimensional data
 - (1000000, 1000000)



	Air-Condition A units	Refrigerator B units
U1	a1	b1
U2	a2	b2
U3	a3	b3
U4	a4	b4
:	:	:
	Σai	Σbi
$CO_2 \text{ emission} = A * \sum ai + B * \sum bi$		

```

412023436986659543855531365332575948179811699844327982845455626433876445565248426198091234867893641941
202343698665954385553136533257594817981169984432798284545562643387644556524842619809123486789364194120
234369866595438555313653325759481798116998443279828454556264338764455652484261980912348678936419412023
436986659543855531123853120234369866595438555313653325759481798116998443279828454556264338764455652484
261980912348678936419412023436986659543855531365332575948179811699844327982845455626433876445565248426
198091234867893641941202343698665954385553136533257594817981169984432798284545562643387644556524842619
80912348678936419412023436986659543855531123853
412023436986659543855531365332575948179811699844327982845455626433876445565248426198091234867893641941
202343698665954385553136533257594817981169984432798284545562643387644556524842619809123486789364194120
234369866595438555313653325759481798116998443279828454556264338764455652484261980912348678936419412023
436986659543855531123853120234369866595438555313653325759481798116998443279828454556264338764455652484
261980912348678936419412023436986659543855531365332575948179811699844327982845455626433876445565248426
198091234867893641941202343698665954385553136533257594817981169984432798284545562643387644556524842619
80912348678936419412023436986659543855531123853

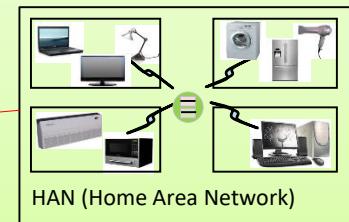
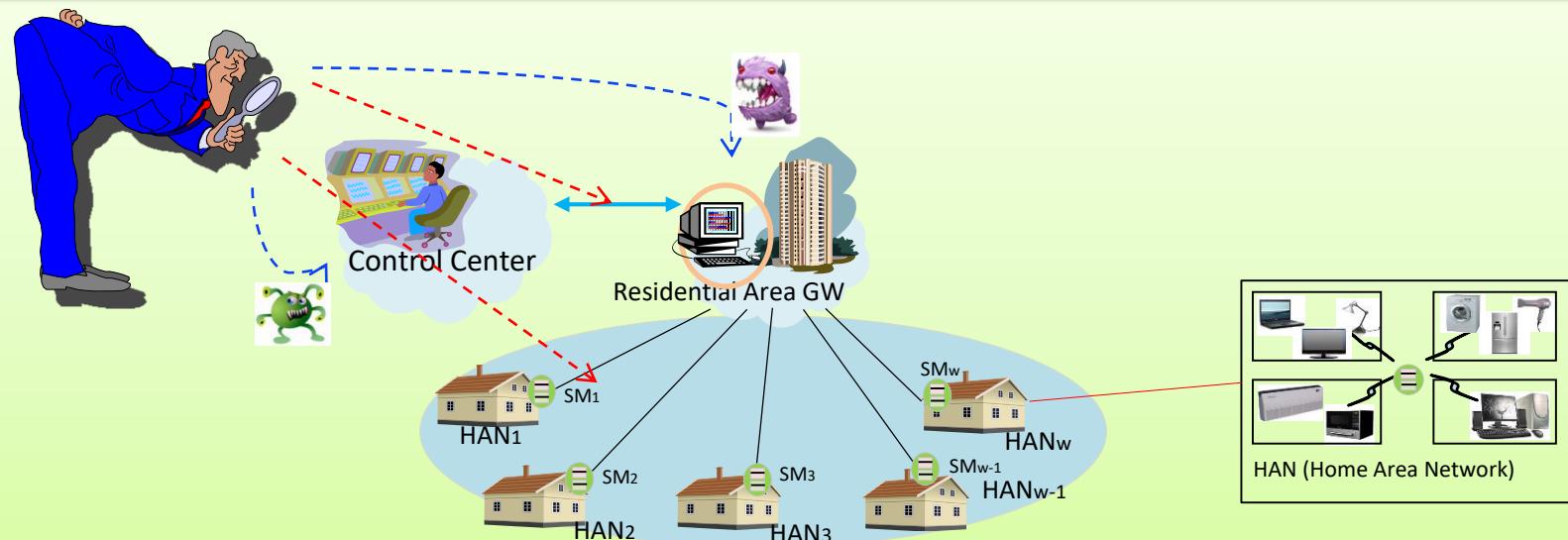
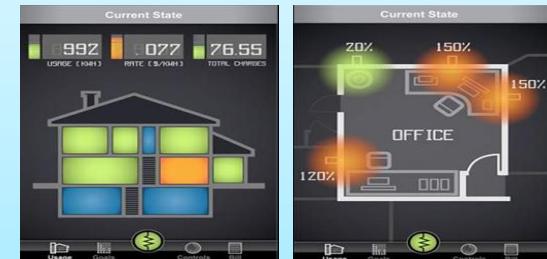
```

Research Challenge

- Is there a solution for multi-dimensional data aggregation in smart grid?
with research goal:
 - Protect residential user privacy
 - Computation efficiency
 - Communication efficiency
- The answer is Yes.
 - EPPA: An Efficient and Privacy-Preserving Aggregation Scheme [1]



Fine-Grained Consumption Data



Privacy Enhancing Technique -- EPPA

- Three Goals
 - Protect residential user privacy
 - Computation efficiency
 - Communication efficiency



- Aggregation Encryption

- For 1 and 2 Goals

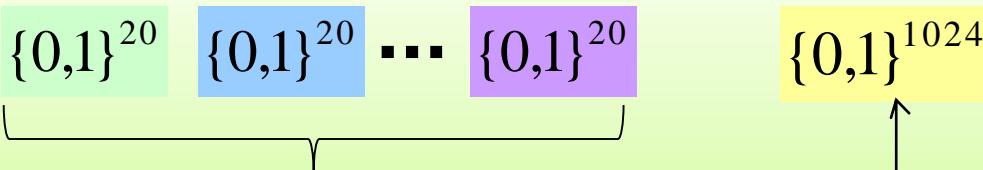
$$E(M_1) \circ E(M_2) \circ \dots \circ E(M_w) = E(M_1 + M_2 + \dots + M_w) \xrightarrow{Dec} \sum_{i=1}^w M_i$$

- Multi-dimensional Aggregation Encryption

- For 1, 2, and 3 Goals

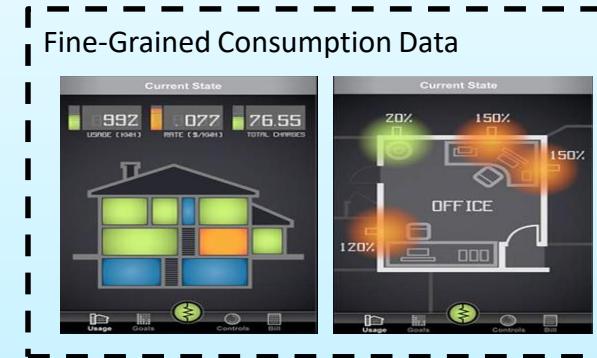
$$E(M_{11}, M_{12}, \dots, M_{1l}) \circ E(M_{21}, M_{22}, \dots, M_{2l}) \circ \dots \circ E(M_{w1}, M_{w2}, \dots, M_{wl})$$

$$= E\left(\sum_{i=1}^w M_{i1}, \sum_{i=1}^w M_{i2}, \dots, \sum_{i=1}^w M_{il}\right) \xrightarrow{Dec} \sum_{i=1}^w M_{i1}, \sum_{i=1}^w M_{i2}, \dots, \sum_{i=1}^w M_{il}$$



41202343698665954385531365332575948179811699844327982845455626433876445565248426198098870423161841879
261420247188869492560931776375033421130982397485150944909106910269861031862704114880866970564902903653
65886743373172081310410519086425479328260139125762403394637326939141202343698665954385531365332575948
1798116998443279828454562643387644556524842619809870423161841879261420247188869492560931776375033421
130982397485150944909106910269861031862704114880866970564902903653658867433731720813104105190864254793
28260139125762403394637326939141202343698665954385531365332575948179811699844327982845455626433876445
565248426198098870423161841879261420247188869492560931776375033421130982397485150944909106910269861031
8627041148808669705649029036536588674337317208131041051908642547932826013912576240339463732693911669705
64902903653658867433731720813104105190864254793282601391257624033946373269391

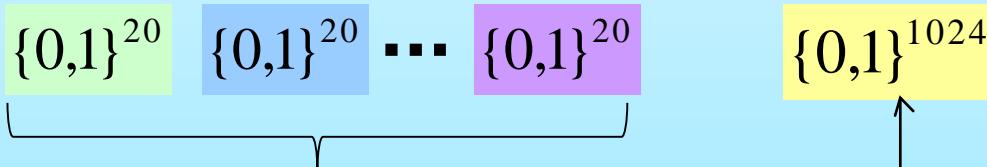
(around 620 decimal digits)



Privacy Enhancing Technique – EPPA (2)

$$\circ E(M_{11}, M_{12}, \dots, M_{1l}) \circ E(M_{21}, M_{22}, \dots, M_{2l}) \circ \dots \circ E(M_{w1}, M_{w2}, \dots, M_{wl})$$

$$= E\left(\sum_{i=1}^w M_{i1}, \sum_{i=1}^w M_{i2}, \dots, \sum_{i=1}^w M_{il}\right) \xrightarrow{Dec} \boxed{\sum_{i=1}^w M_{i1}, \sum_{i=1}^w M_{i2}, \dots, \sum_{i=1}^w M_{il}}$$



4120234369866595438555313653325759481798116998447982845455626433876445565248426198098870423161841879
261420247188869492560931776375033421130982397485150944909106910269861031862704114880866970564902903653
65886743373172081310410519086425479328260139125762403394637326939141202343698659543855531365332575948
179811699844327982845455626433876445565248426198098870423161841879261420247188869492560931776375033421
130982397485150944909106910269861031862704114880866970564902903653658867433731720813104105190864254793
282601391257624033946373269391412023436986659543855531365332575948179811699844327982845455626433876445
56524842619809887042316841879261420247188869492560931776375033421130982397485150944909106910269861031
862704114880866970564902903653658867433731720813104105190864254793282601391257624033946373269391669705
64902903653658867433731720813104105190864254793282601391257624033946373269391 (around 620 decimal digits)

- Research Challenges
 - how to assemble multidimensional aggregated data into one?
 - how to recover multidimensional aggregated data from one?
- Contribution
 - propose an EPPA (Efficient and Privacy-Preserving Aggregation) scheme
 - to address the above challenges

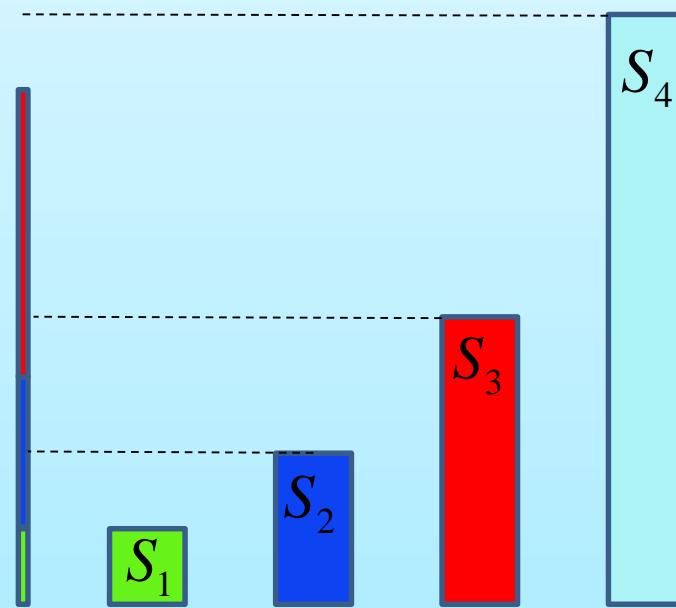


Tech. Background: Superincreasing Sequence

- **Superincreasing Sequence:** In mathematics, a sequence of positive integer numbers (S_1, S_2, \dots, S_n) is called superincreasing if every element of the sequence is greater than the sum of all previous elements in the sequence.

$$S_{w+1} > \sum_{j=1}^w S_j$$

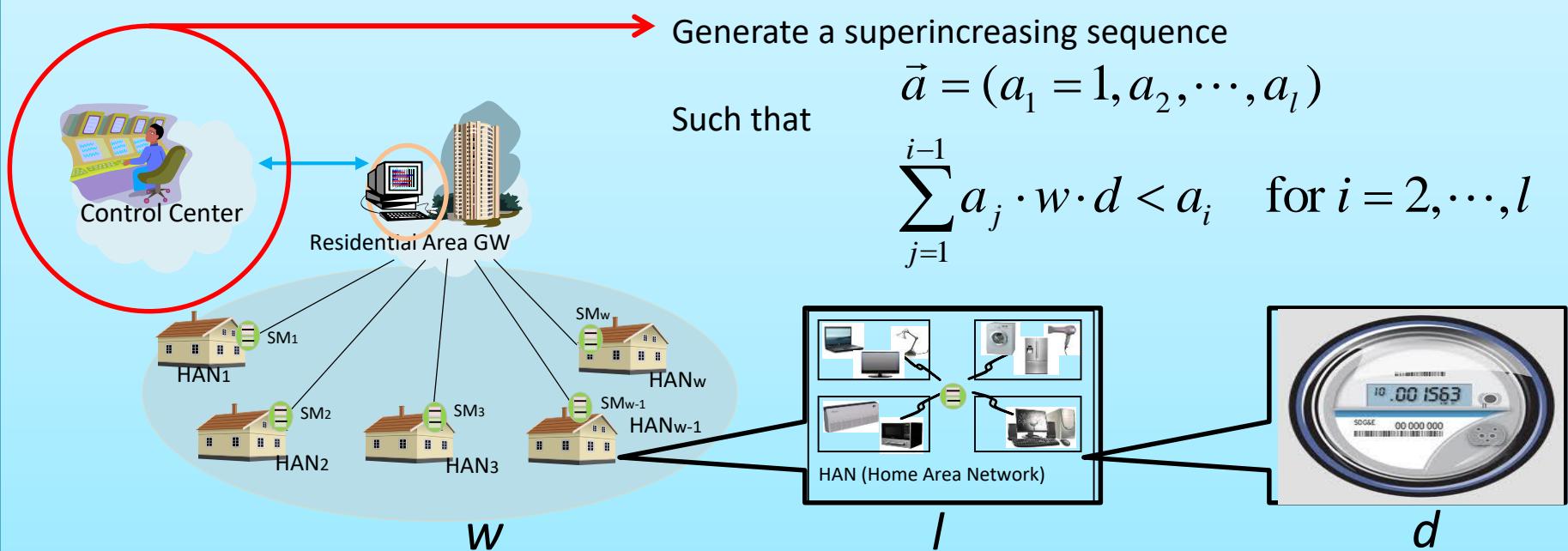
Sum: $1 < 3$
Sum: $1+3 = 4 < 6$
Sum: $1+3+6 = 10 < 13$
Sum: $1+3+6+13 = 23 < 27$
Sum: $1+3+6+13+27 = 50 < 52$



- $(1,3,6,13,27,52)$ is a **superincreasing sequence**

System Assumptions for EPPA

- Max. number of users (HANs) in a residential area is $\leq w$
- Each HAN has I types of electricity usage data (T_1, T_2, \dots, T_I)
- Max. value of T_i is d_i , and $\max(d_i) < d$ (constant)
 - If the collected data is not integer in its original form, it can be easily transformed into an integer $[0, d]$ [1]



EPPA: Key Generation at Control Center



$$p_1, q_1, n = p_1 q_1, \lambda = \text{lcm}(p_1 - 1, q_1 - 1)$$

$$L(u) = \frac{u-1}{n}, g \in \mathbb{Z}_{n^2}^*, \mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$$

$$pk = (n, g), sk = (\lambda, \mu)$$

Paillier
Cryptosystem



$$p_1, q_1, n = p_1 q_1, \lambda = \text{lcm}(p_1 - 1, q_1 - 1)$$

$$L(u) = \frac{u-1}{n}, g \in \mathbb{Z}_{n^2}^*, \mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$$

$$pk = (n, g_1, g_2, \dots, g_l), sk = (\lambda, \mu, \vec{a} = (a_1 = 1, a_2, \dots, a_l))$$

EPPA

A superincreasing sequence

$$\vec{a} = (a_1 = 1, a_2, \dots, a_l)$$

$$\sum_{j=1}^{i-1} a_j \cdot w \cdot d < a_i \quad \text{for } i = 2, \dots, l$$

$$g_i = g^{a_i} \text{ for } i = 1, 2, \dots, l$$

such that

$$\{0,1\}^{20}$$

$$\{0,1\}^{20}$$

$$\cdots \{0,1\}^{20}$$

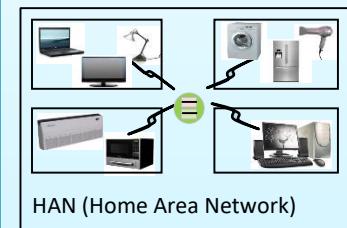
$$\{0,1\}^{1024}$$

Message Space

$$m \in \mathbb{Z}_n \rightarrow \{0,1\}^{|n|}$$

$$\sum_{i=1}^l a_i \cdot w \cdot d < n$$

EPPA: Encryption at HAN



data $m_i = (d_{i1}, d_{i2}, \dots, d_{il})$



$pk = (n, g_1, g_2, \dots, g_l)$

$$m \in Z_n, r \in Z_n^*, c = E(m) = g^m \cdot r^n \bmod n^2$$

Paillier
Cryptosystem



$$m_i = (d_{i1}, d_{i2}, \dots, d_{il}), r_i \in Z_n^*$$

$$c_i = E(m_i) = E(d_{i1}, d_{i2}, \dots, d_{il}) = g_1^{d_{i1}} \cdot g_2^{d_{i2}} \cdots g_l^{d_{il}} \cdot r_i^n \bmod n^2$$

EPPA

$$g_i = g^{a_i} \text{ for } i = 1, 2, \dots, l$$



$$c_i = g^{a_1 d_{i1}} \cdot g^{a_2 d_{i2}} \cdots g^{a_l d_{il}} \cdot r_i^n \bmod n^2$$

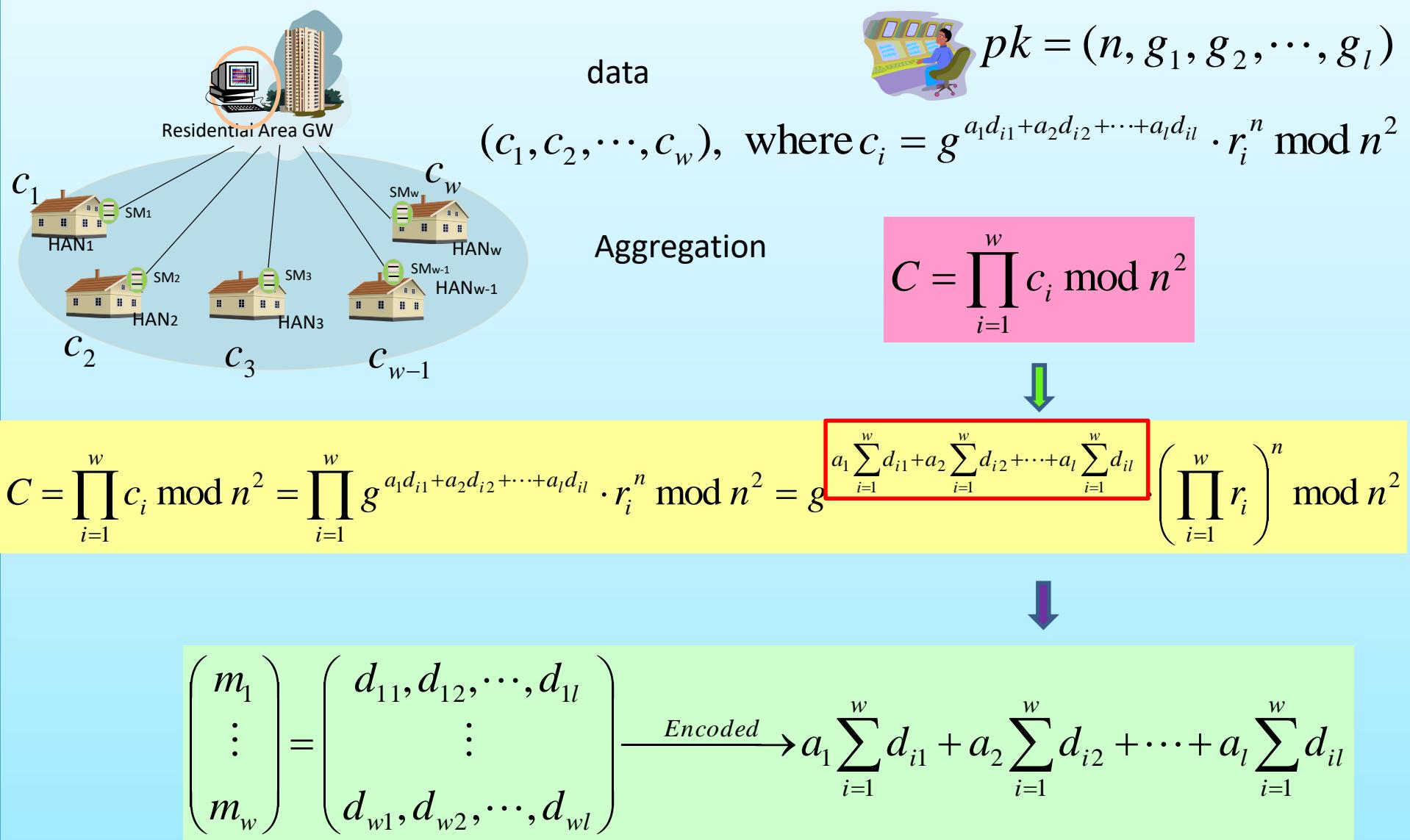


$$c_i = g^{a_1 d_{i1} + a_2 d_{i2} + \cdots + a_l d_{il}} \cdot r_i^n \bmod n^2 = g^{\bar{m}_i} \cdot r_i^n \bmod n^2$$



$$m_i = (d_{i1}, d_{i2}, \dots, d_{il}) \xrightarrow{\text{Encoded}} \bar{m}_i = a_1 d_{i1} + a_2 d_{i2} + \cdots + a_l d_{il}$$

EPPA: Aggregation at Residential Area GW



EPPA: Recovery at Control Center



$$pk = (n, g_1, g_2, \dots, g_l)$$

$$sk = (\lambda, \mu, \bar{a} = (a_1 = 1, a_2, \dots, a_l))$$

$$C = g^{a_1 \sum_{i=1}^w d_{i1} + a_2 \sum_{i=1}^w d_{i2} + \dots + a_l \sum_{i=1}^w d_{il}} \cdot \left(\prod_{i=1}^w r_i \right)^n \mod n^2$$

$$R = \prod_{i=1}^w r_i$$

↓

$$C = g^M \cdot R^n \mod n^2$$

Recover

$$M = a_1 \sum_{i=1}^w d_{i1} + a_2 \sum_{i=1}^w d_{i2} + \dots + a_l \sum_{i=1}^w d_{il}$$

$$c \in Z_{n^2}^*, m = D(c) = L(c^{\lambda \bmod n^2}) \cdot \mu \bmod n$$

$$\sum_{i=1}^l a_i \cdot w \cdot d < n \quad \rightarrow \quad M < n$$

Paillier Cryptosystem

EPPA: Recovery at Control Center (2)



data

$$pk = (n, g_1, g_2, \dots, g_l)$$

$$sk = (\lambda, \mu, \vec{a} = (a_1 = 1, a_2, \dots, a_l))$$

$$M = a_1 \sum_{i=1}^w d_{i1} + a_2 \sum_{i=1}^w d_{i2} + \dots + a_{l-1} \sum_{i=1}^w d_{i(l-1)} + a_l \sum_{i=1}^w d_{il}$$

For example, to recover

$$\sum_{i=1}^w d_{il}$$

Since

$$\sum_{j=1}^{i-1} a_j \cdot w \cdot d < a_i \quad \text{for } i = 2, \dots, l$$

```

1: procedure RECOVER THE AGGREGATED REPORT
Input:  $\vec{a} = (a_1 = 1, a_2, \dots, a_l)$  and  $M$ 
Output:  $(D_1, D_2, \dots, D_l)$ 
2:   Set  $X_l = M$ 
3:   for  $j = l$  to 2 do
4:      $X_{j-1} = X_j \bmod a_j$ 
5:      $D_j = \frac{X_j - X_{j-1}}{a_j} = \sum_{i=1}^w d_{ij}$ 
6:   end for
7:    $D_1 = X_1 = \sum_{i=1}^w d_{i1}$ 
8:   return  $(D_1, D_2, \dots, D_l)$ 
9: end procedure

```

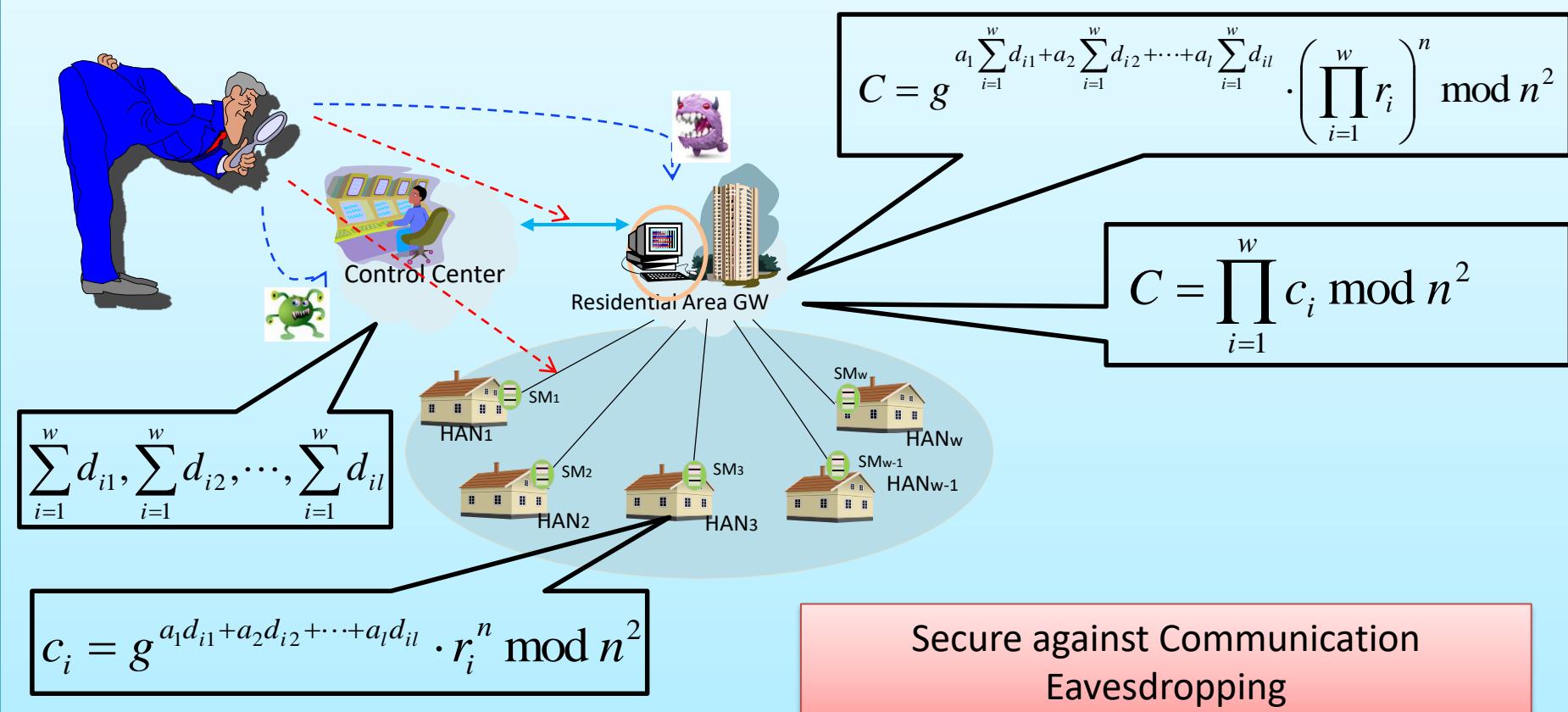
$$a_1 \sum_{i=1}^w d_{i1} + a_2 \sum_{i=1}^w d_{i2} + \dots + a_{l-1} \sum_{i=1}^w d_{i(l-1)} < a_l$$

$$M \bmod a_l$$

$$= a_1 \sum_{i=1}^w d_{i1} + a_2 \sum_{i=1}^w d_{i2} + \dots + a_{l-1} \sum_{i=1}^w d_{i(l-1)}$$

$$\frac{M - (M \bmod a_l)}{a_l} = \sum_{i=1}^w d_{il}$$

Security Analysis

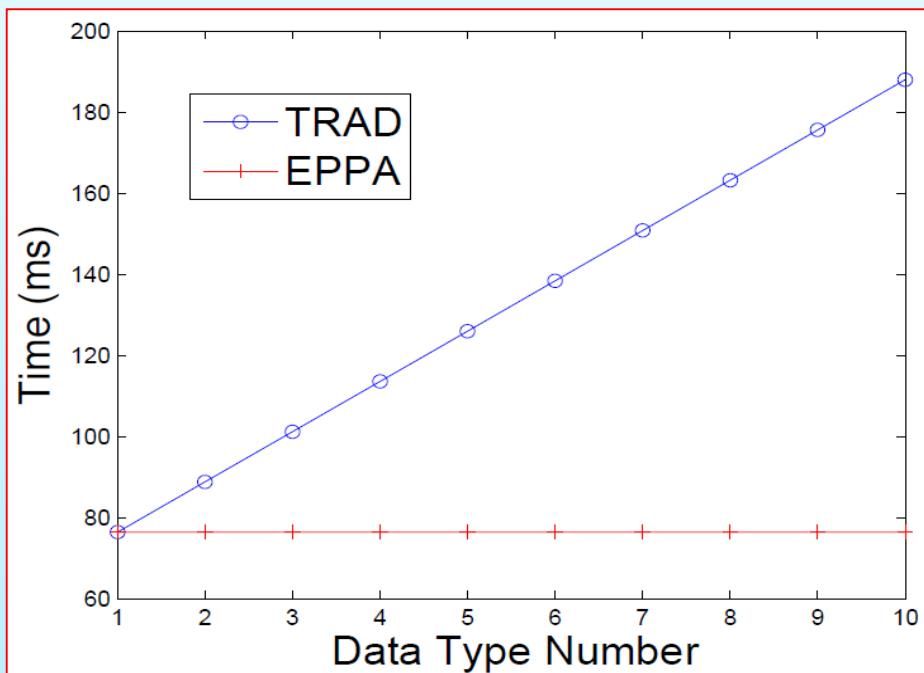


Secure against Malware Eavesdropping at Center

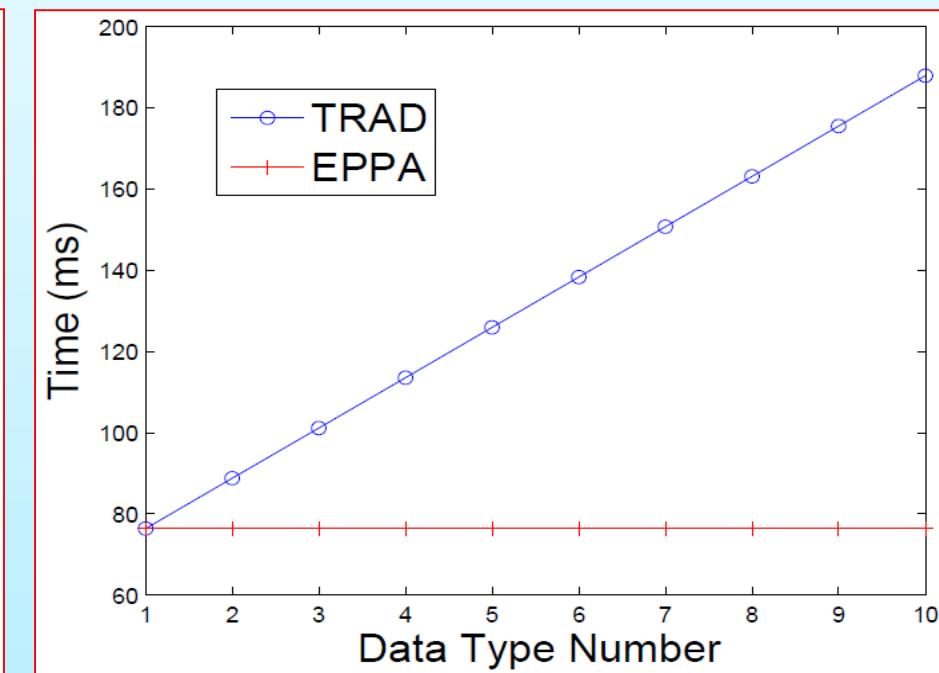
Secure against Malware Eavesdropping at GW

Efficiency Discussion -- Computation

Proposed EPPA & TRAD (traditional one-dimensional aggregation)



Computation Cost of Each HAN



Computation Cost of Control Center

EPPA

$$C = g^{\sum_{i=1}^w d_{i1} + a_2 \sum_{i=1}^w d_{i2} + \dots + a_l \sum_{i=1}^w d_{il}} \cdot \left(\prod_{i=1}^w r_i \right)^n \text{ mod } n^2$$

TRAD

$$C_1 = g^{\sum_{i=1}^w d_{i1}} \cdot \left(\prod_{i=1}^w r_i \right)^n \text{ mod } n^2, C_2, \dots, C_l$$

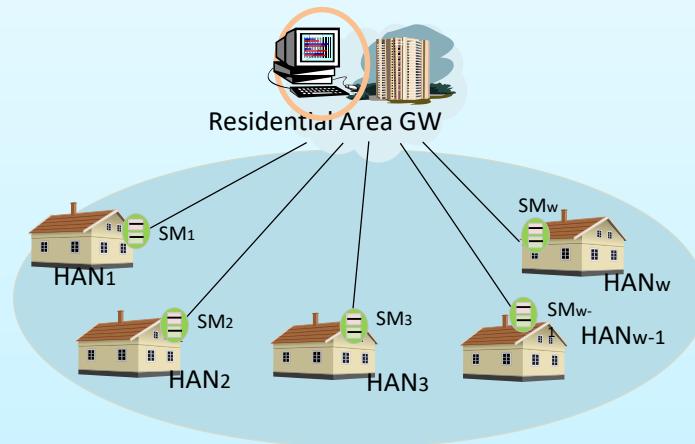
EPPA $c_i = E(m_i) = E(d_{i1}, d_{i2}, \dots, d_{il}) = g^{d_{i1}} \cdot g^{d_{i2}} \cdots g^{d_{il}} \cdot r_i^n \text{ mod } n^2$

TRAD $c_{i1} = g^{d_{i1}} \cdot r_{i1}^n \text{ mod } n^2$ ----- $c_{il} = g^{d_{il}} \cdot r_{il}^n \text{ mod } n^2$

Efficiency Discussion -- Communication

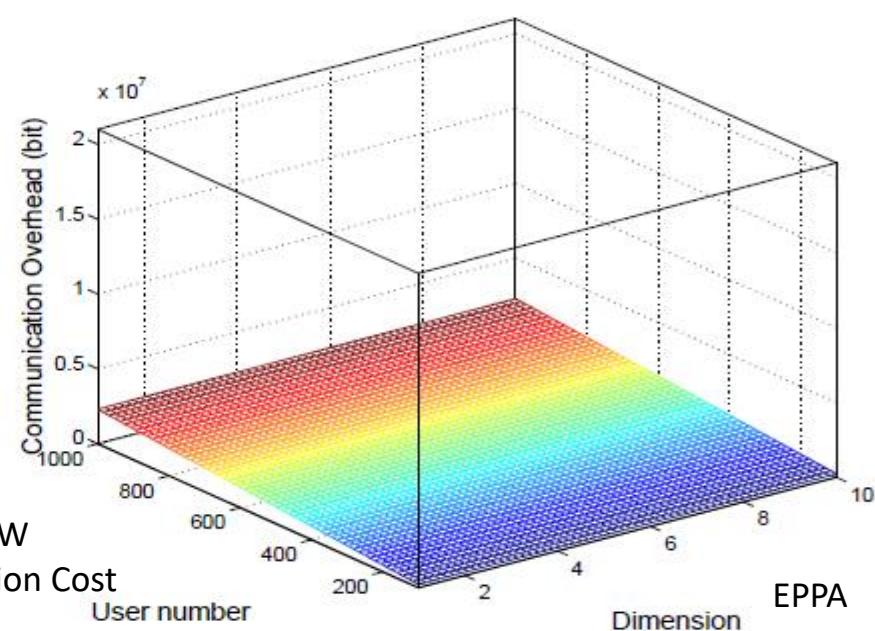
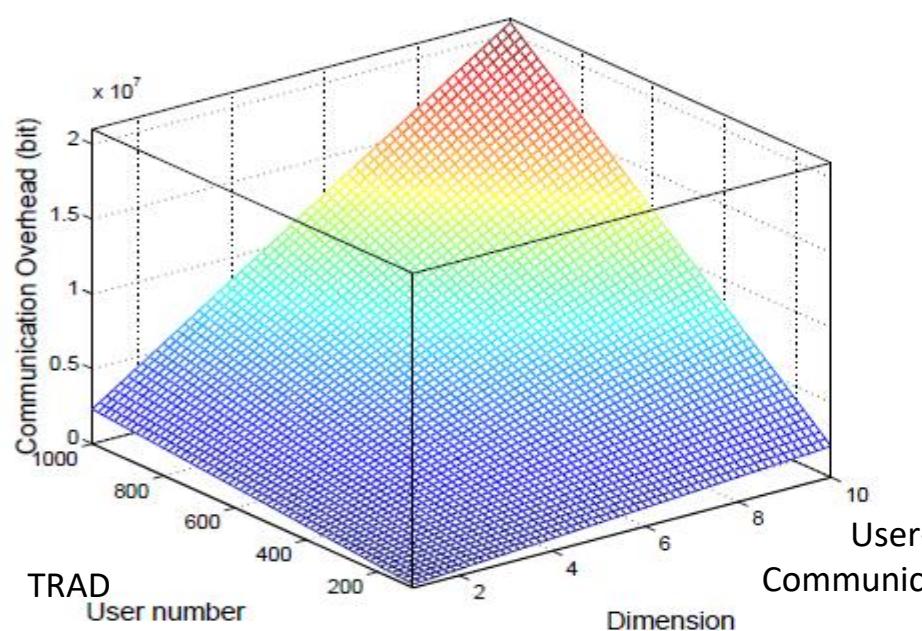
$$c_{i1} = g^{d_{i1}} \cdot r_{i1}^n \mod n^2$$

$$c_{il} = g^{d_{il}} \cdot r_{il}^n \mod n^2$$

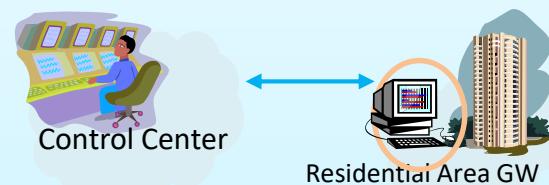


$$c_i = E(m_i) = E(d_{i1}, d_{i2}, \dots, d_{il})$$

$$= g_1^{d_{i1}} \cdot g_2^{d_{i2}} \cdots g_l^{d_{il}} \cdot r_i^n \mod n^2$$

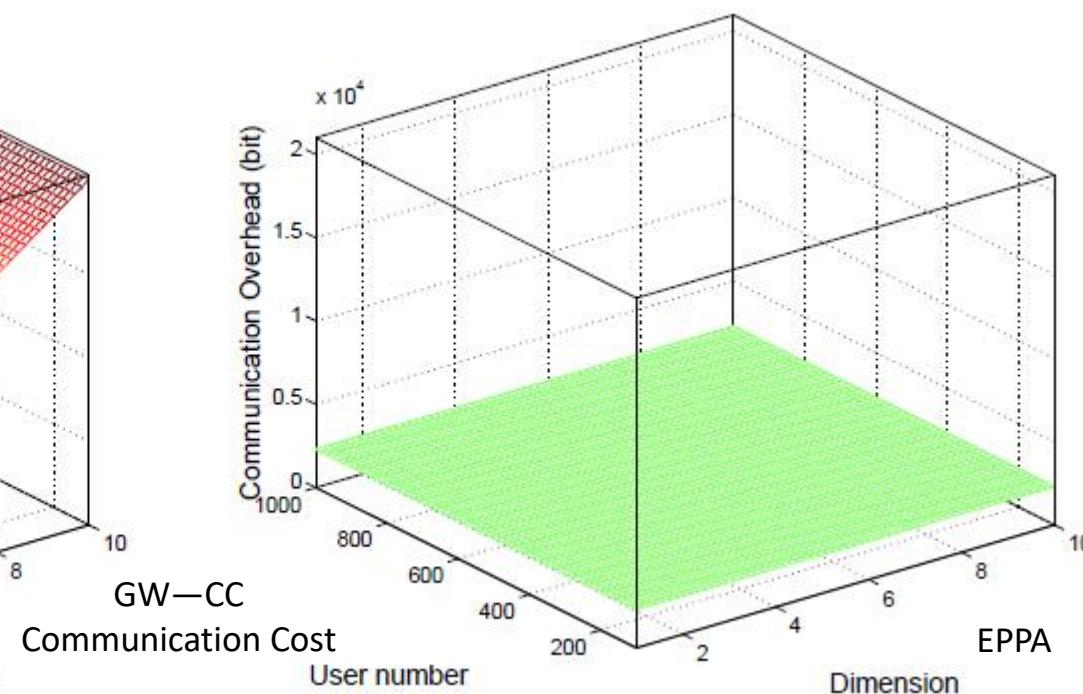
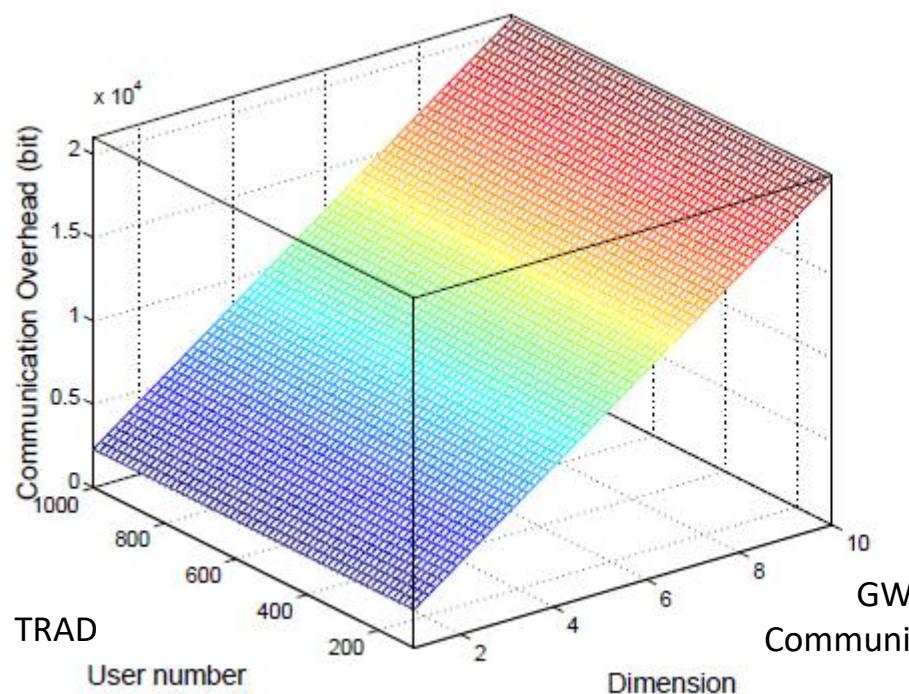


Efficiency Discussion – Communication (2)



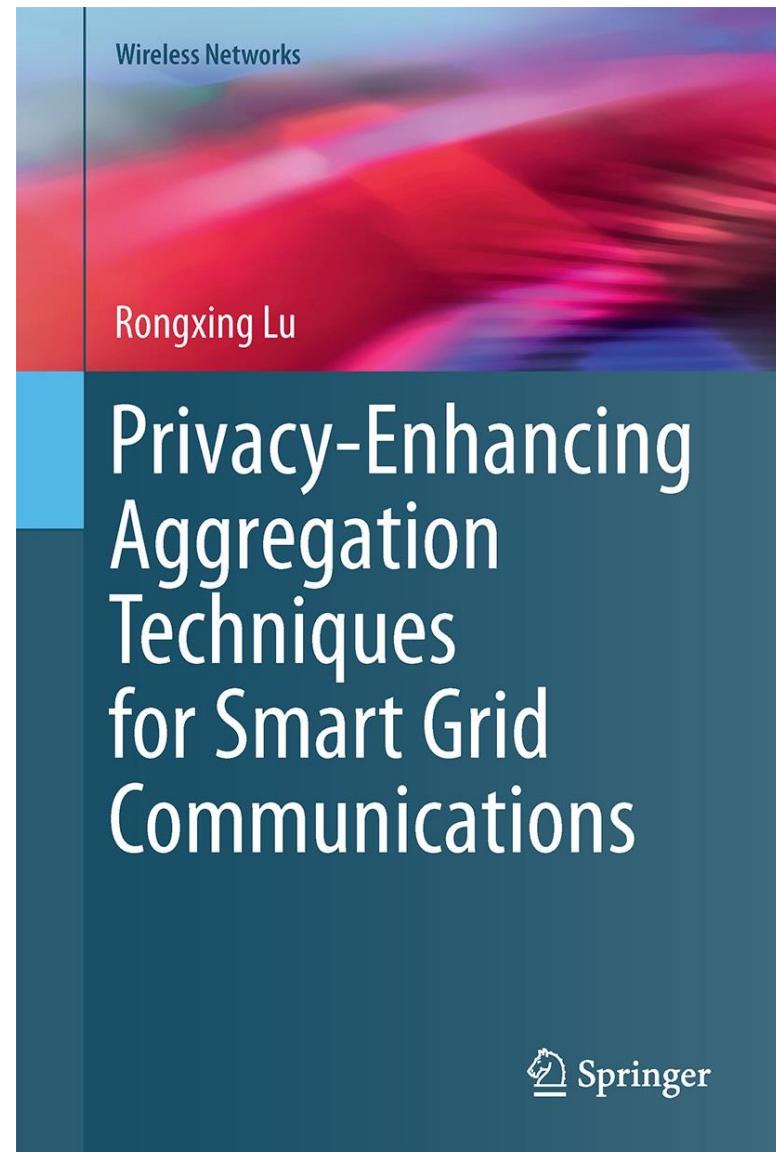
$$C_1 = g^{\sum_{i=1}^w d_{i1}} \cdot \left(\prod_{i=1}^w r_i \right)^n \bmod n^2, C_2, \dots, C_l$$

$$C = g^{a_1 \sum_{i=1}^w d_{i1} + a_2 \sum_{i=1}^w d_{i2} + \dots + a_l \sum_{i=1}^w d_{il}} \cdot \left(\prod_{i=1}^w r_i \right)^n \bmod n^2$$



More References

- Bibliography on Secure Smart Grid Communications
- <http://bbcr.uwaterloo.ca/~rxlu/secureradgridbib>



Thank
you

