CPT-S 415

Big Data

Yinghui Wu

EME B45

CPT_S 415
Big Data

Data security and privacy

- Information Security: basic concepts
- Privacy: basic concepts and comparison with security

Information security

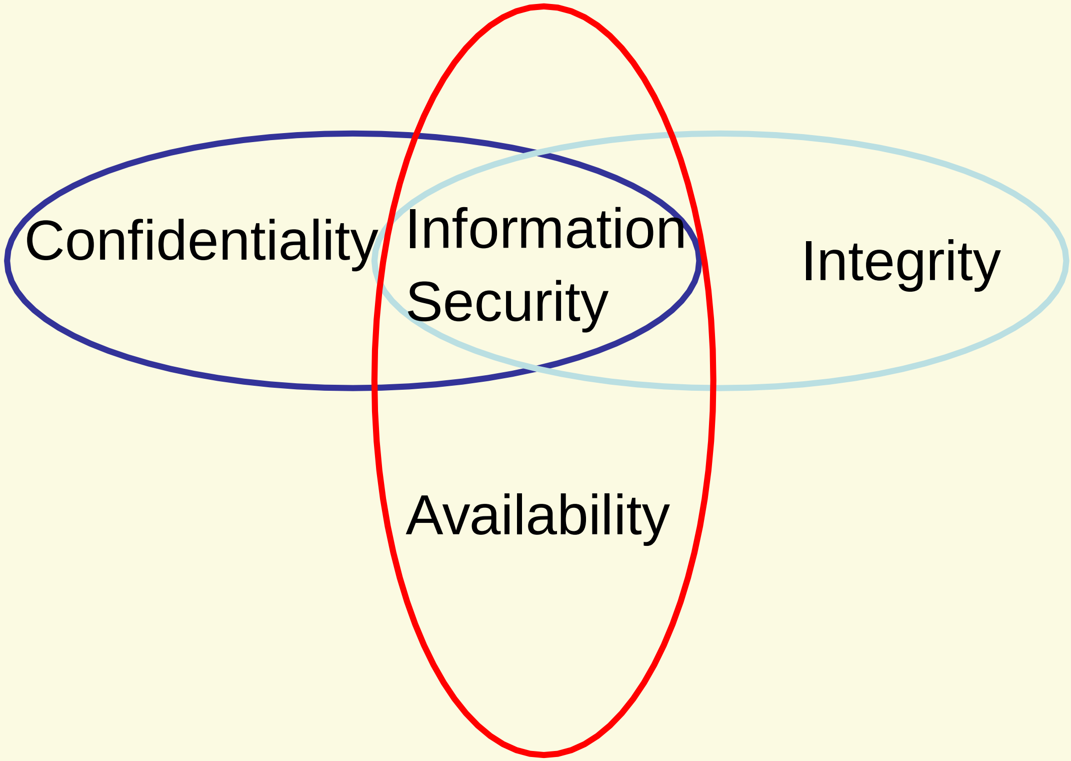
Information Protection - Why?

- Information: important strategic and operational asset for any organization
- Damages and misuses of information affect not only a single user or an application; they may have disastrous consequences on the entire organization
- Additionally, the advent of the Internet as well as networking capabilities has made the access to information much easier

Information Security: Examples

- Consider a payroll database in a corporation, it must be ensured that:
  - salaries of individual employees are not disclosed to arbitrary users of the database
  - salaries are modified by only those individuals that are properly authorized
  - paychecks are printed on time at the end of each pay period
- In a military environment, it is critical that:
  - the target of a missile is not given to an unauthorized user
  - the target is not arbitrarily modified
  - the missile is launched when it is fired

Information Security: Main Requirements



Confidentiality

Information Security

Integrity

Availability

Information Security - main requirements

- *Confidentiality* - information protection from unauthorized read operations

  – the term *privacy* is often used when data to be protected refer to individuals

- *Integrity* - information protection from modifications; it involves several goals:
  - Assuring the integrity of information with respect to the original information (relevant especially in web environment) – often referred to as *authenticity*
  - Protecting information from unauthorized modifications
  - Protecting information from incorrect modifications – referred to as *semantic integrity*

- *Availability* - it ensures that access to information is not denied to authorized subjects

Information Security – additional requirements

- *Information Quality* – not considered traditionally as part of information security but it is very relevant

- *Completeness* – it refers to ensure that subjects receive all information they are entitled to access, according to the stated security policies

Classes of Threats

- Disclosure – unauthorized access to information
  - Snooping, Trojan Horses
    - counted by Confidentiality & Integrity
- Deception – acceptance of false data
  - Modification, spoofing, repudiation of origin, denial of receipt
    - counted by Integrity & availability
- Disruption - interruption or prevention of correct operation
  - Modification
    - counted by Integrity & availability
- Usurpation - unauthorized control of some part of a system
  - Spoofing, Delay, denial of service
    - counted by Integrity & availability

Goals of Security

- Prevention
  - – Prevent attackers from violating security policy
- Detection
  - – Detect attackers’ violation of security policy
- Recovery
  - – Stop attack, assess and repair damage
  - – Continue to function correctly even if attack succeeds

Information Security – How?

- Information must be protected at various levels:
  - The operating system
  - The network
  - The data management system
  - Physical protection is also important

Information Security – Mechanisms

- Confidentiality is enforced by the access control mechanism

- Integrity is enforced by the access control mechanism and by the semantic integrity constraints (domain, type, constraints)

- Availability is enforced by the recovery mechanism and by detection techniques for DoS attacks – an example of which is query flood

Information Security – How?
Additional mechanisms

- *User authentication* - to verify the identity of subjects wishing to access the information

- *Information authentication* - to ensure information authenticity - it is supported by signature mechanisms (e.g., RSA signature keys)
- *Encryption* - to protect information when being transmitted across systems and when being stored on secondary storage
- *Intrusion detection* – to protect against impersonation of legitimate users and also against insider threats

Access Control

- ✓ Problem Statement: authorizing data access scopes (relations, attributes, tuples) to users of DBMS
- ✓ Discretionary access control
  - – Authorization administration policies, ie, granting and revoking authorization (centralized, ownership, etc)
  - – Content-based using views and rewriting for fine-grained access control
  - – Role-based access control: a function with a set of actions, consisting of users members
- ✓ Mandatory access control:
  - – Object and subject classification (eg, top secret, secret, unclassified, etc).

Information Security: A Complete Solution

- It consists of:
  - first defining a *security policy*
  - then choosing some *mechanism* to enforce the policy
  - finally providing *assurance* that both the mechanism and the policy are sound

↓

SECURITY LIFE-CYCLE

Data Privacy

Big data: Is Our Security Keeping Pace?

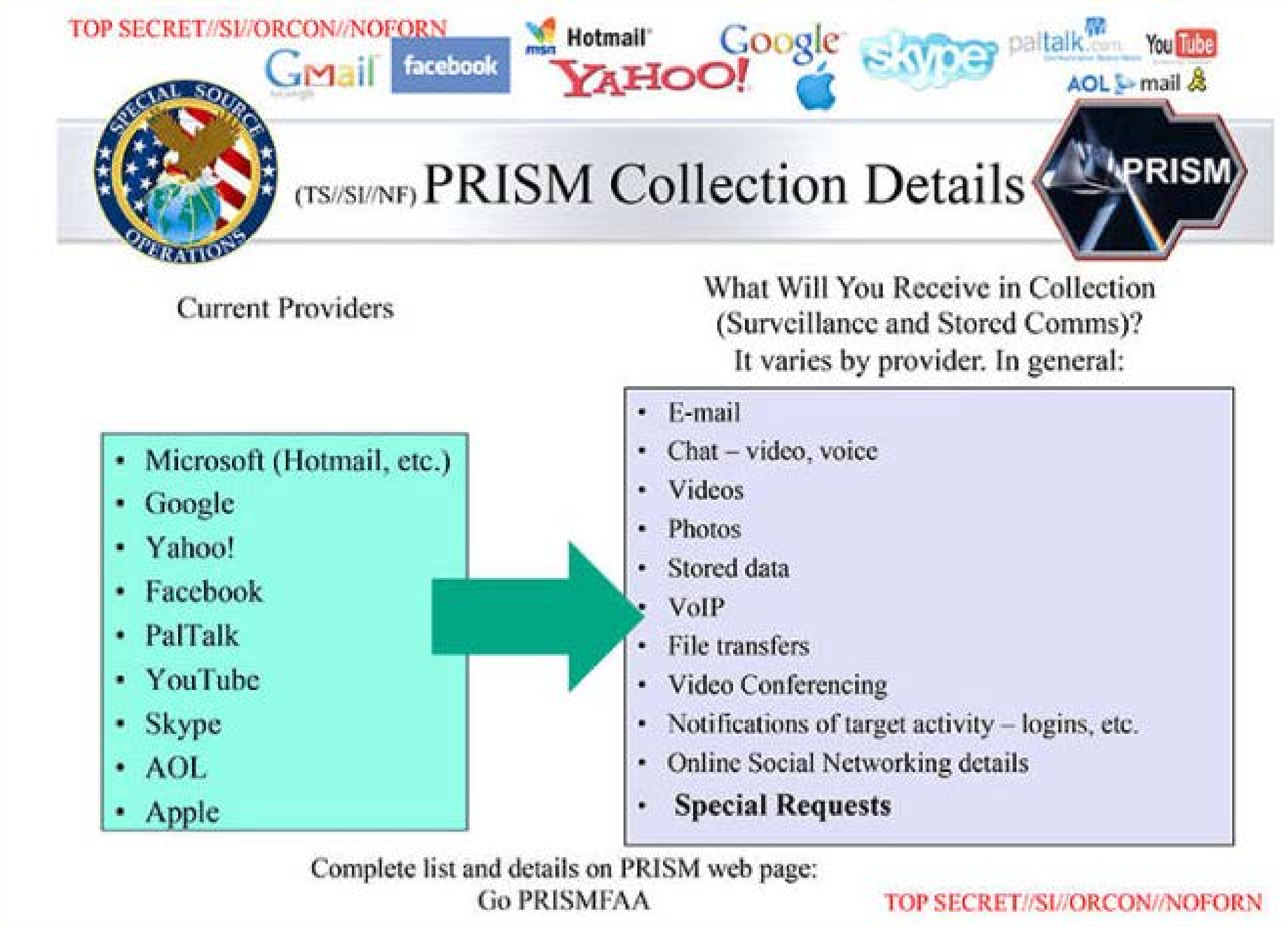
- ✓ Example:

Edward Snowden

He worked for the CIA and then NSA and leaked thousands of classified documents to media outlets.

The documents showed details of a global surveillance program, especially the mass collection of phone data.

PRISM project


Gmail
facebook
msn Hotmail
YAHOO!
Google
skype
paltalk.com
YouTube
AOL mail
SPECIAL SOURCE OPERATIONS
PRISM

TOP SECRET//SI//ORCON//NOFORN

(TS//SI//NF) PRISM Collection Details

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

What Will You Receive in Collection (Surveillance and Stored Comms)?
It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:
Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN

Need for Privacy Guarantees

- ✓ By individuals [Cran *et al.* '99]
  - 99% unwilling to reveal their SSN
  - 18% unwilling to reveal their… favorite TV show
- ✓ By businesses
  - Online consumers worrying about revealing personal data
    held back $15 billion in online revenue in 2001
- ✓ By Federal government
  - Privacy Act of 1974 for Federal agencies
  - Health Insurance Portability and Accountability Act of 1996 (HIPAA)

Need for Privacy Guarantees

✓ By computer industry research (examples)

- Microsoft Research
  - The biggest research challenges:
    - Reliability / Security / Privacy / Business Integrity

    => MS Trustworthy Computing Initiative
  - Topics include: DRM—digital rights management (incl. watermarking surviving photo editing attacks), software rights protection, intellectual property and content protection, database privacy and p.-p. data mining, anonymous e-cash, anti-spyware
- IBM
  - Topics include: pseudonymity for e-commerce, EPA and EPAL—enterprise privacy architecture and language, RFID privacy, p.-p. video surveillance, federated identity management (for enterprise federations), p.-p. data mining and p.-p.mining of association rules, hippocratic (p.-p.) databases, online privacy monitoring

Need for Privacy Guarantees

- ✓ By academic researchers (examples from the U.S.A.)
  - CMU and Privacy Technology Center
    - Latanya Sweeney (k-anonymity, SOS—Surveillance of Surveillances, genomic privacy)
    - Mike Reiter (Crowds – anonymity)
  - Purdue University – CS and CERIAS
    - Elisa Bertino (trust negotiation languages and privacy)
    - Bharat Bhargava (privacy-trust tradeoff, privacy metrics, p.-p. data dissemination, p.-p. location-based routing and services in networks)
    - Chris Clifton (p.-p. data mining)
    - Leszek Lilien (p.-p. data disemination)
  - UIUC
    - Roy Campbell (Mist – preserving location privacy in pervasive computing)
    - Marianne Winslett (trust negotiation w/ controled release of private credentials)
  - U. of North Carolina Charlotte
    - Xintao Wu, Yongge Wang, Yuliang Zheng (p.-p. database testing and data mining)

Definition

- **Privacy** is the ability of a person to control the availability of information about and exposure of him- or herself. It is related to being able to function in society anonymously.

- **Types of privacy** giving raise to special concerns:
  - Political privacy
  - Consumer privacy
  - Medical privacy
  - *Information technology end-user privacy; also called data privacy*
  - Private property

Data Privacy

- Data Privacy problems exist *wherever uniquely identifiable data relating to a person or persons are collected and stored, in digital form or otherwise*.

- Improper or non-existent disclosure control can be the root cause for privacy issues.

- The most common sources of data affected by data privacy issues:
  - *Health information*
  - *Criminal justice*
  - *Financial information*
  - *Genetic information*

Data Privacy

- The challenge in data privacy is to share data while protecting the personally identifiable information.
  - Consider the example of health data which are collected from hospitals in a district; it is standard practice to share this only in aggregate form
  - The idea of sharing the data in aggregate form is to ensure that only non-identifiable data are shared.

- The legal protection of the right to privacy in general and of data privacy in particular varies greatly around the world.

Technologies with Privacy Concerns

- Biometrics (DNA, fingerprints, iris) and face recognition
- Video surveillance, ubiquitous networks and sensors
- Cellular phones
- Personal Robots
- DNA sequences, Genomic Data

Threats to Privacy [cf. Simone Fischer-Hübner]

1) Threats to privacy at application level

- Threats to collection / transmission of large quantities of personal data
  - Health Networks / Public administration Networks
  - Research Networks / Electronic Commerce / Teleworking
  - Distance Learning / Private use

  - Example: Information infrastructure for a better healthcare
    [cf. Danish "INFO-Society 2000"- or Bangemann-Report]
    - National and European healthcare networks for the interchange of information
    - Interchange of (standardized) electronic patient case files
    - Systems for tele-diagnosing and clinical treatment

Threat to Privacy

[cf. Simone Fischer-Hübner]

2) Threats to privacy at communication level

- ✓ Threats to anonymity of sender / forwarder / receiver
- ✓ Threats to anonymity of service provider
- ✓ Threats to privacy of communication
  - – E.g., via monitoring / logging of transactional data
    - Extraction of user profiles & its long-term storage

3) Threats to privacy at system level

- ✓ E.g., threats at system access level

4) Threats to privacy in audit trails

Threat to Privacy

[cf. Simone Fischer-Hübner]

- ✓ Identity theft – the most serious crime against privacy
- ✓ Threats to privacy – another view
  - – Aggregation and data mining
  - – Poor system security
  - – Government threats
    - Gov't has a lot of people's most private data
      - – Taxes / homeland security / etc.
    - People's privacy vs. homeland security concerns

Copyright 2006 by Randy Glasbergen.
www.glasbergen.com

"The identity I stole was a fake!
Boy, you just can't trust people these days!"

  - – The Internet as privacy threat
    - Unencrypted e-mail / web surfing / attacks
  - – Corporate rights and private business
    - Companies may collect data that U.S. gov't is *not* allowed to
  - – Privacy for sale - many traps
    - "Free" is not free…
      - – E.g., accepting frequent-buyer cards reduces your privacy

Approaches in Privacy-Preserving Information Management

- Anonymization Techniques
  - Have been investigated in the areas of networks (see “the Anonymity Terminology” by Andreas Pfitzman) and databases (see the notion of “k-anonymity” by L. Sweeney)
- Privacy-Preserving Data Mining
- P3P policies (platform for privacy preference)
  - Are tailored to the specification of privacy practices by organizations and to the specification user privacy preferences
- Hippocratic Databases *(Rakesh.A et al, VLDB 02)*
  - Are tailored to support privacy policies
- Fine-Grained Access Control Techniques
- Private Information Retrieval Techniques

Privacy vs Security

- Privacy is not just confidentiality and integrity of user data
- Privacy includes other requirements:
  - Support for user preferences
  - Support for obligation execution
  - Usability
  - Proof of compliance

Data Anonymization

Inference - Example

Name	Sex	Programme	Units	Grade Ave
Alma	F	MBA	8	63
Bill	M	CS	15	58
Carol	F	CS	16	70
Don	M	MIS	22	75
Errol	M	CS	8	66
Flora	F	MIS	16	81
Gala	F	MBA	23	68
Homer	M	CS	7	50
Igor	M	MIS	21	70

Policy: average grade of a single student cannot be disclosed

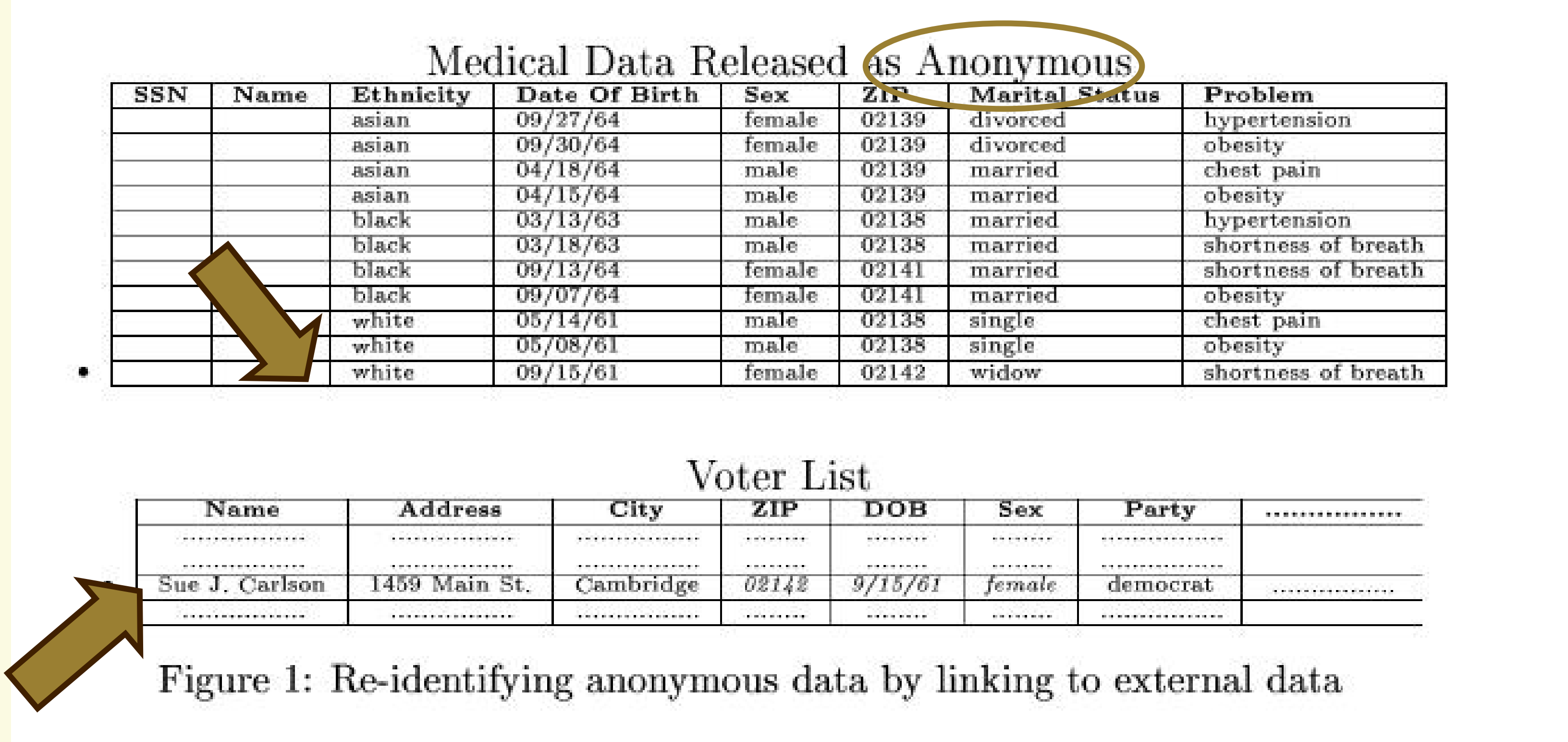
Inference - Example

- However, statistical summaries can be disclosed
- Suppose that an attacker knows that Carol is a female CS student
- By combining the results of the following legitimate queries:
  - Q1: SELECT Count (*) FROM Students WHERE Sex =‘F’ AND Programme = ‘CS’
  - Q2: SELECT Avg (Grade Ave) FROM Students WHERE Sex =‘F’ AND Programme = ‘CS’

The attacker learns from Q1 that there is only one female student so the value 70 returned by Q2 is precisely her average grade

Latanya Sweeney’s Attack (1997)

Massachusetts hospital discharge dataset



Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset

Data Anonymization

- ✓ Problem: protecting Personally Identifiable Information (PII) and their sensitive attributes

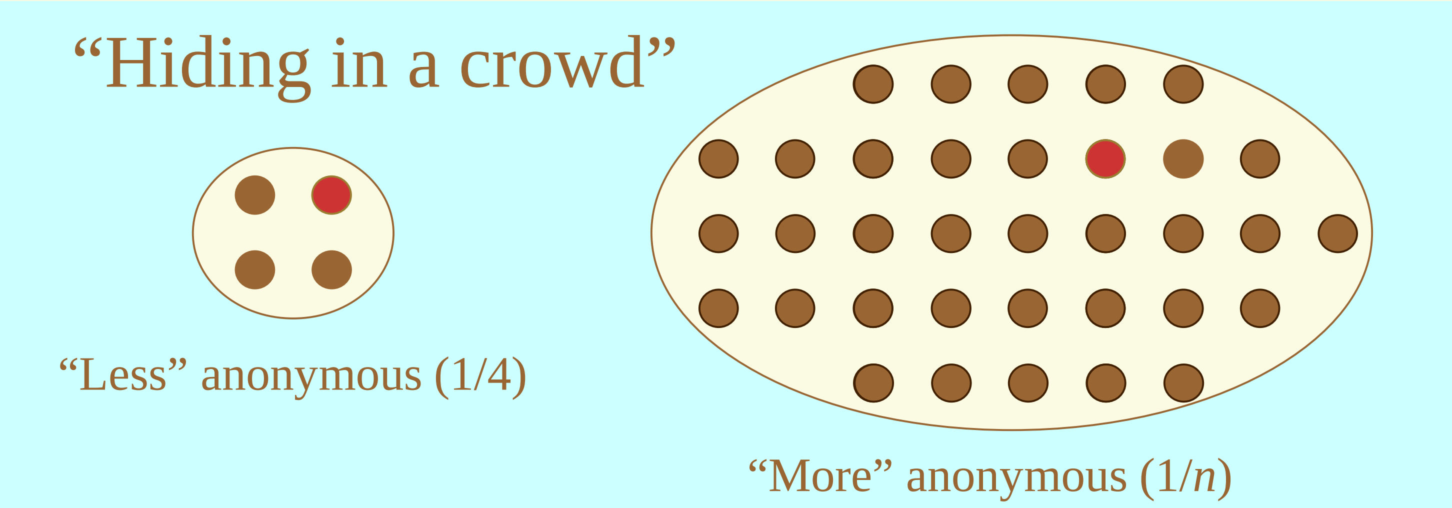
Quasi-identifier			Sensitive
DOB	*Gender*	*Zipcode*	*Disease*
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

Quasi-identifiers need to be generalized or suppressed

Quasi-identifiers are sets of attributes that can be linked with external data to uniquely identify an individual

Anonymity Set Size Metrics

- ✓ The larger set of indistinguishable entities, the lower probability of identifying any one of them
  - Can use to ”anonymize” a selected private attribute value within the domain of its all possible values



Anonymity Set

- Anonymity set A

  A = {(s₁, p₁), (s₂, p₂), …, (sₙ, pₙ)}

  - s_i: subject *i* who might access private data

    or: *i*-th possible value for a private data attribute
  - p_i: probability that s_i accessed private data

    or: probability that the attribute assumes the *i*-th possible value

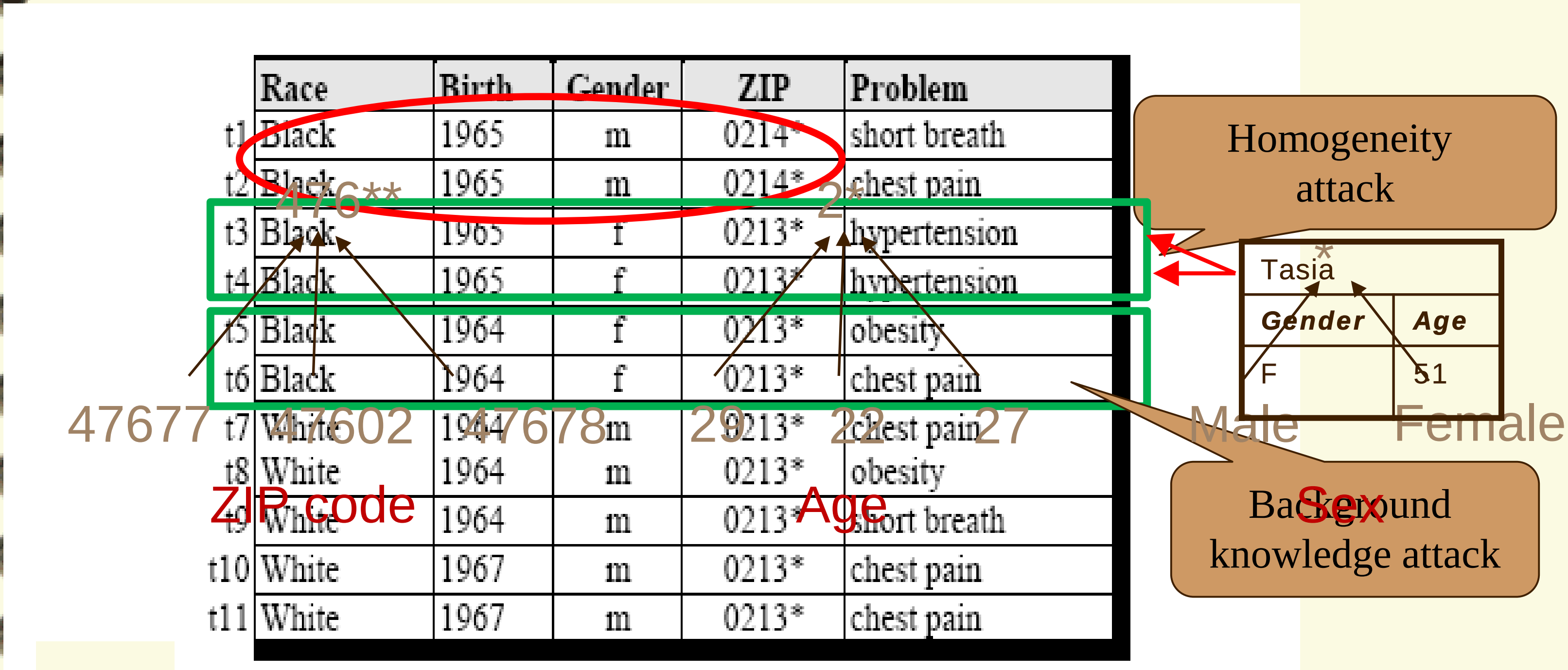
- Effective anonymity set size is

$$L = |A| \sum_{i=1}^{|A|} \min(p_i, 1/|A|)$$

  - Maximum value of L is |A| iff all p_i’s are equal to 1/|A|
  - L below maximum when distribution is skewed
    - skewed when p_i’s have different values

Solution: k-Anonymity [Samarati et al. TR'98]

- ✓ Quasi-identifiers indistinguishable among k individuals
- ✓ Implemented by building generalization hierarchy or partitioning multi-dimensional data space



	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Homogeneity attack

Background knowledge attack

Tasia	
Gender	Age
F	51

Example of *k*-anonymity, where *k*=2 and QI={*Race, Birth, Gender, ZIP*}

Enhanced Solution: l-Diversity [Machanavajjhala et al. ICDE’06]

- At least l values for sensitive attributes in each equivalence class

Bob	
Zip	***Age***
47678	27

Similarity attack

A 3-diverse patient table

Zipcode	*Age*	*Salary*	***Disease***
476**	2*	20K	Gastric Ulcer
476**	2*	25K	Gastritis
476**	2*	30K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Skewness attack

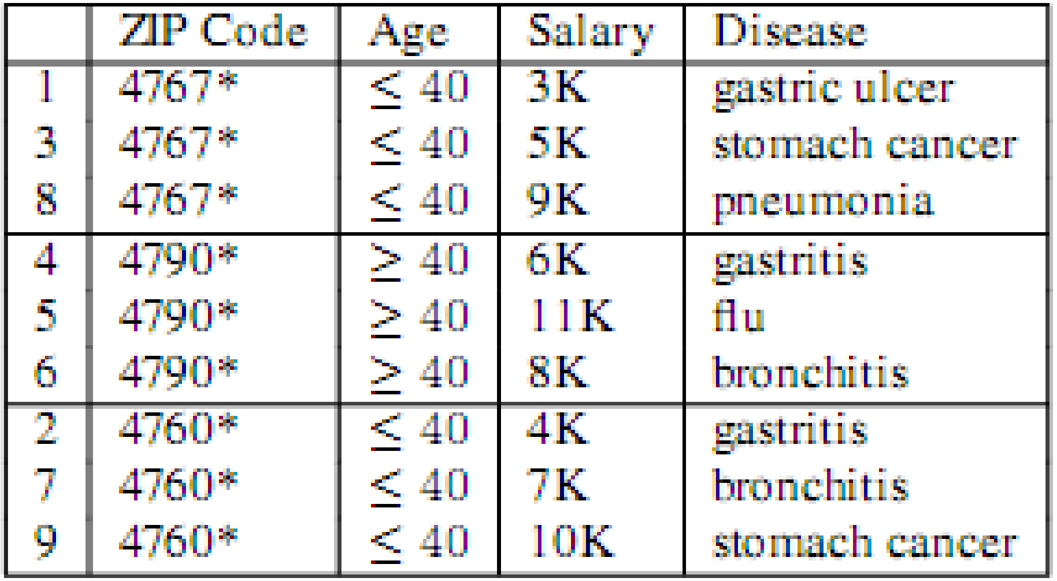
Conclusion

- Bob’s salary is in [20k,40k], which is relatively low
- Bob has some stomach-related disease

Enhanced Solution: t-Closeness [Li et al. ICDE’07]

“Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database” ->

- Distance between overall distribution of sensitive attribute values and distribution of sensitive attribute values in an equivalence class bounded by t



	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

What Does Attacker Know?

Bob is Caucasian and I heard he was admitted to hospital with flu…

This is against the rules!
“flu” is not a quasi-identifier

Yes… and this is yet another problem with k-anonymity

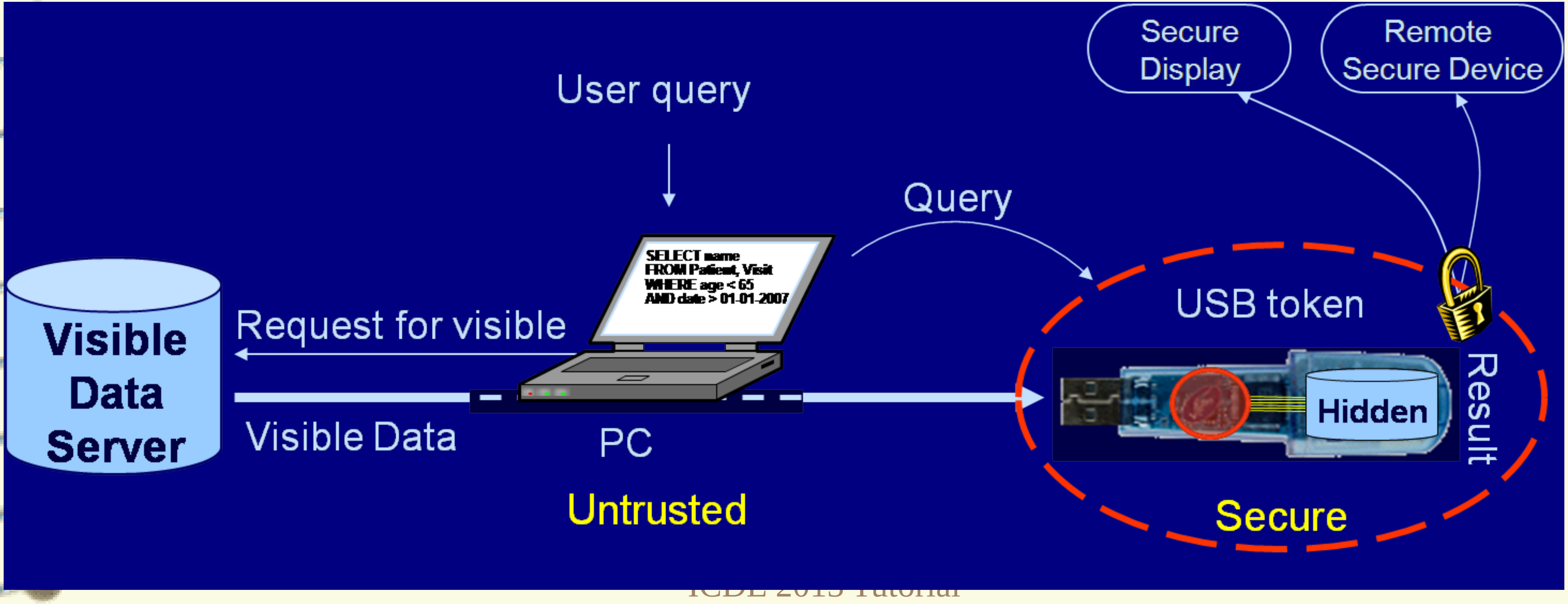
Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	[illegible]	Flu
[illegible]	[illegible]	HIV+	Shingles
[illegible]	[illegible]	[illegible]	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

k-Anonymity: limitations

- ✓ Syntactic
  - – Focuses on data transformation, not on what can be learned from the anonymized dataset
  - – “k-anonymous” dataset can leak sensitive information
- ✓ “Quasi-identifier” fallacy
  - – Assumes a priori that attacker will not know certain information about his target
- ✓ Relies on locality
  - – Possible harmful to the utility of many real-world datasets

Secure Devices for Privacy [Anciaux et al. SIGMOD'07]

- Problem: protecting private data during queries involving both private (hidden) and public (visible) data
- Solution: carry private data in a secure USB key, ensure private data never leaves the USB key, and only public data flows to the key
- Query optimization for small RAM USB key



```
SELECT name
FROM Patient, Visit
WHERE age < 65
AND date > 01-01-2007
```

Data privacy in future

Privacy in Pervasive Computing

- ✓ In pervasive computing environments, socially-based paradigms (incl. trust) will play a big role
- ✓ People surrounded by zillions of computing devices of all kinds, sizes, and aptitudes [“Sensor Nation: Special Report,” *IEEE Spectrum*, vol. 41, no. 7, 2004]
  - Most with limited / rudimentary capabilities
    - Quite small, e.g., RFID tags, smart dust
  - Most embedded in artifacts for everyday use, or even human bodies
    - Possible both beneficial and detrimental (even apocalyptic) consequences
- ✓ Danger of malevolent *opportunistic* sensor networks — pervasive devices self-organizing into huge spy networks
  - Able to spy anywhere, anytime, on everybody and everything
  - Need means of detection & neutralization
    - To tell which and how many snoops are active, what data they collect, and who they work for
      - An advertiser? a nosy neighbor? Big Brother?
    - Questions such as “Can I trust my refrigerator?” will not be jokes
      - The refrigerator snitching on its owner’s dietary misbehavior for her doctor

Using Trust for Privacy Protection

- ✓ Privacy = entity’s ability to control the availability and exposure of information about itself
  - extended the subject of privacy from a person in the original definition [“Internet Security Glossary,” *The Internet Society, Aug. 2004*] to an entity— including an organization or software
    - Important in pervasive computing
- ✓ Privacy and trust are closely related
  - Trust is a socially-based paradigm
  - Privacy-trust tradeoff: Entity can trade privacy for a corresponding gain in its partners’ trust in it
  - The scope of an entity’s privacy disclosure should be proportional to the benefits expected from the interaction
    - As in social interactions
    - E.g.: a customer applying for a mortgage must reveal much more personal data than someone buying a book

Selected Publications

- ✓ “Private and Trusted Interactions,” by B. Bhargava and L. Lilien.
- ✓ “On Security Study of Two Distance Vector Routing Protocols for Mobile Ad Hoc Networks,” by W. Wang, Y. Lu and B. Bhargava, Proc. of IEEE Intl. Conf. on Pervasive Computing and Communications (PerCom 2003), Dallas-Fort Worth, TX, March 2003. http://www.cs.purdue.edu/homes/wangwc/PerCom03wangwc.pdf
- ✓ “Fraud Formalization and Detection,” by B. Bhargava, Y. Zhong and Y. Lu, Proc. of 5th Intl. Conf. on Data Warehousing and Knowledge Discovery (DaWaK 2003), Prague, Czech Republic, September 2003. http://www.cs.purdue.edu/homes/zhong/papers/fraud.pdf
- ✓ “Trust, Privacy, and Security. Summary of a Workshop Breakout Session at the National Science Foundation Information and Data Management (IDM) Workshop held in Seattle, Washington, September 14 - 16, 2003” by B. Bhargava, C. Farkas, L. Lilien and F. Makedon, CERIAS Tech Report 2003-34, CERIAS, Purdue University, November 2003. http://www2.cs.washington.edu/nsf2003 or https://www.cerias.purdue.edu/tools_and_resources/bibtex_archive/archive/2003-34.pdf
- ✓ “e-Notebook Middleware for Accountability and Reputation Based Trust in Distributed Data Sharing Communities,” by P. Ruth, D. Xu, B. Bhargava and F. Regnier, Proc. of the Second International Conference on Trust Management (iTrust 2004), Oxford, UK, March 2004. http://www.cs.purdue.edu/homes/dxu/pubs/iTrust04.pdf
- ✓ “Position-Based Receiver-Contention Private Communication in Wireless Ad Hoc Networks,” by X. Wu and B. Bhargava, submitted to the Tenth Annual Intl. Conf. on Mobile Computing and Networking (MobiCom’04), Philadelphia, PA, September - October 2004. http://www.cs.purdue.edu/homes/wu/HTML/research.html/paper_purdue/mobi04.pdf

Selected Publications

1. *The American Heritage Dictionary of the English Language*, 4th ed., Houghton Mifflin, 2000.
2. B. Bhargava et al., *Trust, Privacy, and Security: Summary of a Workshop Breakout Session at the National Science Foundation Information and Data Management (IDM) Workshop held in Seattle,Washington, Sep. 14–16, 2003*, tech. report 2003-34, Center for Education and Research in Information Assurance and Security, Purdue Univ., Dec. 2003;

   www.cerias.purdue.edu/tools_and_resources/bibtex_archive/archive/2003-34.pdf.
3. “Internet Security Glossary,” *The Internet Society*, Aug. 2004; www.faqs.org/rfcs/rfc2828.html.
4. B. Bhargava and L. Lilien “Private and Trusted Collaborations,” to appear in *Secure Knowledge Management* (SKM 2004)*: A Workshop*, 2004.
5. “Sensor Nation: Special Report,” *IEEE Spectrum*, vol. 41, no. 7, 2004.
6. R. Khare and A. Rifkin, “Trust Management on the World Wide Web,” *First Monday*, vol. 3, no. 6, 1998; www.firstmonday.dk/issues/issue3_6/khare.
7. M. Richardson, R. Agrawal, and P. Domingos,“Trust Management for the Semantic Web,” *Proc. 2nd Int’l Semantic Web Conf.*, LNCS 2870, Springer-Verlag, 2003, pp. 351–368.
8. P. Schiegg et al., “Supply Chain Management Systems—A Survey of the State of the Art,” *Collaborative Systems for Production Management: Proc. 8th Int’l Conf. Advances in Production Management Systems* (APMS 2002), IFIP Conf. Proc. 257, Kluwer, 2002.
9. N.C. Romano Jr. and J. Fjermestad, “Electronic Commerce Customer Relationship Management: A Research Agenda,” *Information Technology and Management*, vol. 4, nos. 2–3, 2003, pp. 233–258.

Course evaluation

✓ For students:

• Direct to myWSU portal. The center of the page includes a BLUE COURSE EVALUATIONS window.

• Please complete the evaluations – thanks!