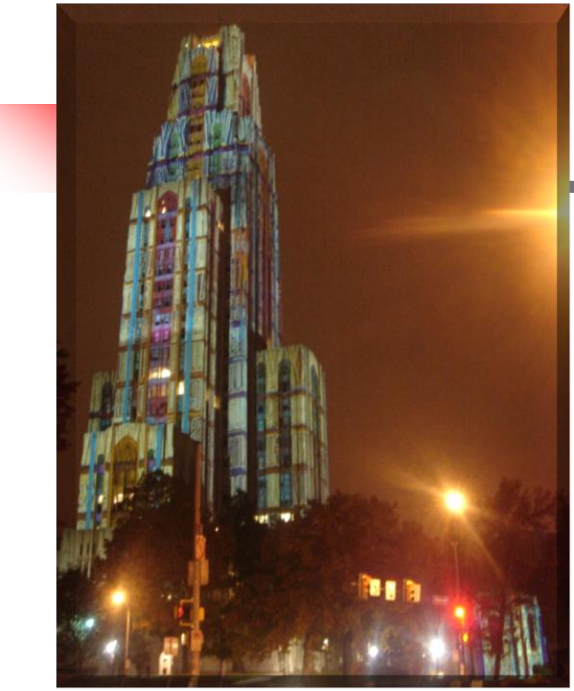
IS 2150 / TEL 2810
Information Security & Privacy



James Joshi
Associate Professor, SIS

Lecture 11
April 10, 2013

Information Privacy

(Contributed by Amirreza Masoumzadeh)

What is privacy?

- Hard to define
- “Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others”
  - Alan Westin, Privacy and Freedom, 1967

OECD Guidelines on the Protection of Privacy (1980)

- Collection limitation
- Data quality
- Purpose specification
- Use limitation
- Security safeguards
- Openness
- Individual participation
- Accountability

http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html#part2

FTC Fair Information Practice Principles

- Notice/Awareness
- Choice/Consent
- Access/Participation
- Integrity/Security
- Enforcement/Redress

http://www.ftc.gov/reports/privacy3/fairinfo.shtm

Privacy Laws

- EU: Comprehensive
  - European Directive on Data Protection
- US: Sector specific
  - HIPAA (Health Insurance Portability and Accountability Act of 1996)
    - Protect individually identifiable health information
  - COPPA (Children's Online Privacy Protection Act of 1998)
    - Address collection of personal information from children under 13, how to seek verifiable parental consent from their parents, etc.
  - GLB (Gramm-Leach-Bliley-Act of 1999)
    - Requires financial institutions to provide consumers with a privacy policy notice, including what info collected, where info shared, how info used, how info protected, opt-out options, etc.

Online Privacy Seal Programs (1)

- WebTrust
  - Developed by the American Institute of Certified Public Accountants and the Canadian Institute of Chartered Accountants
  - Privacy standards established by the Online Privacy Alliance, the EU, and Canada with regard to business practices and information privacy, transaction integrity, and security
- TRUSTe
  - Founded by Electronic Frontier Foundation and CommerceNet Consortium, Inc.
  - Adherence to TRUSTe's privacy policies of disclosure, choice, access, and security
  - Ongoing oversight and alternative dispute resolution processes

http://www.uschamber.com/issues/technology/online-privacy-seal-programs

Online Privacy Seal Programs (2)

- BBBOnLine
  - Developed by the Council of Better Business Bureaus
  - Features verification, monitoring and review, consumer dispute resolution, enforcement mechanisms, and an educational component
- The Platform for Privacy Preferences (P3P)
  - Developed by W3C
  - Enables Websites to express their privacy practices in a standard format that can be retrieved automatically and interpreted easily by user agents

DATA ANONYMIZATION

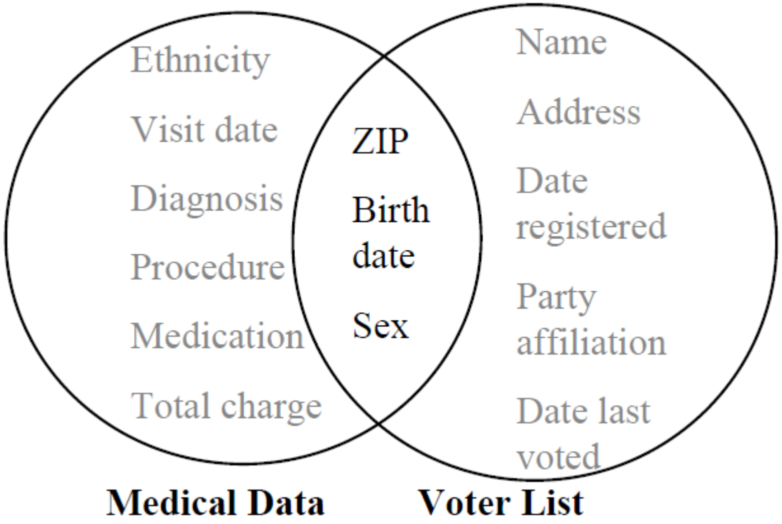
Some slides barrowed from Vitaly Shmatikov

Data Collection & Publishing

- Health-care datasets
  - Clinical studies, hospital discharge databases …
- Genetic datasets
  - 1000 genome, HapMap, deCode …
- Demographic datasets
  - U.S. Census Bureau, sociology studies …
- Search logs, recommender systems, social networks, blogs …
  - AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon …

Linking Attack

- 87% of US population uniquely identifiable by 5-digit ZIP code, gender, DOB [using 1990 US census summary data]
- A practical attack [Sweeney2002]



- Massachusetts governor's hospital record re-identified
  - 6 with same DOB, 3 men, only one with same ZIP code

Quasi-identifier

- Identifier attributes
  - e.g., Name, SSN, address, phone no., etc.
  - A naïve anonymization method will always remove these
- Quasi-identifier attributes
  - 5-digit ZIP code, gender, DOB
  - Combination of attributes that can be used for linking attack
- Other attributes

k-Anonymity

- Each record must be indistinguishable with at least *k*-1 other records with respect to the quasi-identifier
- Linking attack cannot be performed with confidence > 1/*k*
- Formal definition [Samarati2001]
  - Let $T(A_1, \ldots, A_n)$ be a table and *QI* be a quasi-identifier associated with it. *T* is said to satisfy *k*-anonymity wrt *QI* iff each sequence of values in *T[QI]* appears at least with *k* occurrences in *T[QI]*.
    - (*T[QI]* is the projection of *T* on quasi-identifier attributes)

k-Anonymity: Example

- *k=2* and *QI*={*Race, Birth, Gender, ZIP*}

Equivalency Class

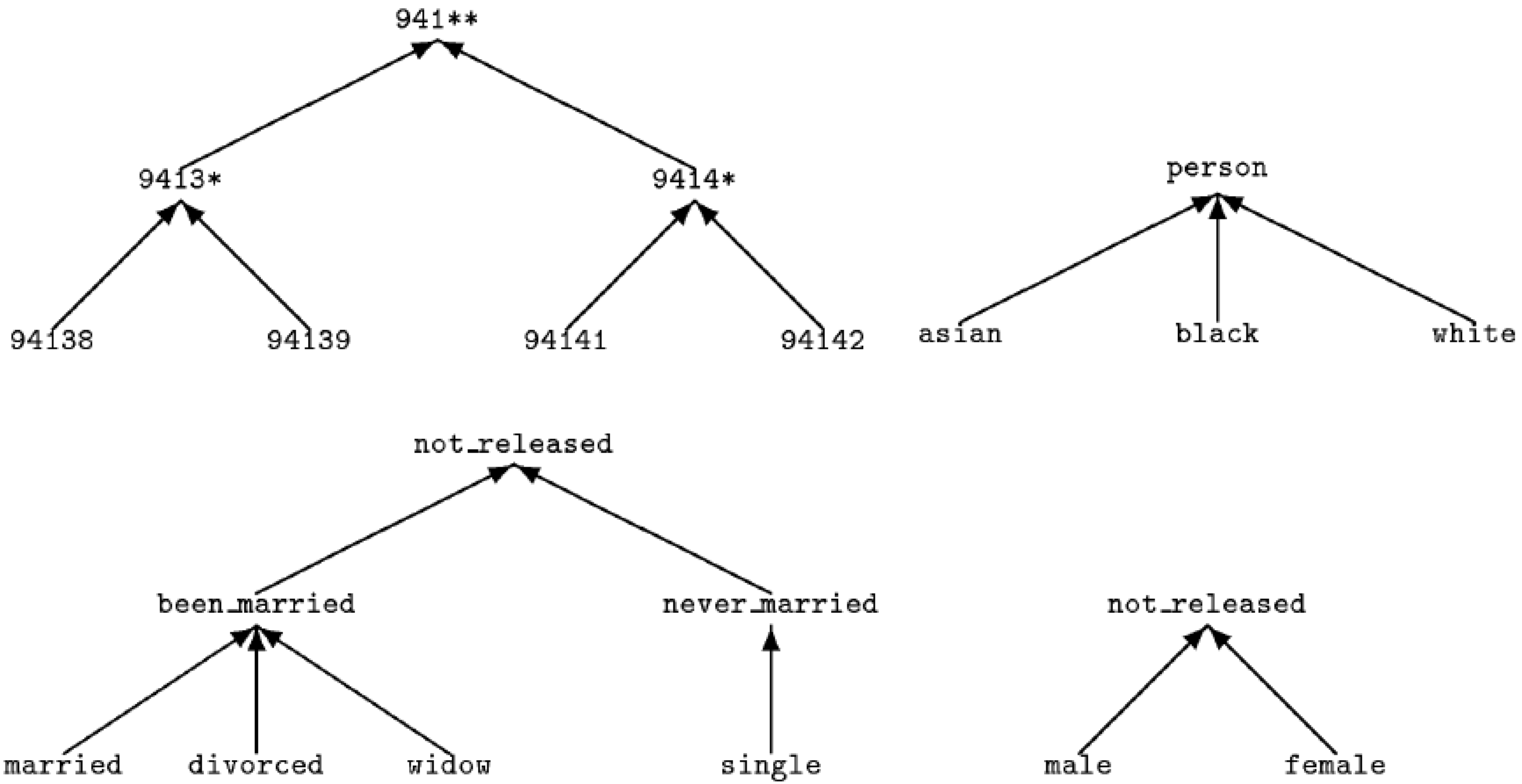
	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White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

Achieving *k*-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get *k* identical values
  - Partition ordered-value domains into intervals
- Suppression
  - Remove some records
  - When generalization causes too much information loss
    - This is common with "outliers"
- Lots of algorithms in the literature
  - Aim to produce "useful" anonymizations

  … usually without any clear notion of utility

Generalization Hierarchy



941**

9413*

9414*

94138

94139

94141

94142

person

asian

black

white

not_released

been_married

never_married

married

divorced

widow

single

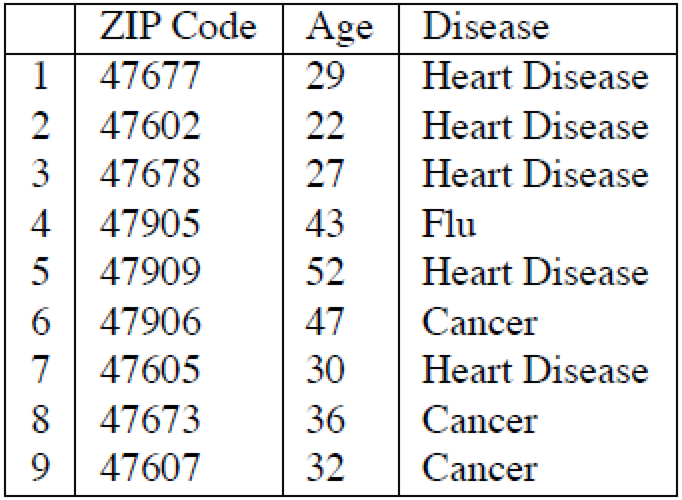
not_released

male

female

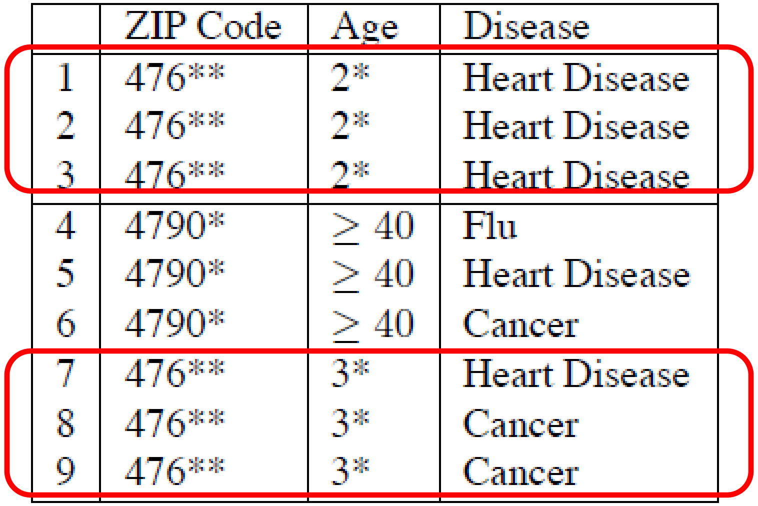
k-Anonymity Is Not Enough

- *k*-anonymity protects against **identity disclosure**, but not **attribute disclosure**!



	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table 1. Original Patients Table



	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table 2. A 3-Anonymous Version of Table 1

- Lack of diversity in sensitive attributes of an equivalency class can reveal sensitive attributes

l-Diversity

- A table is said to have *l*-diversity if every equivalence class of the table has *l*-diversity
  - i.e., there are at least *l* “well-represented” values for the sensitive attribute
- Distinct *l*-diversity
  - Each equivalence class has at least *l* well-represented sensitive values
  - Does not prevent probabilistic inference attacks

…	**Disease**
	…
	HIV
	HIV
	…
	HIV
	pneumonia
	bronchitis
	…

10 records

8 records have HIV

2 records have other values

l-Diversity: Skewness Attack

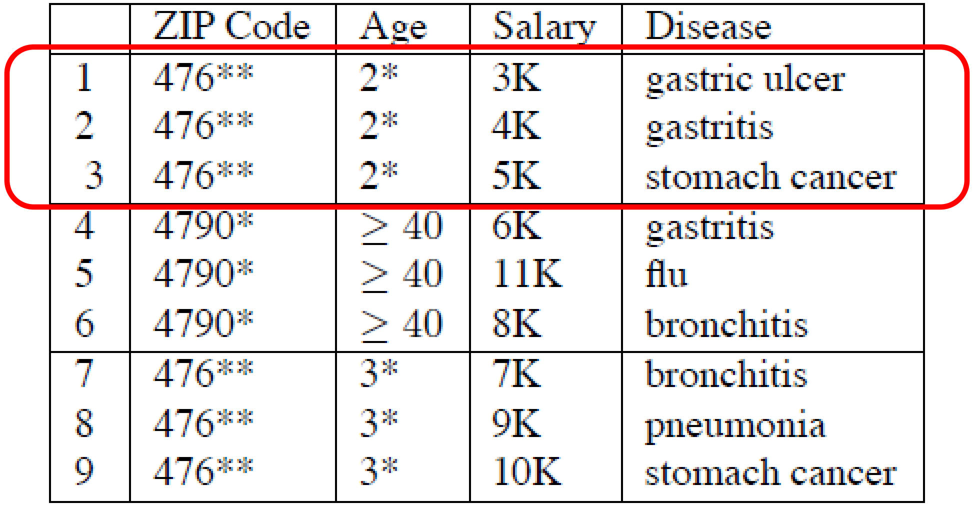
- Example
  - One sensitive attribute with two values: HIV+(1%)/HIV-(99%)
  - Suppose one class has equal number of HIV+ and HIV-
  - Satisfies any *2*-diversity requirement
  - Anyone in the class has 50% probability of being HIV+ (compare it to 1% chance in overall population)

- Issue: When the overall distribution is skewed, satisfying *l*-diversity does not prevent attribute disclosure

l-Diversity: Similarity Attack

- Bob (ZIP=47621, Age=26)

- Leakage of sensitive info
  - Low salary [3K,5K]
  - Stomach-related disease



	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

- Issue: *l*-Diversity does not take into account the semantical closeness of sensitive values

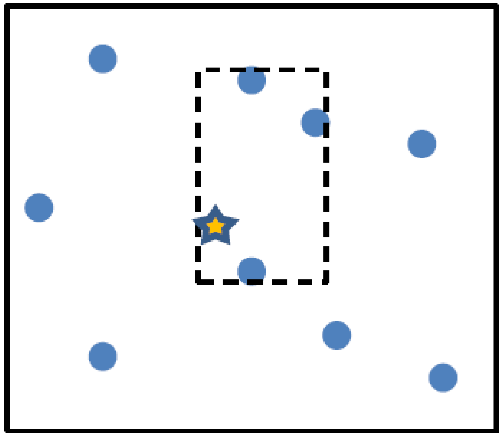
PRIVACY IN LOCATION-BASED SERVICES

Location-Based Services

- Location-Based Service (LBS)
  - A service that is offered based on a user’s location
- Privacy risks
  - Tracking a user
  - Identifying a user based on location
- Service/Privacy tradeoff
  - Report perturbed location
  - cloaking/obfuscation
    - A region containing the actual location is reported (i.e., generalization of location)

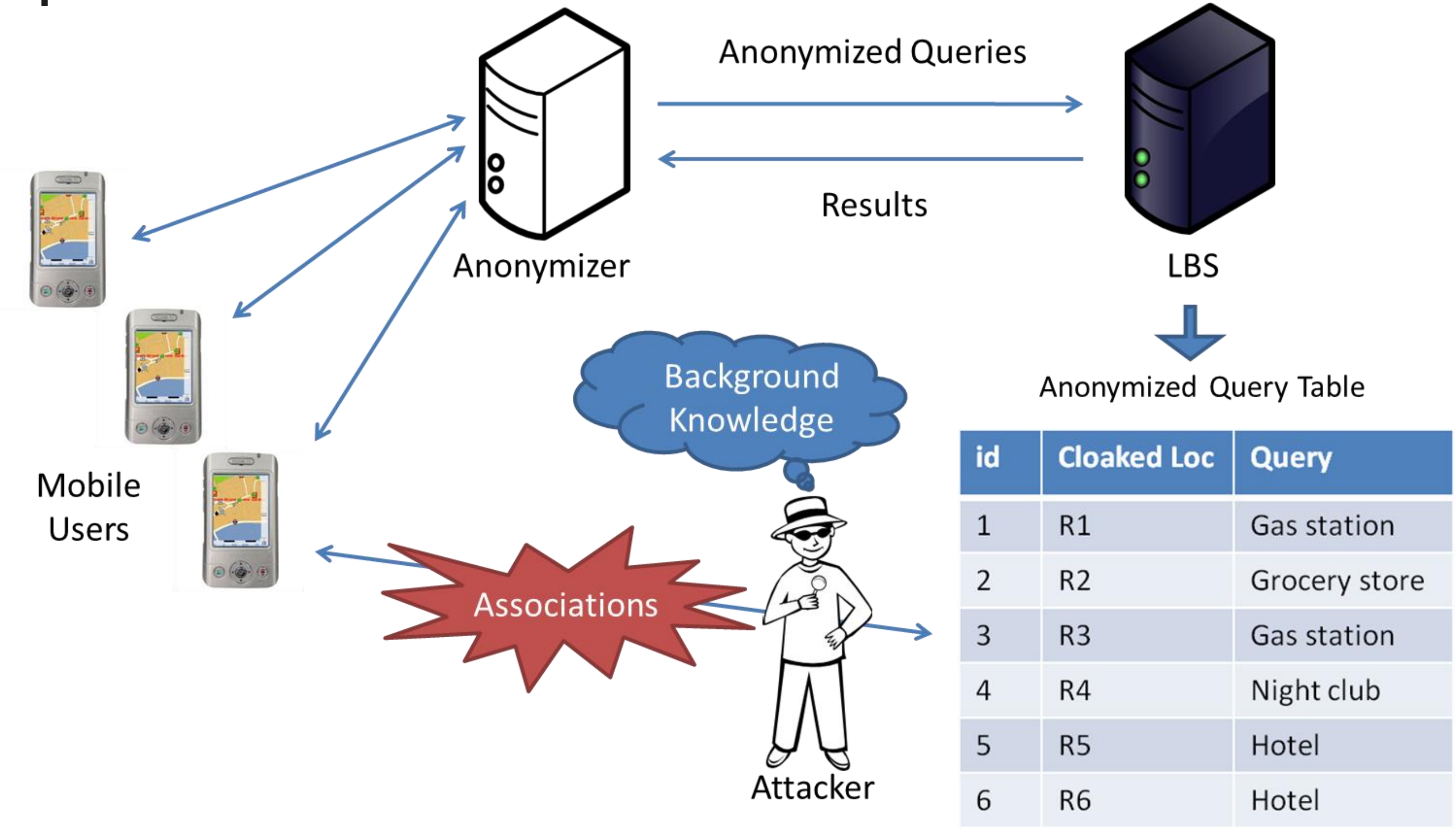
Location *k*-Anonymity

- Submitted cloaked region must contain at least *k* users
  - Collect and submit *k* queries together
  - If not enough queries to group with
    - Drop the query (may not be acceptable)
    - Generate enough dummy (fake) queries (raises service cost)



- Different users may have different privacy requirements, service level needs
  - Important distinction from traditional *k*-anonymity

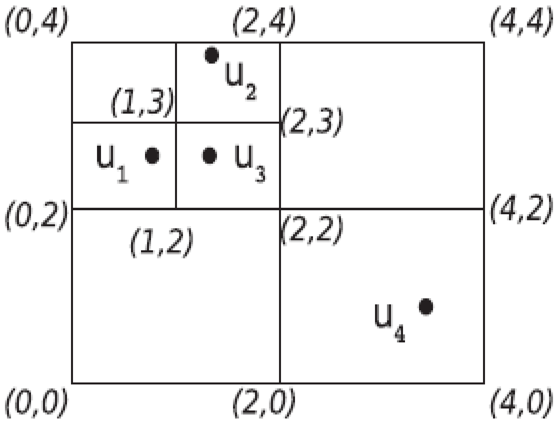
LBS Anonymization: Threat Model

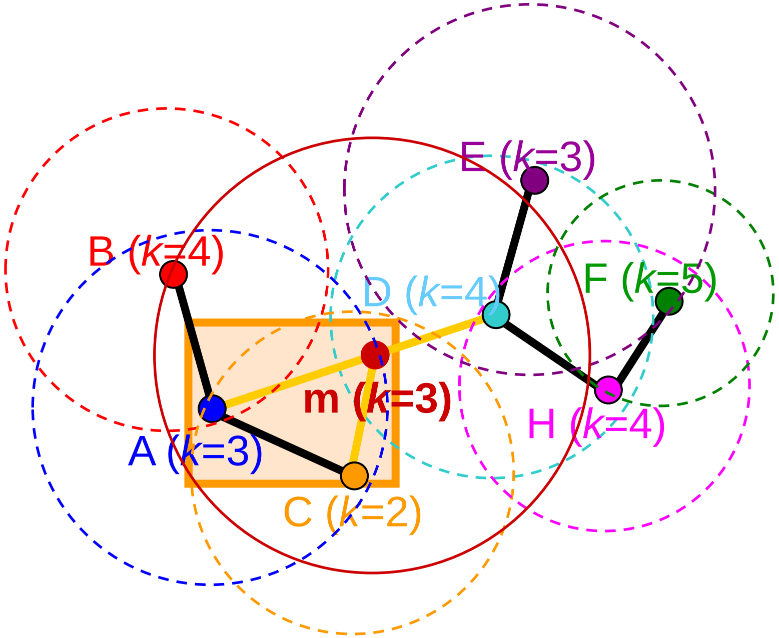


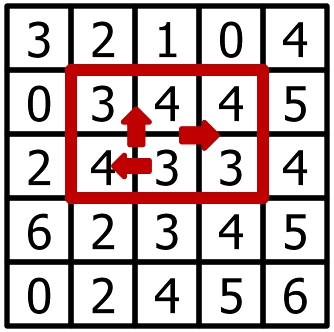
id	Cloaked Loc	Query
1	R1	Gas station
2	R2	Grocery store
3	R3	Gas station
4	R4	Night club
5	R5	Hotel
6	R6	Hotel

Location *k*-Anonymization

- Various algorithms
  - Nearest neighbor *k*-anonymization
  - Quad-tree spatial cloaking
  - CliqueCloak
  - Privacy Grid







3	2	1	0	4
0	3	4	4	5
2	4	3	3	4
6	2	3	4	5
0	2	4	5	6

PRIVACY IN SOCIAL NETWORKING SYSTEMS

Social Networking Systems

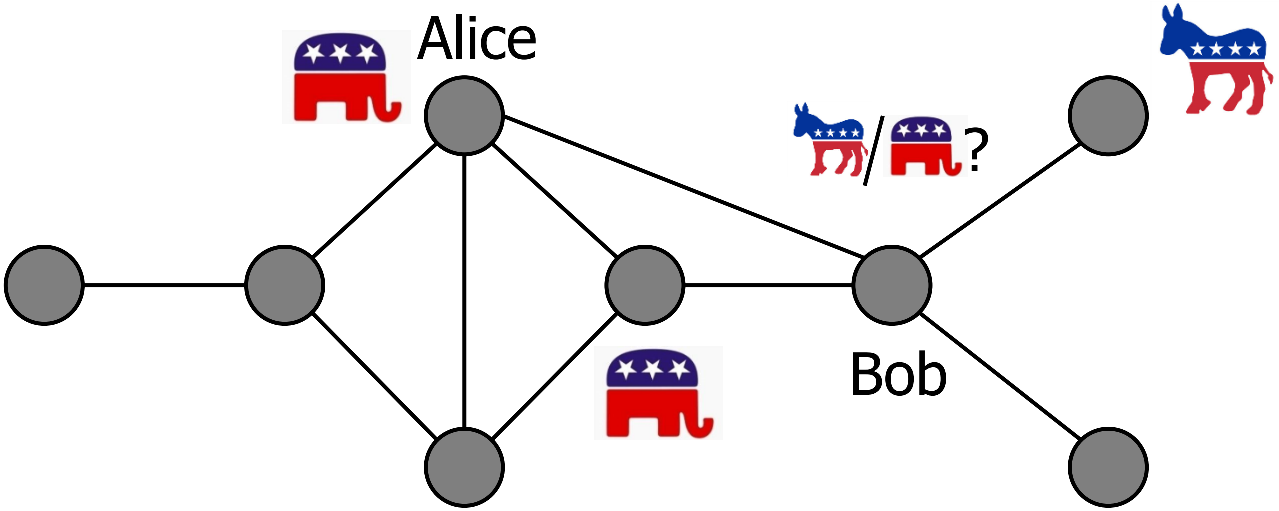
- Social networking systems (Online social networks)
  - Facebook, Orkut, LinkedIn, Twitter, Buzz, etc.
- Social network: a collection of
  - Social entities, e.g., people in Facebook, and
  - Relations among them, e.g., friendship relation in Facebook
  - Basically, a graph
    - Nodes / vertices / actors
    - Links / edges / relations

Users’ Challenges in Privacy Control

- Not enough control features
- Configuring a policy is a complicated task for an ordinary user
  - Hundreds of just directly linked friends
  - Magnitude of information objects: profile, status, posts, photos, etc.
  - Third party apps
- Even if you have the tool and knowledge to use it, still hard to determine your ideal protection preferences!

Privacy Risks in Releasing SNs

- Identity disclosure
- Link disclosure
- Attribute disclosure



Social Network Anonymization

- Generalization
  - Cluster nodes, usually based on communities
  - Replace a cluster with a hyper node
  - Only report hyper nodes, incl. summarized structural properties, and their links
- Perturbation
  - Insert/delete edges in a network to meet a privacy goal such as
    - Degree *k*-anonymity
    - Neighborhood *k*-anonymity
    - Community *k*-anonymity