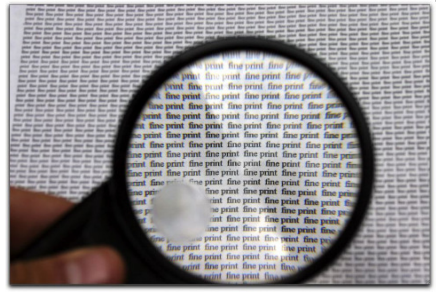
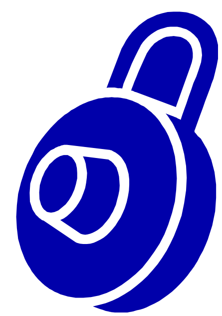



Anonymized Data: Generation, Models, Usage

Graham Cormode

Divesh Srivastava

{graham,divesh}@research.att.com

Outline

Part 1

- Introduction to Anonymization and Uncertainty
- Tabular Data Anonymization

Part 2

- Set and Graph Data Anonymization
- Models of Uncertain Data
- Query Answering on Anonymized Data
- Open Problems and Other Directions

Why Anonymize?

- For Data Sharing
  - Give real(istic) data to others to study without compromising privacy of individuals in the data
  - Allows third-parties to try new analysis and mining techniques not thought of by the data owner
- For Data Retention and Usage
  - Various requirements prevent companies from retaining customer information indefinitely
  - E.g. Google progressively anonymizes IP addresses in search logs
  - Internal sharing across departments (e.g. billing $\rightarrow$ marketing)

Why Privacy?

- Data subjects have inherent right and expectation of privacy
- "Privacy" is a complex concept (beyond the scope of this tutorial)
  - What exactly does "privacy" mean? When does it apply?
  - Could there exist societies without a concept of privacy?
- Concretely: at collection "small print" outlines privacy rules
  - Most companies have adopted a privacy policy
  - E.g. AT&T privacy policy att.com/gen/privacy-policy?pid=2506

- Significant legal framework relating to privacy
  - UN Declaration of Human Rights, US Constitution
  - HIPAA, Video Privacy Protection, Data Protection Acts

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: Netflix Prize

- Raw data: 100M dated ratings from 480K users to 18K movies
- Why released: improve predicting ratings of unlabeled examples
- How anonymized: exact details not described by Netflix
  - All direct customer information removed
  - Only subset of full data; dates modified; some ratings deleted,
  - Movie title and year published in full
- Attacks: dataset is claimed vulnerable [Narayanan Shmatikov 08]
  - Attack links data to IMDB where same users also rated movies
  - Find matches based on similar ratings or dates in both
- Consequences: rich source of user data for researchers
  - Unclear how serious the attacks are in practice

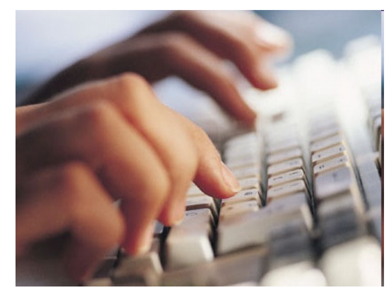
Case Study: AOL Search Data

- 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
  - NY Times identified 4417749 as 62yr old GA widow [Barbaro Zeller 06]
- Consequences: CTO resigned, two researchers fired
  - Well-intentioned effort failed due to inadequate anonymization

Three Abstract Examples

- "Census" data recording incomes and demographics
  - Schema: **(SSN, DOB, Sex, ZIP, Salary)**
  - Tabular data—best represented as a table

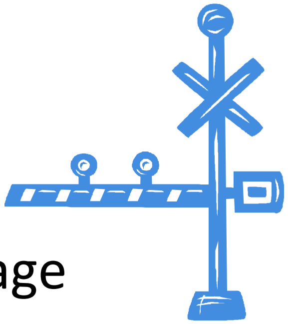
- "Video" data recording movies viewed
  - Schema: **(Uid, DOB, Sex, ZIP), (Vid, title, genre), (Uid, Vid)**
  - Graph data—graph properties should be retained



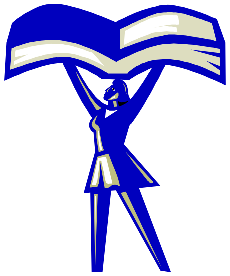
- "Search" data recording web searches
  - Schema: **(Uid, Kw1, Kw2, …)**
  - Set data—each user has different set of keywords
- Each example has different anonymization needs

Models of Anonymization

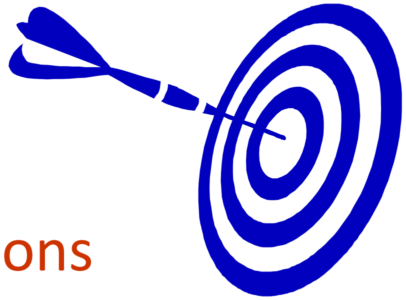
- Interactive Model (akin to statistical databases)
  - Data owner acts as "gatekeeper" to data
  - Researchers pose queries in some agreed language
  - Gatekeeper gives an (anonymized) answer, or refuses to answer



- "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
  - Our focus in this tutorial – seems to model most real releases

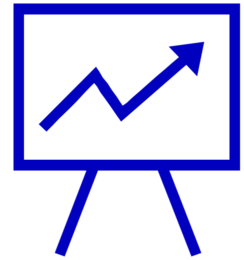


Objectives for Anonymization



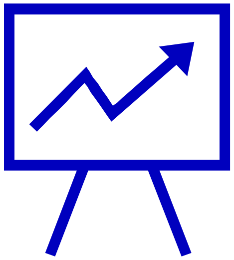
- Prevent (high confidence) inference of associations
  - Prevent inference of salary for an individual in “census”
  - Prevent inference of individual’s viewing history in “video”
  - Prevent inference of individual’s search history in “search”
  - All aim to prevent linking sensitive information to 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)

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

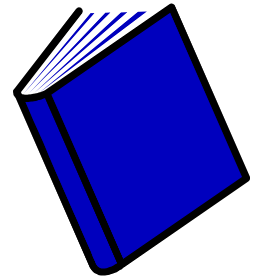
Measures of Utility



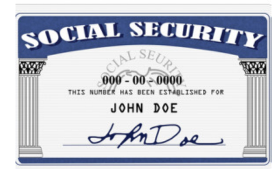
- Define a surrogate measure and try to optimize
  - Often based on the "information loss" of the anonymization
  - Simple example: number of rows suppressed in a table
- Give a guarantee for all queries in some fixed class
  - Hope the class is representative, so other uses have low distortion
  - Costly: 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

at&t

Definitions of Technical Terms

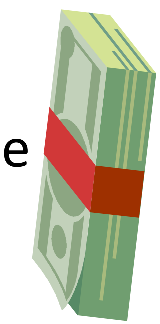


- Identifiers–uniquely identify, e.g. Social Security Number (SSN)
  - Step 0: remove all identifiers
  - Was not enough for AOL search data



- 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 [Sweeney 02]

- 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
  - In “video”, association between user and video is sensitive
  - SA can be identifying: bonus may identify salary…



Summary of Anonymization Motivation

- Anonymization needed for safe data sharing and retention
  - Many legal requirements apply
- Various privacy definitions possible
  - Primarily, prevent inference of sensitive information
  - Under some assumptions of background knowledge
- Utility of the anonymized data needs to be carefully studied
  - Different data types imply different classes of query

- Our focus: publishing model with careful utility consideration
  - Data types: tables (census), sets and graphs (video & search)

Anonymization as Uncertainty

- We view anonymization as 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
- Much recent work on anonymization and uncertainty separately
  - Here, we aim to bring them together
  - More formal framework for anonymization
  - New application to drive uncertainty

Possible Worlds

- Uncertain Data typically represents multiple possible worlds
  - Each possible world corresponds to a database (or graph, 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?

Outline

Part 1

- Introduction to Anonymization and Uncertainty
- **Tabular Data Anonymization**

Part 2

- Set and Graph Data Anonymization
- Models of Uncertain Data
- Query Answering on Anonymized Data
- Open Problems and Other Directions

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 $\rightarrow$ 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
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
11-1-111	1/21/76
33-3-333	2/28/76

- 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

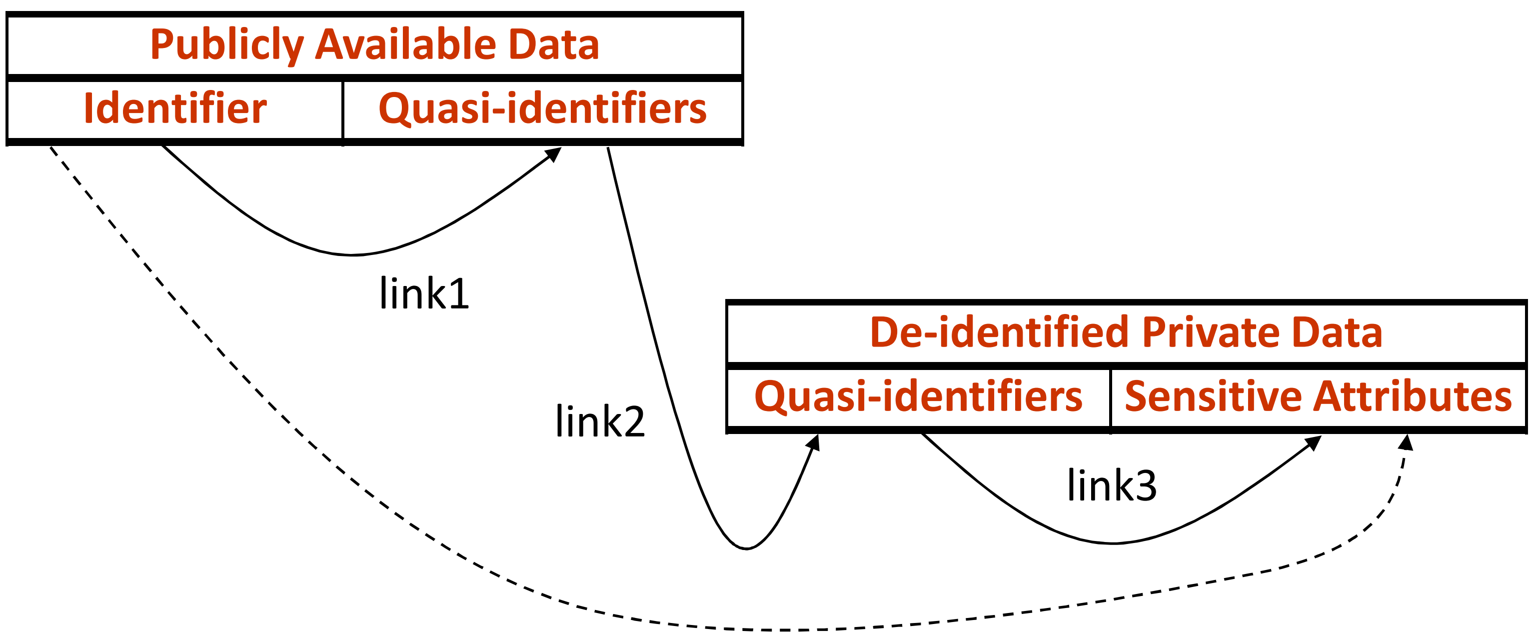
DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

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 [Sweeney 02]

Tabular Data: Linking Attack

- Observation: Identifier $\rightarrow$ SA is a composition of link1, link2, link3
  - Generalization-based techniques weaken link2
  - Permutation-based techniques weaken link3



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., ZIP = 537** → ZIP ∈ {53700, …, 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
  - If distribution of perturbation is given, it implicitly defines a model

k-Anonymization [Samarati, Sweeney 98]

- 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

- Huge literature: 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 (e.g., numerical attributes)

Algorithm	Model	Properties	Complexity
Samarati 01	G+TS, FD, HB	One exact, binary search	$O(2^{\lvert QI \rvert})$
Sweeney 02	G+TS, FD, HB	Exact, exhaustive	$O(2^{\lvert QI \rvert})$
Bayardo+ 05	G+TS, FD, PB	Exact, top-down	$O(2^{\lvert QI \rvert})$
LeFevre+ 05	G+TS, FD, HB	All exact, bottom-up cube	$O(2^{\lvert QI \rvert})$

Computing k-Anonymizations

- Huge literature: 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

Algorithm	Model	Properties	Complexity
Iyengar 02	G+TS, FD, PB	Heuristic, stochastic search	No bounds
Winkler 02	G+TS, FD, HB	Heuristic, simulated annealing	No bounds
Fung+ 05	G, FD, PB	Heuristic, top-down	No bounds

Computing k-Anonymizations

- Huge literature: 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

Algorithm	Model	Properties	Complexity
Meyerson+ 04	S, L	NP-hard, O(k log k) approximation	$O(n^{2k})$
Aggarwal+ 05a	S, L	O(k) approximation	$O(kn^2)$
Aggarwal+ 05b	G, L, HB	O(k) approximation	$O(kn^2)$
LeFevre+ 06	G, MD, PB	Constant-factor approximation	O(n log n)

Incognito [LeFevre+ 05]

- 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 [LeFevre+ 05]

- Every full-domain generalization described by a "domain vector"
  - B0={1/21/76, 2/28/76, 4/13/86} → B1={76-86}
  - S0={M, F} → S1={*}
  - Z0={53715,53710,53706,53703}→ Z1={5371*,5370*}→ 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 [LeFevre+ 05]

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

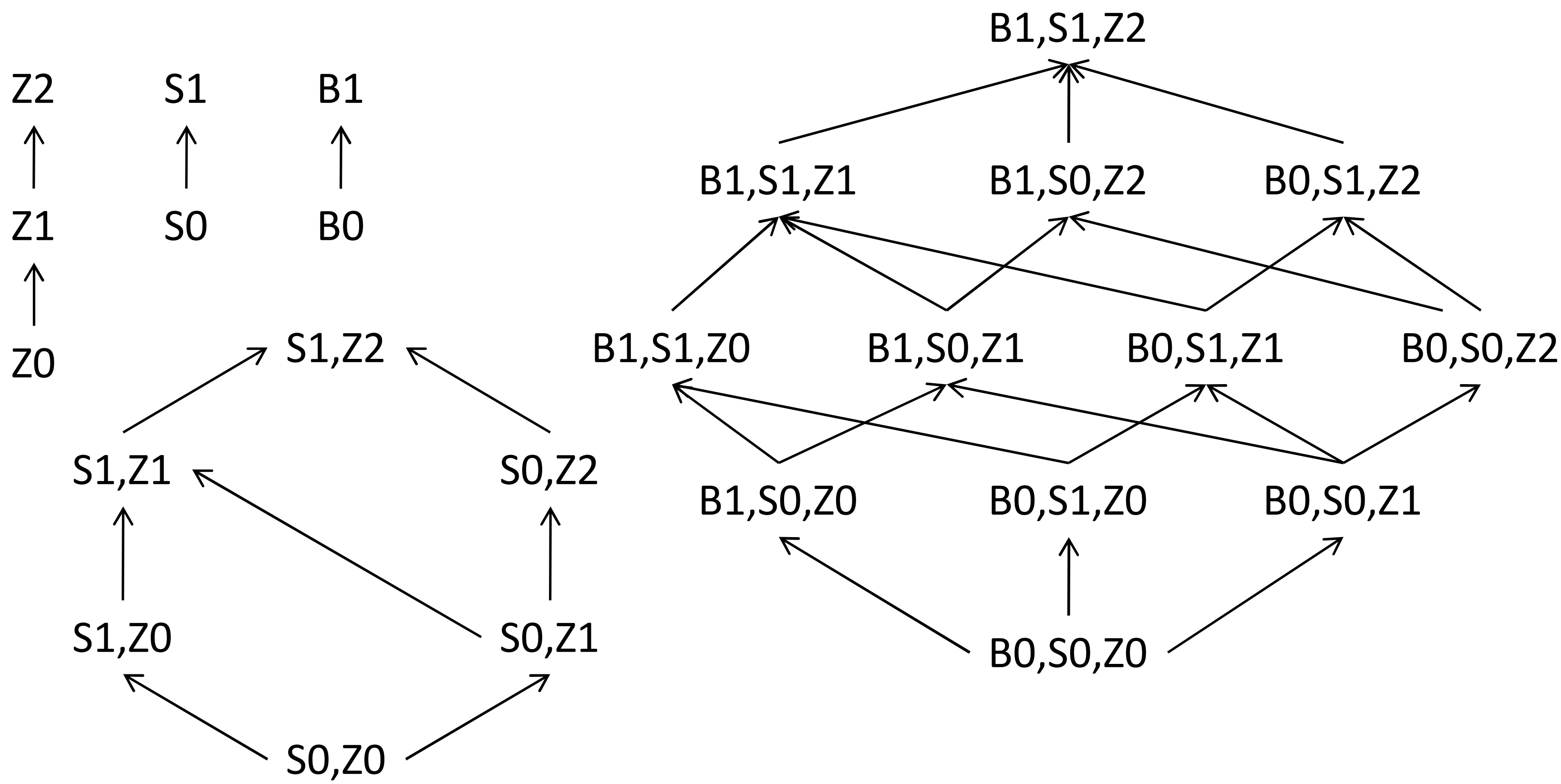
DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

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

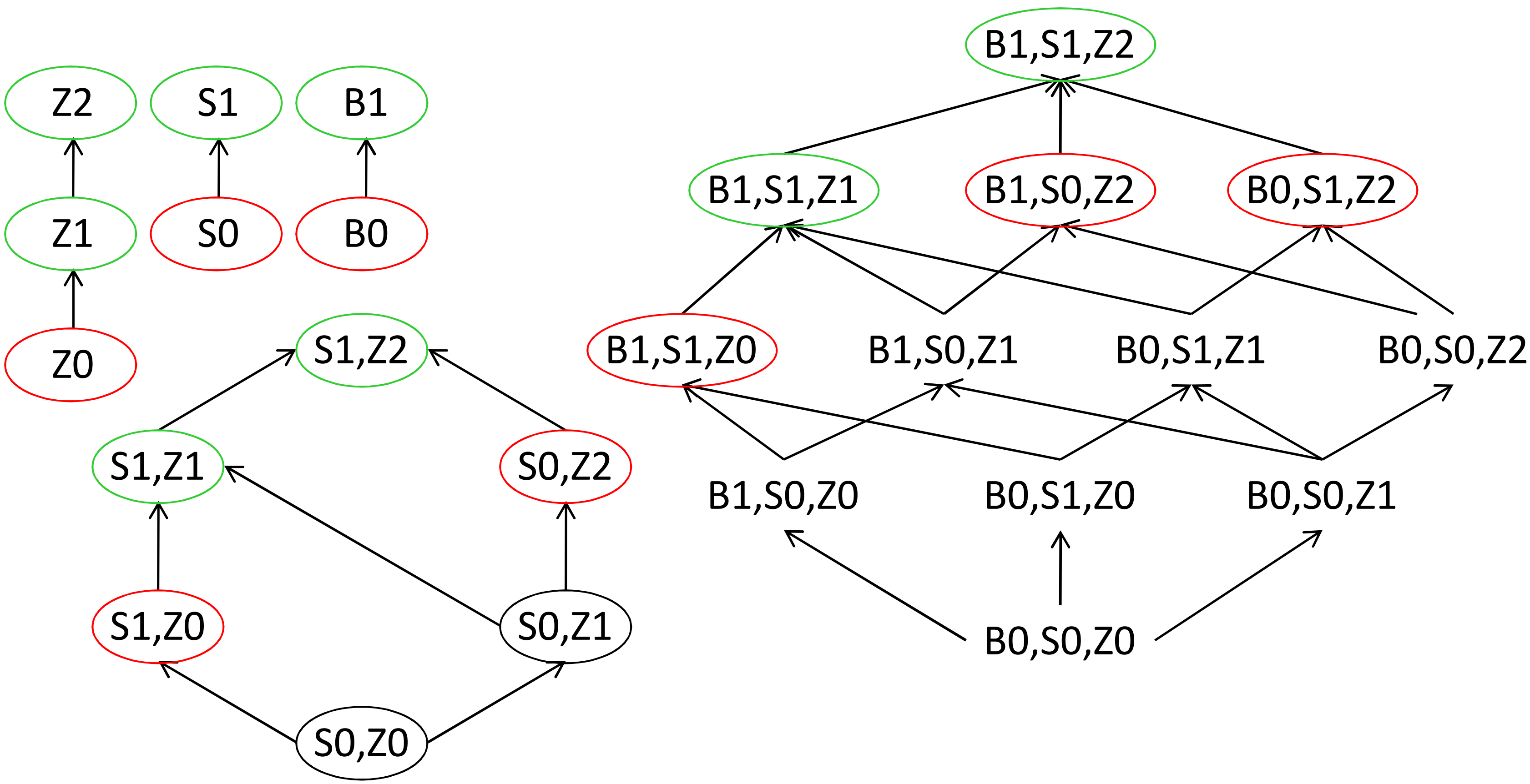
Incognito [LeFevre+ 05]

- Lattice of domain vectors



Incognito [LeFevre+ 05]

- Lattice of domain vectors



Incognito [LeFevre+ 05]

- Subset Property: If table T is k-anonymous wrt 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
- 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

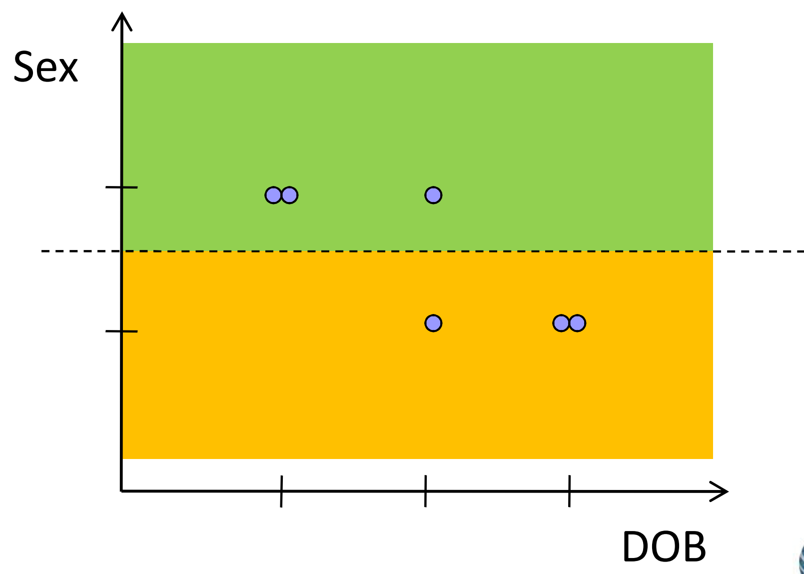
Mondrian [LeFevre+ 06]

- 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 = (total tuples / total groups) / k
- Efficient: greedy O(n log n) heuristic for NP-hard problem
- Quality guarantee: solution is a constant-factor approximation

Mondrian [LeFevre+ 06]

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

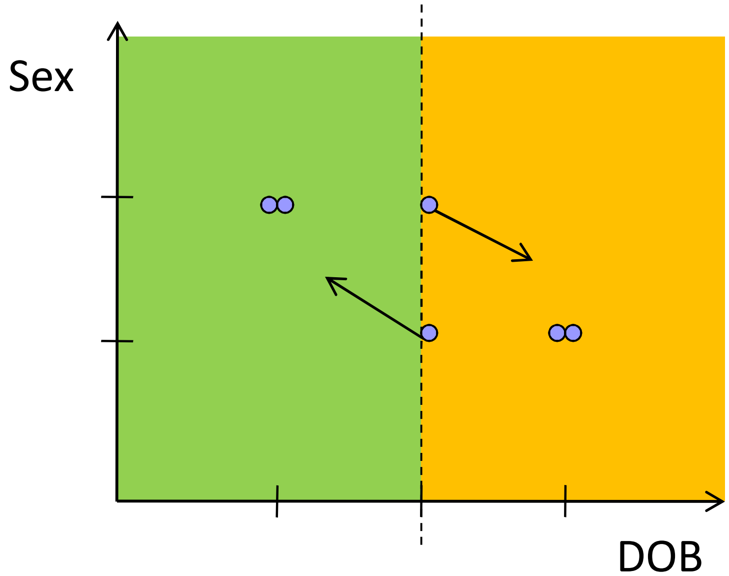
DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000



Mondrian [LeFevre+ 06]

- Uses ideas from spatial kd-tree construction
  - QI tuples = points in a multi-dimensional space
  - Hyper-rectangles with ≥ 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 [Machanavajjhala+ 06]

- Issue: k-anonymity requires each tuple in (the multiset) T[QI] to appear ≥ 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 [Machanavajjhala+ 06]

- Issue: k-anonymity requires each tuple in (the multiset) T[QI] to appear ≥ 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
  - For some groupings, no loss of privacy

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 [Machanavajjhala+ 06]

- *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, entropy(g) ≥ log(*l*)
  - Recursive (c,*l*)-diversity: for each QI group g with m SA values, and r_i the i'th highest frequency, $r_1 < c\ (r_l + r_{l+1} + \ldots + r_m)$
  - Folk *l*-diversity: for each QI group g, no SA value should occur more than 1/*l* fraction of the time = Recursive(1/*l*, 1)-diversity

l-Diversity [Machanavajjhala+ 06]

- 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, entropy(g) ≥ log(*l*)
  - *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 [Machanavajjhala+ 06]

- 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 [Li+ 07]

- 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 [Xiao+ 06, Zhang+ 07]

- 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 [Xiao+ 06, Zhang+ 07]

- 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

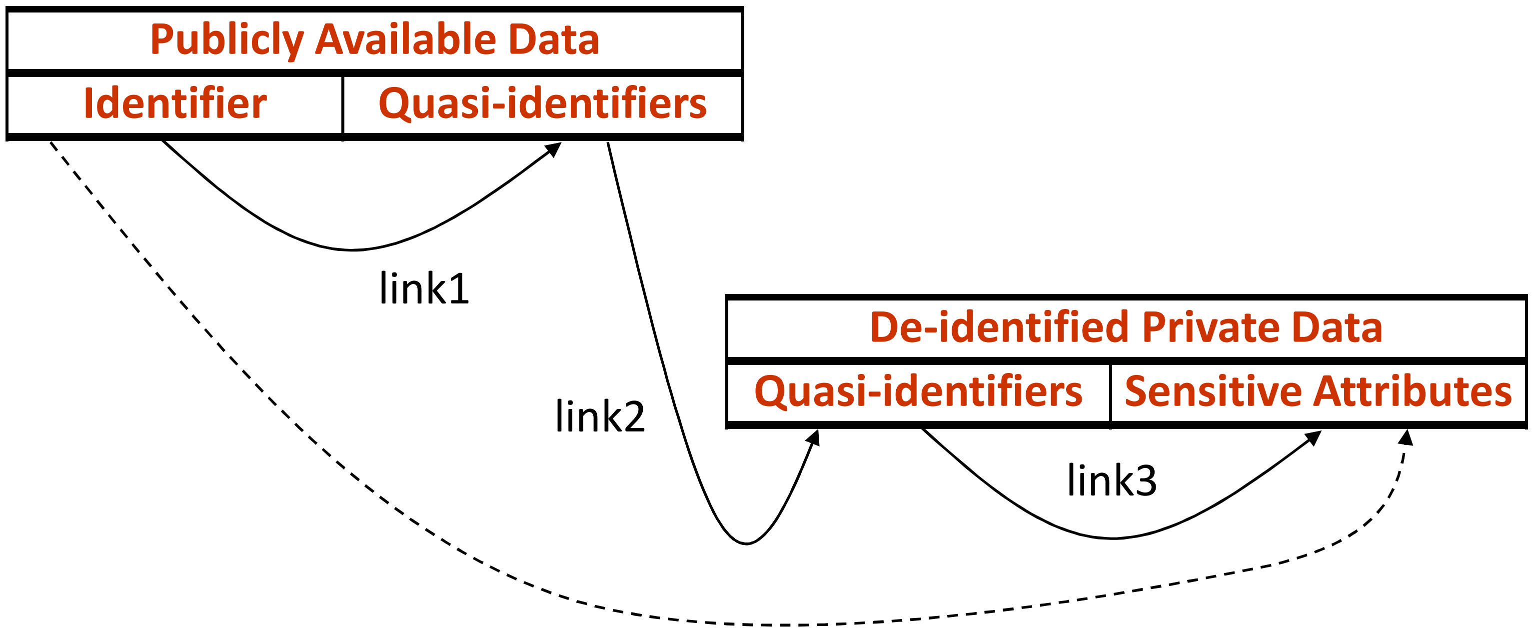
→

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 $\rightarrow$ SA is a composition of link1, link2, link3
  - Generalization-based techniques weaken link2

- Alternative: Weaken link 3 (QI $\rightarrow$ SA association in private data)



Permutation: Basics [Xiao+ 06, Zhang+ 07]

- 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
1/21/76	M	53715
4/13/86	F	53715
2/28/76	M	53703
1/21/76	M	53703
4/13/86	F	53706
2/28/76	F	53706

Salary
60,000
75,000
65,000
50,000
70,000
55,000

at&t

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
1/21/76	M	53715
4/13/86	F	53715
2/28/76	M	53703
1/21/76	M	53703
4/13/86	F	53706
2/28/76	F	53706

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

Salary
60,000
75,000
65,000
50,000
70,000
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 [Xiao+ 06]: form *l*-diverse groups
  - Hash SA values into buckets
  - Iteratively pick 1 value from each of the *l* most populated buckets

- Permutation [Zhang+ 07]: 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

DOB	Sex	ZIP
1/21/76	M	53715
4/13/86	F	53715
2/28/76	M	53703
1/21/76	M	53703
4/13/86	F	53706
2/28/76	F	53706

Salary
60,000
75,000
65,000
50,000
70,000
55,000

No! →

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

Tabular Anonymization and Uncertainty

- Generalization + Suppression: natural representation and efficient reasoning using Uncertain Database models
- Permutation:
  - Can be represented with c-tables, MayBMS in a tedious way
  - Weaker knowledge can be represented in Trio model
- New research: working models to precisely handle permutation
  - Bijection as a primitive?

Recent Attacks and Uncertainty

- Minimality Attack [Wong+ 07]:
  - Uses knowledge of anonymization algorithm to argue some possible worlds are not consistent with output
- deFinetti Attack [Kifer 09]:
  - 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

Outline

Part 1

- Introduction to Anonymization and Uncertainty
- Tabular Data Anonymization

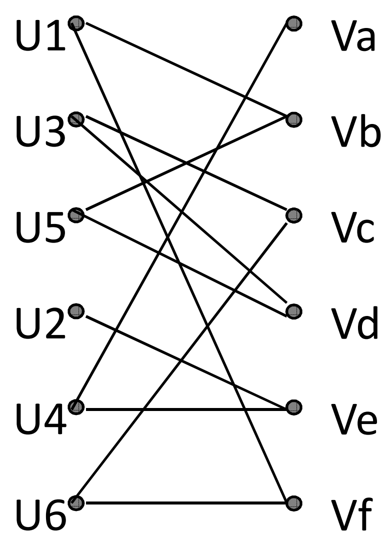
Part 2

- **Set and Graph Data Anonymization**
- Models of Uncertain Data
- Query Answering on Anonymized Data
- Open Problems and Other Directions

Graph (Multi-Tabular) Data Example

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



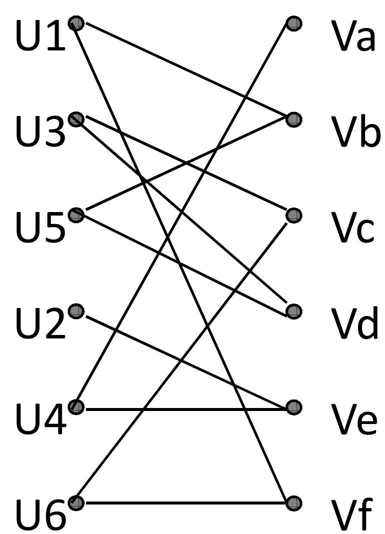
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- Similar associations arise in medical data (Patient, Symptoms), search logs (User, Keyword)

Graph (Multi-Tabular) Data Example

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



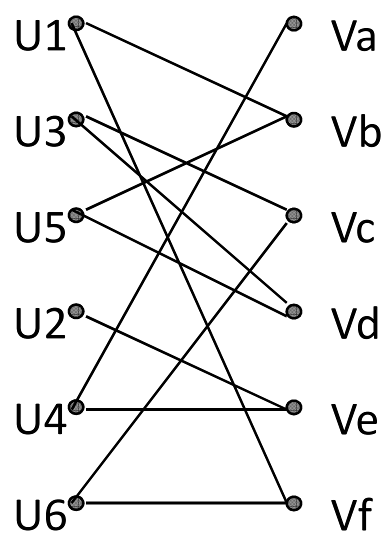
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- Releasing Uid → Vid association violates individual's privacy, possibly for a subset of videos across all users

Graph (Multi-Tabular) Data Example

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



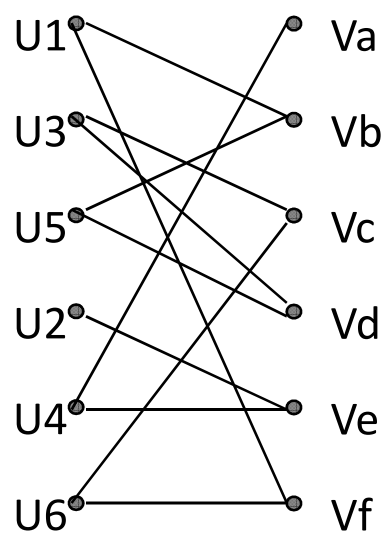
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- Releasing Uid → Vid association violates individual's privacy, possibly for different subsets of videos for different users

Graph (Multi-Tabular) Data Example

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706

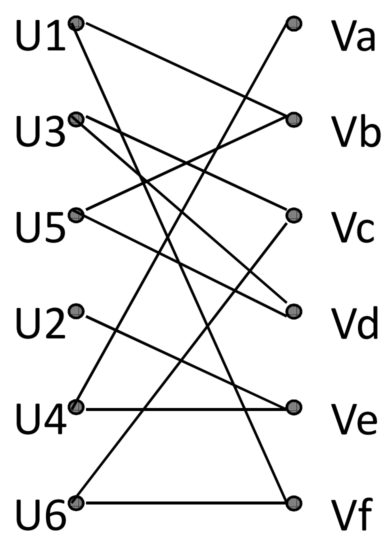


Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- Releasing Uid → Vid association violates individual's privacy, possibly for different subsets of videos for different users

Graph Data: Multi-table Linking Attack

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



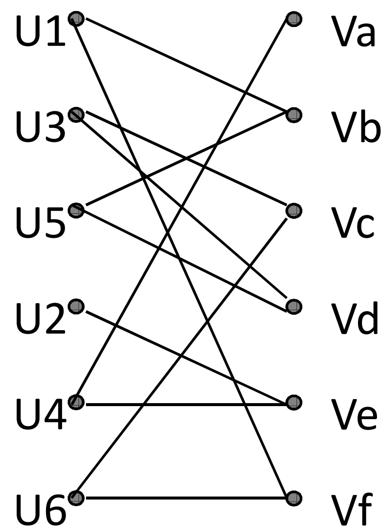
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

SSN	DOB	Sex	Title
11-1-111	1/21/76	M	apartment

Graph Data: Homogeneity Attack

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	*	537**
U2	4/13/86	*	537**
U3	2/28/76	*	537**
U4	1/21/76	*	537**
U5	4/13/86	*	537**
U6	2/28/76	*	537**



Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

SSN	DOB	Sex	ZIP
33-3-333	2/28/76	M	53703

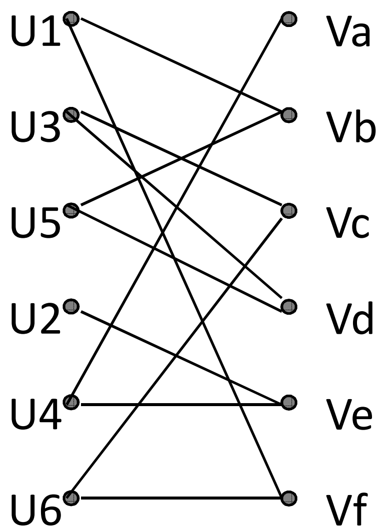
Graph Data Anonymization

- Goal: publish anonymized and useful version of graph data
- Privacy goals
  - Hide associations involving private entities in graph
  - Allow for static attacks (inferred from published graph)
  - Allow for learned edge attacks (background public knowledge)
- Useful queries
  - Queries on graph structure ("Type 0")
  - Queries on graph structure + entity attributes ("Types 1, 2")

Graph Data: Type 0 Query

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



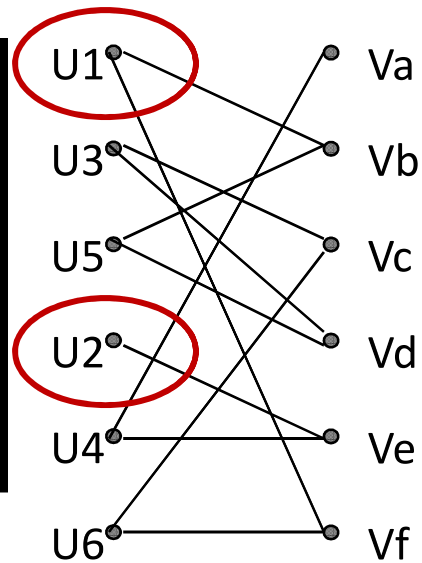
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- What is the average number of videos viewed by users? 11/6

Graph Data: Type I Query

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



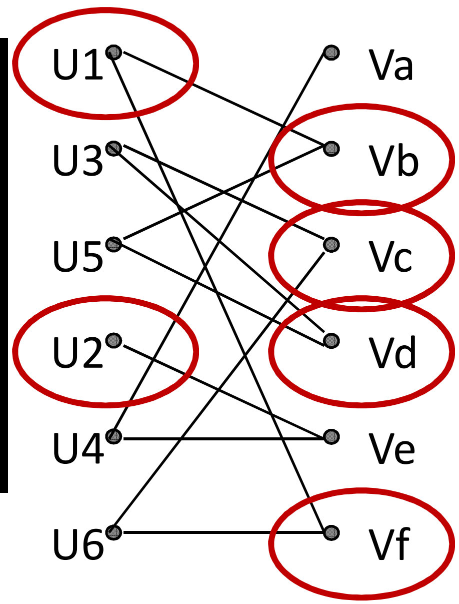
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- What is the average number of videos viewed by users in the 53715 ZIP? 3/2

Graph Data: Type 2 Query

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- What is the average number of comedy videos viewed by users in the 53715 ZIP? 1

(h,k,p)-Coherence [Xu+ 08]

- Universal private videos, model graph using sets in a single table
  - Public video set akin to high-dimensional quasi-identifier
  - Allow linking attack through public video set

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{ }
U2	4/13/86	F	53715	{ }	{SV}
U3	2/28/76	M	53703	{HG, In}	{ }
U4	1/21/76	M	53703	{ }	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{ }
U6	2/28/76	F	53706	{HG, LB}	{ }

(h,k,p)-Coherence [Xu+ 08]

- New privacy model parameterized by "power" (p) of attacker
  - (h,k,p)-coherence: for every combination S of at most p public items in a tuple of table T, at least k tuples must contain S and no more than h % of these tuples should contain a common private item
- Is the following table (50%,2,1)-coherent? **Yes**

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{ }
U2	4/13/86	F	53715	{ }	{SV}
U3	2/28/76	M	53703	{HG, In}	{ }
U4	1/21/76	M	53703	{ }	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{ }
U6	2/28/76	F	53706	{HG, LB}	{ }

(h,k,p)-Coherence [Xu+ 08]

- New privacy model parameterized by "power" (p) of attacker
  - (h,k,p)-coherence: for every combination S of at most p public items in a tuple of table T, at least k tuples must contain S and no more than h % of these tuples should contain a common private item
- Is the following table (50%,2,2)-coherent? **No**

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{ }
U2	4/13/86	F	53715	{ }	{SV}
U3	2/28/76	M	53703	{HG, In}	{ }
U4	1/21/76	M	53703	{ }	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{ }
U6	2/28/76	F	53706	{HG, LB}	{ }

(h,k,p)-Coherence [Xu+ 08]

- Greedy algorithm to achieve (h,k,p)-coherence
  - Identify minimal "moles" using an Apriori algorithm
  - Suppress item that minimizes normalized "information loss"
- To achieve (50%,2,2)-coherence
  - Pick minimal "mole" {HG, In}, suppress HG globally

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{ }
U2	4/13/86	F	53715	{ }	{SV}
U3	2/28/76	M	53703	{~~HG~~, In}	{ }
U4	1/21/76	M	53703	{ }	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{ }
U6	2/28/76	F	53706	{~~HG~~, LB}	{ }

(h,k,p)-Coherence [Xu+ 08]

- Greedy algorithm to achieve (h,k,p)-coherence
  - Identify minimal “moles” using an Apriori algorithm
  - Suppress item that minimizes normalized “information loss”
- To achieve (50%,2,2)-coherence
  - Pick minimal “mole” {Ap, LB}, suppress Ap globally

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{~~Ap~~, LB}	{ }
U2	4/13/86	F	53715	{ }	{SV}
U3	2/28/76	M	53703	{~~HG~~, In}	{ }
U4	1/21/76	M	53703	{ }	{HC, SV}
U5	4/13/86	F	53706	{~~Ap~~, In}	{ }
U6	2/28/76	F	53706	{~~HG~~, LB}	{ }

Properties of (h,k,p)-Coherence

- Preserves support of item sets present in anonymized table
  - Critical for computing association rules from anonymized table
  - But, no knowledge of some items present in original table
- Vulnerable to linking attack with negative information
  - Table is (50%,2,2)-coherent, but {LB, ¬Ap} identifies U4

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB, In}	{ }
U2	4/13/86	F	53715	{Ap, LB}	{SV}
U3	2/28/76	M	53703	{HG, FW}	{ }
U4	1/21/76	M	53703	{LB, In}	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{ }
U6	2/28/76	F	53706	{HG, FW}	{ }

(h,k,p)-Coherence and Uncertainty

- Intuition: An (h,k,p)-coherent T' represents the set of all "possible world" tables T_i s.t. T' is an (h,k,p)-coherent suppression of T_i
  - Need to identify number of suppressed items in each public item set
  - Obtain T_i from T' by adding non-suppressed items from universe

Graph Data Anonymization [Ghinita+ 08]

- Universal private videos, model graph as a single sparse table

Uid	DOB	Sex	ZIP	Ap	HG	In	LB	HC	SV
U1	1/21/76	M	53715	1	0	0	1	0	0
U2	4/13/86	F	53715	0	0	0	0	0	1
U3	2/28/76	M	53703	0	1	1	0	0	0
U4	1/21/76	M	53703	0	0	0	0	1	1
U5	4/13/86	F	53706	1	0	1	0	0	0
U6	2/28/76	F	53706	0	1	0	1	0	0

- Permutation-based approach, cluster tuples based on similarity of public video vectors, ensure diversity of private videos

Graph Data Anonymization [Ghinita+ 08]

- Clustering: reorder rows and columns to create a band matrix
  - Specifically to improve utility of queries
- ≤ 1 occurrence of each private video in a group to get *l*-diversity
  - Group private-video tuple with *l*-1 adjacent "non-conflicting" tuples

Uid	DOB	Sex	ZIP	LB	Ap	HG	In
U2	4/13/86	F	53715	0	0	0	0
U1	1/21/76	M	53715	1	1	0	0
U6	2/28/76	F	53706	1	0	1	0
U5	4/13/86	F	53706	0	1	0	1
U3	2/28/76	M	53703	0	0	1	1
U4	1/21/76	M	53703	0	0	0	0

HC	SV
0	1
0	0
0	0
0	0
0	0
1	1

Properties of [Ghinita+ 08]

- Permutation-based approach is good for query accuracy
  - No loss of information via generalization or suppression
- Experimental study measured KL-divergence (surrogate measure) of anonymized data from original data
  - Compared to permutation grouping found via Mondrian
  - Observed that KL-divergence via clustering was appreciably better
- Uncertainty model is the same as for tabular data!

k^m-Anonymization [Terrovitis+ 08]

- No *a priori* distinction between public and private videos
  - Allow linking attack using any item set, remaining items are private
  - Model graph using public item set = private item set in a single table
- Simplified model for personalized privacy (e.g., AOL search log)
  - Each user has own (but unknown) set of public and private items

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{Ap, LB}
U2	4/13/86	F	53715	{SV}	{SV}
U3	2/28/76	M	53703	{HG, In}	{HG, In}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{Ap, In}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}

k^m-Anonymization [Terrovitis+ 08]

- New privacy model parameterized by "power" (m) of attacker
  - k^m-anonymity: for every combination S of at most m public items in a tuple of table T, at least k tuples must contain S
  - Note: no diversity condition specified on private items
- Is the following table k^m-anonymous, m=2? **No**

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{Ap, LB}
U2	4/13/86	F	53715	{SV}	{SV}
U3	2/28/76	M	53703	{HG, In}	{HG, In}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{Ap, In}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}

k^m-Anonymization [Terrovitis+ 08]

- k^m-anonymity: for every combination S of at most m public items in a tuple of table T, at least k tuples must contain S
- Is the following table k^m-anonymous, m=1? **No**
  - Recall that the graph was (50%,2,1)-coherent
- Observation: (h,k,p)-coherence does not imply k^p-anonymity

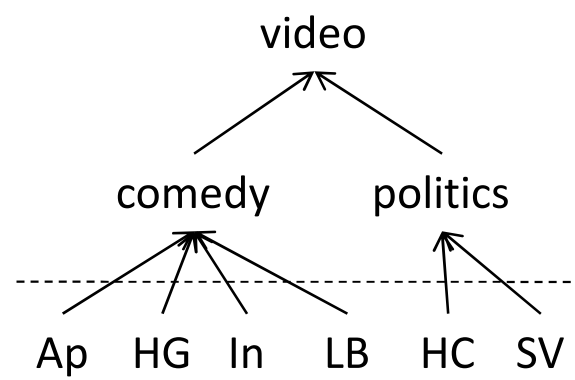
Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{Ap, LB}
U2	4/13/86	F	53715	{SV}	{SV}
U3	2/28/76	M	53703	{HG, In}	{HG, In}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{Ap, In}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}

at&t

k^m-Anonymization [Terrovitis+ 08]

- k^m-anonymization: given a generalization hierarchy on items, a table T' is a k^m-anonymization of table T if T' is k^m-anonymous and is obtained by generalizing items in each tuple of T
  - Search space defined by a cut on the generalization hierarchy

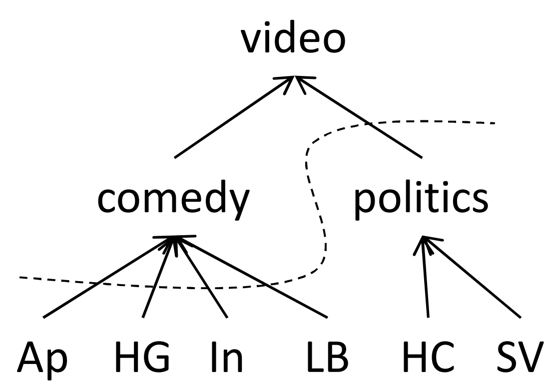
Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{Ap, LB}
U2	4/13/86	F	53715	{SV}	{SV}
U3	2/28/76	M	53703	{HG, In}	{HG, In}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{Ap, In}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}



k^m-Anonymization [Terrovitis+ 08]

- k^m-anonymization: given a generalization hierarchy on items, a table T' is a k^m-anonymization of table T if T' is k^m-anonymous and is obtained by generalizing items in each tuple of T
  - Search space defined by a cut on the generalization hierarchy
  - Global recoding (but not full-domain): k^m-anonymous (k=2, m=1)

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB}	{Ap, LB}
U2	4/13/86	F	53715	{politics}	{politics}
U3	2/28/76	M	53703	{HG, In}	{HG, In}
U4	1/21/76	M	53703	{politics}	{politics}
U5	4/13/86	F	53706	{Ap, In}	{Ap, In}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}



k^m-Anonymization [Terrovitis+ 08]

- Optimal k^m-anonymization minimizes NCP metric
  - Bottom-up, breadth-first exploration of lattice of hierarchy cuts
  - NCP: based on % of domain items covered by recoded values
- Heuristic based on Apriori principle
  - If itemset of size i causes privacy breach, so does itemset of size $i+1$
  - Much faster than optimal algorithm, very similar NCP value
- Issues:
  - k^m-anonymization vulnerable to linking attack with negative info
  - k^m-anonymization vulnerable to lack of diversity

k-Anonymization [He Naughton 09]

- Motivation: k^m vulnerable to linking attack with negative info
  - Table satisfies 2^2-anonymity, but {LB, ¬HG} identifies U1

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{LB}	{LB}
U2	4/13/86	F	53715	{HC, SV}	{HC, SV}
U3	2/28/76	M	53703	{HG}	{HG}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{HG, LB}	{HG, LB}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}

k-Anonymization [He Naughton 09]

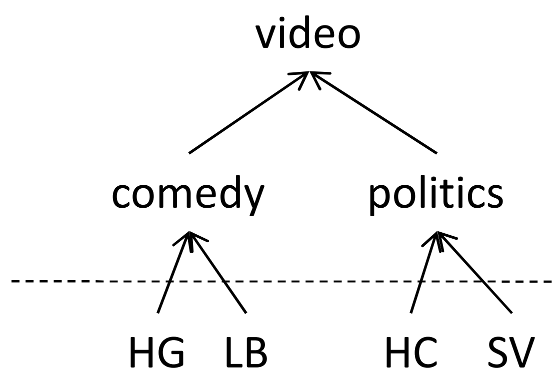
- "Old" solution (k-anonymity): for every public item set S in a tuple of table T, at least k tuples must have S as its public item set
  - Is k-anonymity = k^{max}-anonymity?
  - No! Table is 2^2-anonymous, but not 2-anonymous

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{LB}	{LB}
U2	4/13/86	F	53715	{HC, SV}	{HC, SV}
U3	2/28/76	M	53703	{HG}	{HG}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{HG, LB}	{HG, LB}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}

k-Anonymization [He Naughton 09]

- k-anonymization: given a generalization hierarchy on items, a table T' is a k-anonymization of table T if T' is k-anonymous and is obtained by generalizing items in each tuple of T
  - Search space defined by cuts on the generalization hierarchy

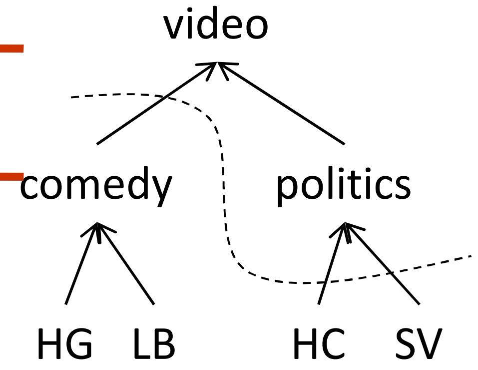
Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{LB}	{LB}
U2	4/13/86	F	53715	{HC, SV}	{HC, SV}
U3	2/28/76	M	53703	{HG}	{HG}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{HG, LB}	{HG, LB}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}



k-Anonymization [He Naughton 09]

- k-anonymization: given a generalization hierarchy on items, a table T' is a k-anonymization of table T if T' is k-anonymous and is obtained by generalizing items in each tuple of T
  - Search space defined by cuts on the generalization hierarchy
  - Local recoding: k-anonymous (k=2)

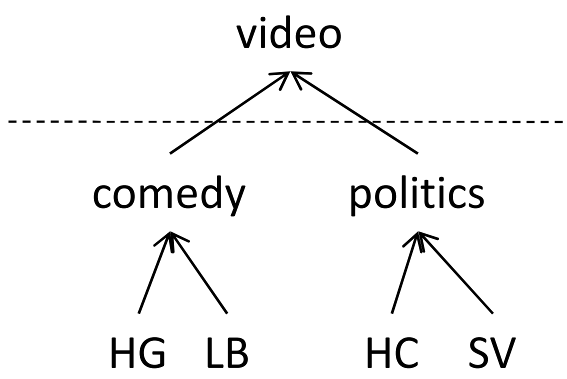
Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{comedy}	{comedy}
U2	4/13/86	F	53715	{HC, SV}	{HC, SV}
U3	2/28/76	M	53703	{comedy}	{comedy}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{HG, LB}	{HG, LB}
U6	2/28/76	F	53706	{HG, LB}	{HG, LB}



k-Anonymization [He Naughton 09]

- Greedy partitioning algorithm
  - Top-down exploration of lattice of hierarchy cuts
  - Local recoding → each equivalence class uses its own hierarchy cut
  - Much faster than bottom-up algorithm using global recoding
  - Lower information loss (NCP) than bottom-up algorithm

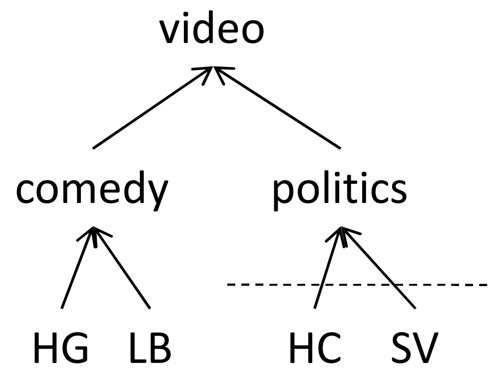
Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{comedy}	{comedy}
U2	4/13/86	F	53715	{politics}	{politics}
U3	2/28/76	M	53703	{comedy}	{comedy}
U4	1/21/76	M	53703	{politics}	{politics}
U5	4/13/86	F	53706	{comedy}	{comedy}
U6	2/28/76	F	53706	{comedy}	{comedy}



k-Anonymization [He Naughton 09]

- Greedy partitioning algorithm
  - Top-down exploration of lattice of hierarchy cuts
  - Local recoding → each equivalence class uses its own hierarchy cut
  - Much faster than bottom-up algorithm using global recoding
  - Lower information loss (NCP) than bottom-up algorithm

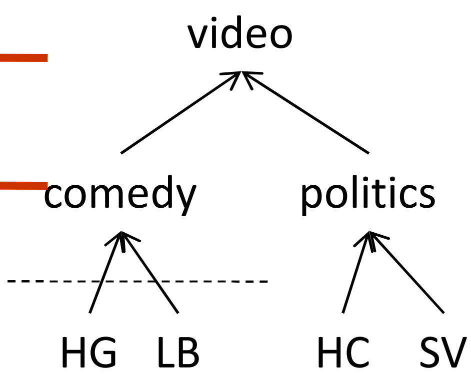
Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{comedy}	{comedy}
U2	4/13/86	F	53715	{HC, SV}	{HC, SV}
U3	2/28/76	M	53703	{comedy}	{comedy}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{comedy}	{comedy}
U6	2/28/76	F	53706	{comedy}	{comedy}



k-Anonymization [He Naughton 09]

- Greedy partitioning algorithm
  - Top-down exploration of lattice of hierarchy cuts
  - Local recoding → each equivalence class uses its own hierarchy cut
  - Much faster than bottom-up algorithm using global recoding
  - Lower information loss (NCP) than bottom-up algorithm

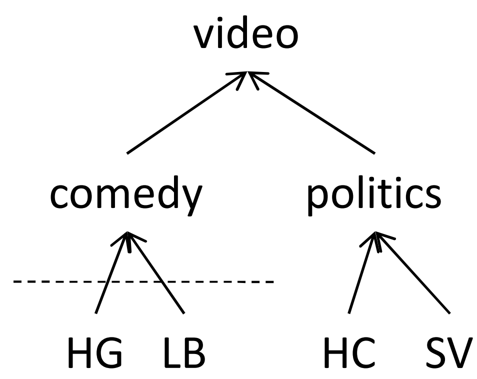
Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{LB}	{LB}
U2	4/13/86	F	53715	{HC, SV}	{HC, SV}
U3	2/28/76	M	53703	{HG}	{HG}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{LB, HG}	{LB, HG}
U6	2/28/76	F	53706	{LB, HG}	{LB, HG}



k-Anonymization [He Naughton 09]

- Greedy partitioning algorithm
  - Top-down exploration of lattice of hierarchy cuts
  - Local recoding → each equivalence class uses its own hierarchy cut
  - Much faster than bottom-up algorithm using global recoding
  - Lower information loss (NCP) than bottom-up algorithm

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{comedy}	{comedy}
U2	4/13/86	F	53715	{HC, SV}	{HC, SV}
U3	2/28/76	M	53703	{comedy}	{comedy}
U4	1/21/76	M	53703	{HC, SV}	{HC, SV}
U5	4/13/86	F	53706	{LB, HG}	{LB, HG}
U6	2/28/76	F	53706	{LB, HG}	{LB, HG}



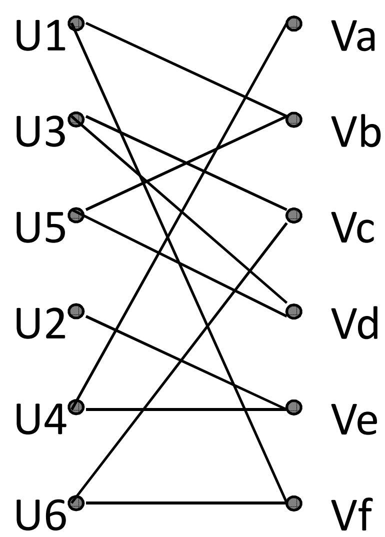
k^m-/k-Anonymization and Uncertainty

- Intuition: A k^m-/k-anonymized table T' represents the set of all "possible world" tables T_i s.t. T' is a k^m-/k-anonymization of T_i
- The table T from which T' was originally derived is one of the possible worlds
- Answer queries by assuming that each specialization of a generalized value is equally likely

Graph (Multi-Tabular) Data Example

- Video data recording videos viewed by users

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



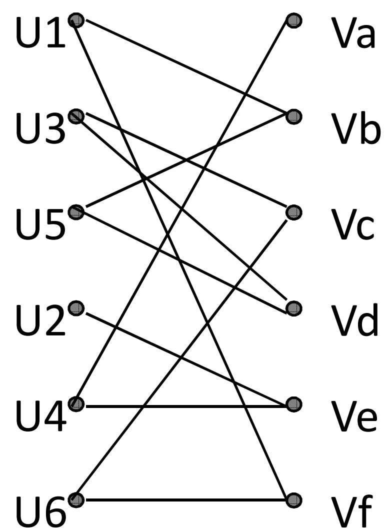
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

- Similar associations arise in medical data (Patient, Symptoms), search logs (User, Keyword)

(k, l)-Anonymity [Cormode+ 08]

- No *a priori* distinction between public and private videos
- Intuition: retain graph structure, permute entity→node mapping
  - Adding, deleting edges can change graph properties

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706

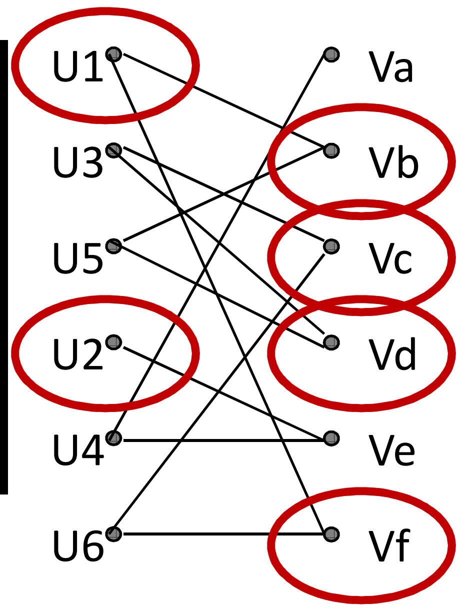


Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

(k, l)-Anonymity [Cormode+ 08]

- Assumption: publishing censored graph does not violate privacy
- Censored graph of limited utility to answer queries
  - Average number of comedy videos viewed by users in 53715? 1

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

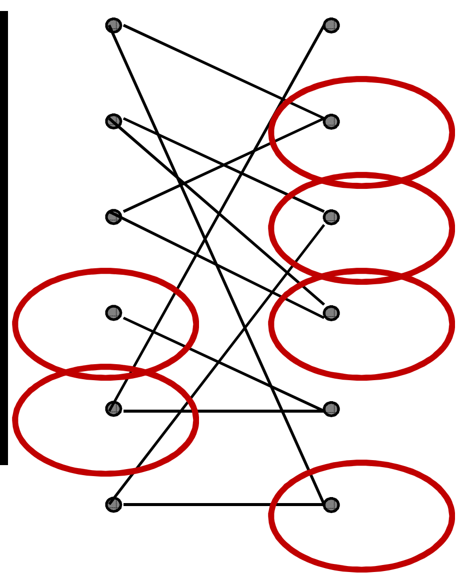
at&t

(k, l)-Anonymity [Cormode+ 08]

- Assumption: publishing censored graph does not violate privacy

- Censored graph of limited utility to answer queries
  - Average number of comedy videos viewed by users in 53715? 0

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706

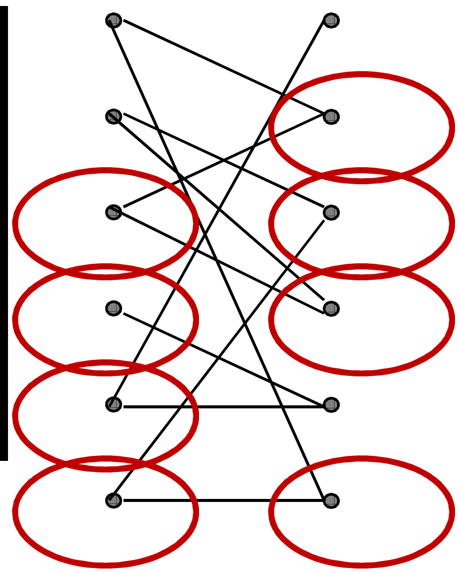


Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

(k, l)-Anonymity [Cormode+ 08]

- Assumption: publishing censored graph does not violate privacy
- Censored graph of limited utility to answer queries
  - Average number of comedy videos viewed by users in 53715? 2

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706

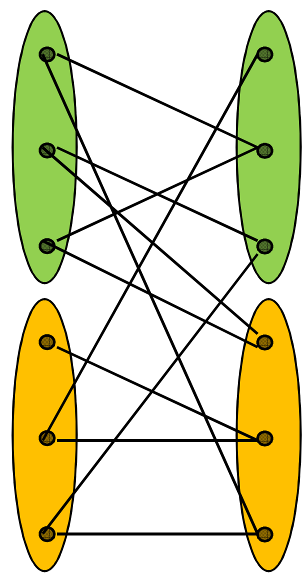


Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

(k, l)-Anonymity [Cormode+ 08]

- Goal: Improve utility: (k, l) grouping of bipartite graph (V, W, E)
  - Partition V (W) into non-intersecting subsets of size ≥ k (l)
  - Publish edges E' that are isomorphic to E, where mapping from E to E' is anonymized based on partitions of V, W

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

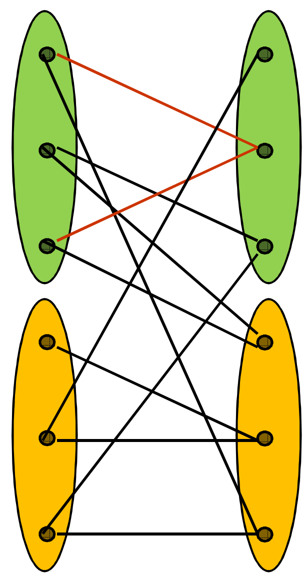
(k, l)-Anonymity [Cormode+ 08]

- Issue: some (k, l) groupings (e.g., local clique) leak information
  - Low density of edges between pair of groups not sufficient
  - Low density may not be preserved after few learned edges
- Solution: safe (k, l) groupings
  - Nodes in same group of V have no common neighbors in W
  - Requires node and edge sparsity in bipartite graph
- Properties of safe (k, l) groupings
  - Safe against static attacks
  - Safe against attackers who know a limited number of edges

(k, l)-Anonymity [Cormode+ 08]

- Safe (k, l) groupings
  - Nodes in same group of V have no common neighbors in W
  - Essentially a diversity condition
- Example: unsafe (3, 3) grouping

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



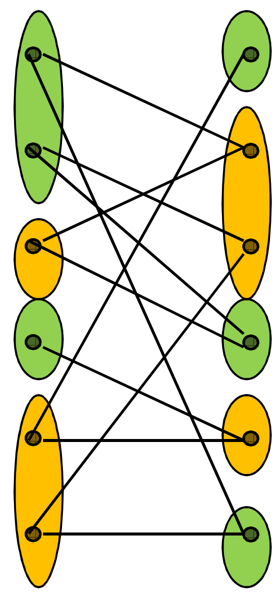
Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

at&t

(k, l)-Anonymity [Cormode+ 08]

- Safe (k, l) groupings
  - Nodes in same group of V have no common neighbors in W
  - Essentially a diversity condition
- Example: safe (3, 3) grouping

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

(k, l)-Anonymity [Cormode+ 08]

- Static Attack Privacy: In a safe (k, l) grouping, there are k*l possible identifications of entities with nodes and an edge is in at most a 1/max(k, l) fraction of such possible identifications
  - Natural connection to Uncertainty

- Learned Edge Attack Privacy: Given a safe (k, l) grouping, if an attacker knows $r < \min(k, l)$ true edges, the most the attacker can infer corresponds to a $(k - r, l - r)^{*(r, r)}$ -grouped graph

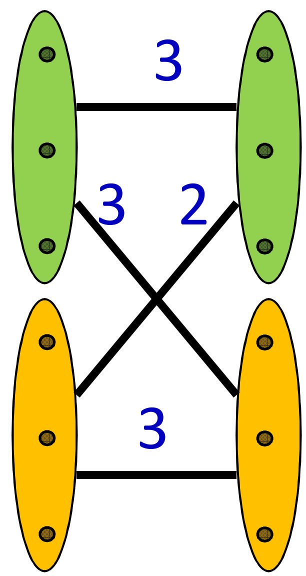
(k, l)-Anonymity [Cormode+ 08]

- Type 0 queries: answered exactly
- Theorem: Finding the best upper and lower bounds for answering a Type 2 aggregate query is NP-hard
  - Upper bound: reduction from set cover
  - Lower bound: reduction from maximum independent set
- Heuristic for Type 1, 2 queries
  - Reason with each pair of groups, aggregate results
  - Complexity is $O(|E|)$

Partition [Hay+ 08]

- Partition nodes into groups as before
- Publish only number of edges between pairs of groups

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706



Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy

Partition and Uncertainty

- Encodes a larger space of possible worlds than (k, l)-anonymity
  - Removes information about correlation of edges with nodes

- Increased privacy: identifying node does not identify other edges

- Reduced utility: more variance over possible worlds
  - Accuracy lower than for corresponding (k, l)-anonymization

Other Graph Anonymization Techniques

- Much recent work on anonymizing social network graph data
  - [Backstrom+ 07] study active, passive attacks on fully censored data
  - [Narayanan+ 09] link fully censored data with public sources
  - [Zhou+ 08] define privacy based on one-step neighborhood
  - [Zhou+09] define privacy based on full node reachability graph
  - [Korolova+ 08] analyze attacks when attacker “buys” information
  - [Zheleva+ 07] use machine learning to infer sensitive edges

- Topic of continued interest to the community
  - More papers in ICDE 2010 and beyond…

Outline

Part 1

- Introduction to Anonymization and Uncertainty
- Tabular Data Anonymization

Part 2

- Set and Graph Data Anonymization
- **Models of Uncertain Data**
- Query Answering on Anonymized Data
- Open Problems and Other Directions

Representing Uncertainty in Databases

- Almost every DBMS represents some uncertainty…
  - NULL can represent an unknown value
- Foundational work in the 1980s
  - Work on (possibilistic) c-tables [Imielinski Lipski 84]
- Resurgence in interest in recent years
  - For lineage and provenance
  - For general uncertain data management
  - Augment possible worlds with probabilistic models

Uncertain Database Systems

- Uncertain Databases proposed for a variety of applications:
  - Handling and querying (uncertain, noisy) sensor readings
  - Data integration with (uncertain, fuzzy) mappings
  - Processing output of (uncertain, approximate) mining algorithms
- To this list, we add anonymized data
  - A much more immediate application
  - Generates new questions and issues for UDBMSs
  - May require new primitives in systems

Conditional Tables

- Conditional Tables (c-tables) form a powerful representation
  - Allow variables within rows
  - Each assignment of variables to constants yields a possible world
  - Extra column indicates condition that row is present
  - May have additional global conditions

DOB	Sex	ZIP	Salary	Condition
1/21/76	M	**X**	50,000	**true**
Y	F	53715	55,000	$(Y=4/13/86) \vee (Y=1/21/76)$
2/28/76	M	53703	**Z**	$Z \in \{55{,}000, 60{,}000, 65{,}000\}$
Y	M	**W**	6000	$W \neq X \wedge (Y=4/13/86)$

at&t

Conditional Tables

- C-tables are a very powerful model
  - Conditions with variables in multiple locations become complex
  - Even determining if there is one non-empty world is NP Hard
  - Anonymization typically results in more structured examples
- Other simpler variations have been proposed
  - Limit where variables can occur (e.g. only in conditions)
  - Limit clauses to e.g. only have (in)equalities
  - Only global, no local conditions
- C-tables with Boolean variables only in conditions are complete
  - Capable of representing any possible set of base tables

Probabilistic c-tables

- Can naturally add probabilistic interpretation to c-tables
  - Specify probability distributions over variables

DOB	Sex	ZIP	Salary	Condition
1/21/76	M	**X**	50,000	**true**
Y	F	53715	55,000	**$(Y=4/13/86) \vee (Y=1/21/76)$**
2/28/76	M	53703	**Z**	**$Z \in \{55,000, 60,000, 65,000\}$**
Y	M	**W**	6000	**$W \neq X \wedge (Y=4/13/86)$**

z	Pr[Z=z]
55,000	0.2
60,000	0.6

x	Pr[X=x]
53703	0.5
53715	0.5

- Probabilistic c-tables are complete for distributions over tables
  - Also closed under relational algebra
  - Even when variables restricted to boolean

Uncertain Database Management System

- Several systems for working with uncertain data
  - TRIO, MayBMS, Orion, Mystiq, BayesStore, MCDB…
- Do not always expose a complete model to users
  - Complete models (eg probabilistic c-tables) hard to understand
  - May present a “working model” to the user
  - Working models can still be closed under a set of operations
- Working models specified via tuples and conditions
  - Class of conditions defines models
  - E.g. possible existence; exclusivity rules

Working Models of Uncertain Data

- Attribute-level uncertainty
  - Some attributes within a tuple are uncertain, have a pdf
  - Each tuple is independent of others in same relation
- Tuple-level uncertainty
  - Each tuple has some probability of occurring
  - Rules define mutual exclusions between tuples
- More complex graphical models have also been proposed
  - Capture correlations across values in a tuple, or across tuples
- General models
  - Can represent any distribution by listing probability for each world
  - May be large and unwieldy in the worst case

MayBMS model (Cornell/Oxford)

- U-relational database, using c-tables with probabilities [AJKO 08]
  - No global conditions, only local conditions of form X=c (var=const)
  - Only consider set valued variables

DOB	Sex	ZIP	Salary	Prob	Condition
1/21/76	M	53715	50,000	1	X=1
4/13/86	F	53715	55,000	0.5	Y=1
1/21/76	F	53715	55,000	0.5	Y=2
2/28/76	M	53703	55,000	0.6	Z=1
2/28/76	M	53703	60,000	0.6	Z=1

y	Pr[Y=y]
1	0.5
2	0.5

  - Probability of a world is product of tuple probabilities
  - Any world distribution can be represented via correlated tuples
- Possible query answers found exactly, probabilities approximated

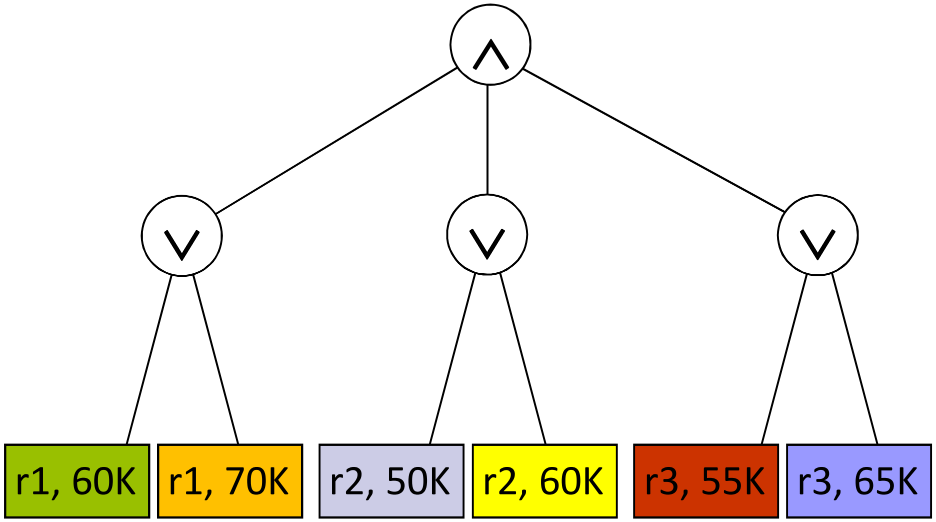
Trio Model (Stanford)

- Some certain attributes, others specified as alternatives [BSHW 06]

ZIP	Sex	(DOB, Salary)
53715	M	(1/21/76, 50,000) : 1
53715	F	(4/13/86, 55,000) : 0.5 || (1/21/76, 55,000) : 0.5
53703	M	(2/28/76, 55,000) : 0.2 || (2/28/76, 60,000) : 0.6

- Last column gives joint distribution of uncertain attributes
  - Attribute level uncertainty model
- System tracks the *lineage* of tuples in derived tables
  - Similar to the conjunction of variable assignments in a c-table

AND/XOR model (Maryland)



- Tree representation of data
  - Leaves are possible tuples
  - Internal nodes are ANDs or XORs
- Easy to compute probabilities in this model [Li Deshpande 09]
  - Based on use of generating functions
- Can easily encode moderately complex correlations of tuples
  - Still not completely natural to capture e.g. bijection semantics

Other systems

- MYSTIQ (U. Washington)
  - Targeted at integrating multiple databases
- Orion (Purdue)
  - Explicit support for continuous dbns as attributes
- MCDB (Florida)
  - Monte Carlo approach to query answering via “tuple bundles”
- BayesStore (Berkeley)
  - Sharing graphical models (Bayesian networks) across attributes

Summary of Uncertain Databases

- Anonymization is an important source of uncertain data
  - Seems to have received only limited attention thus far
- Complete models can represent any possible dbn over tables
  - Probabilistic c-tables with boolean variables in conditions suffice
- Simpler "working models" adopted by nascent systems
  - Offering discrete dbns over attribute values, presence/absence
- Exact (aggregate) querying possible, but often approximate
  - Approximation needed to avoid exponential blow-ups

- Our focus: representing and querying anonymized data
  - Identifying limitations of existing systems for this purpose

Outline

Part 1

- Introduction to Anonymization and Uncertainty
- Tabular Data Anonymization

Part 2

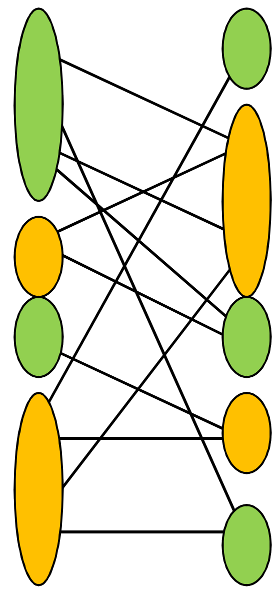
- Set and Graph Data Anonymization
- Models of Uncertain Data
- **Query Answering on Anonymized Data**
- Open Problems and Other Directions

Monte Carlo Methods

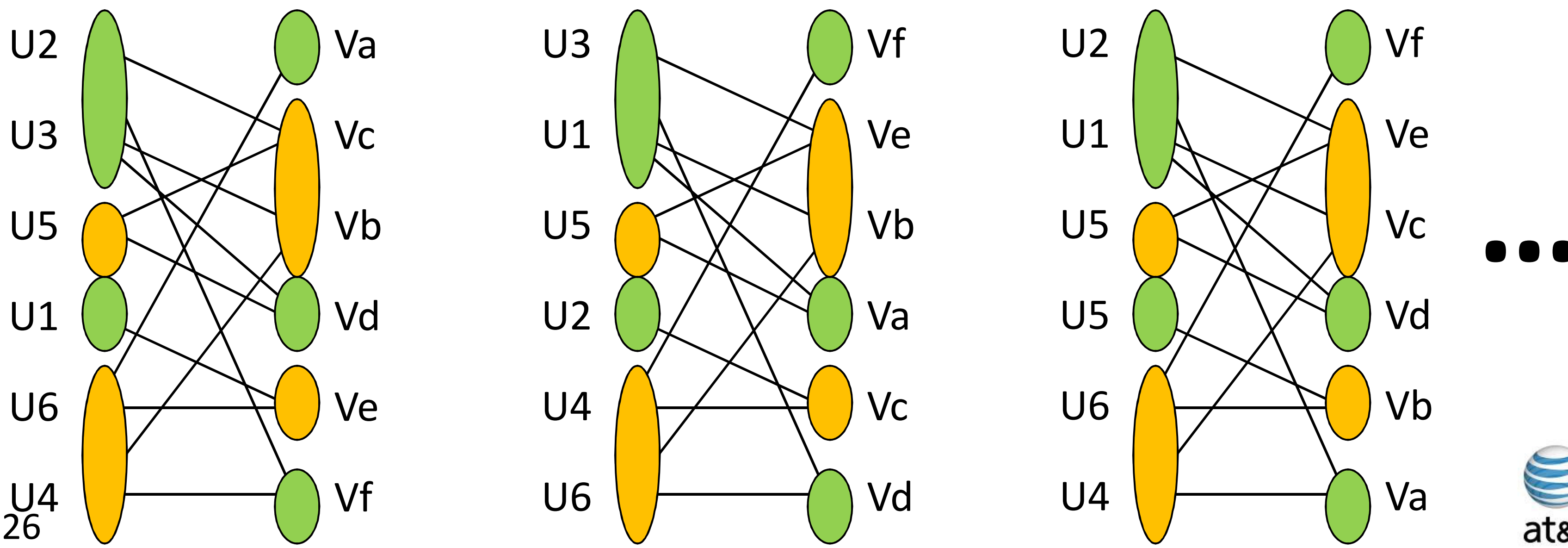
- Efficient approximations given by generic Monte-Carlo approach
  - Sample N possible worlds according to possible world dbn
  - Evaluate query on each possible world
- Distribution of sample query answers approximates true dbn
  - Average of sample query answers gives mean (in expectation)
  - Median, quantiles of sample answers behave likewise
- Can bound accuracy of these estimates:
  - Pick $N = O(1/\varepsilon^2 \log 1/\delta)$ for parameters ε, δ
  - Sample median corresponds to $(0.5 \pm \varepsilon)$ quantile w/prob $1-\delta$
  - Cumulative distributions are close: $\forall x.\ |F(x) - F_{sample}(x)| < \varepsilon$

Monte Carlo Example on Graph Data

Uid	DOB	Sex	ZIP
U1	1/21/76	M	53715
U2	4/13/86	F	53715
U3	2/28/76	M	53703
U4	1/21/76	M	53703
U5	4/13/86	F	53706
U6	2/28/76	F	53706

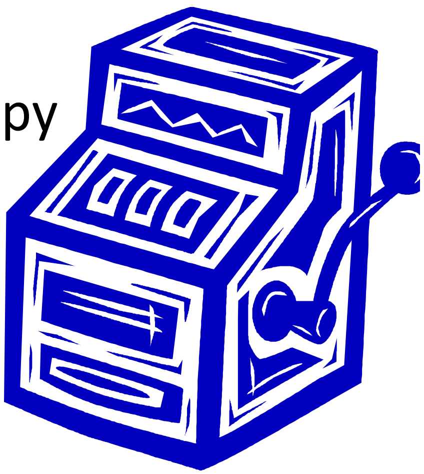


Vid	Title	Genre
Va	hanging chads	politics
Vb	apartment	comedy
Vc	holy grail	comedy
Vd	incredibles	comedy
Ve	stolen votes	politics
Vf	life of brian	comedy



Monte Carlo Efficiency

- Naively evaluating query on N sampled worlds can be slow
  - N typically 10s to 1000s for high accuracy
- Can exploit redundancy in the sample
  - If same world sampled many times, only use one copy
  - Scale estimates accordingly
- MCDB [JPXJWH '08]: Monte Carlo Database
  - Tracks sample as "bundle of tuples" for efficiency
  - Evaluates query only once over all sampled tuples
  - Postpones sampling from parametric dbns as long as possible
  - Significant time savings possible in practice



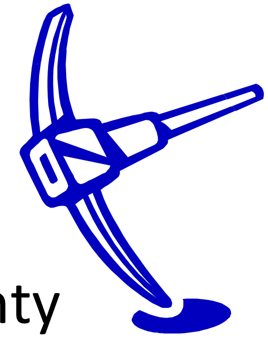
Karp-Luby

- Uniform sampling gives bad estimates for unlikely events
  - A given tuple may appear in very few sampled worlds
- For tuple conditions specified in Disjunctive Normal Form
  - $C_1 \vee C_2 \vee \ldots C_m$ for clauses $C_i = (l_1 \wedge l_2 \wedge \ldots)$
- Karp-Luby alg approximates no. of satisfying assignments [KL83]
  - Let S_i denote set of satisfying assignments to clause C_i
  - Sample clause i with probability $|S_i|/\sum_{i=1}^{m} |S_i|$
  - Uniformly sample an assignment τ that satisfies C_i
  - Compute $c(\tau)$ = number of clauses satisfied by τ
  - Estimate $X(\tau) = \sum_{i=1}^{m} |S_i| \,/\, c(\tau)$

Karp-Luby analysis

- $E[X(\tau)]$ is number of satisfying assignments
- Variance is bounded: $Var[X(\tau)] \leq m^2 E^2[X(\tau)]$
- Taking the mean of $O(m^2/\varepsilon^2)$ estimates gives $(1 \pm \varepsilon)$ approx
  - Gives relative error, not additive error (better for small probs)
- Used in MayBMS system for estimating confidence of tuples
  - Accounts for the different (overlapping) conditions for presence

Mining Anonymized Data



- Most mining problems are well-defined with uncertainty
  - Correspond to an optimization problem over possible worlds
- Can hope for accurate answers despite anonymization
  - Mining looks for global patterns, which have high support
  - Ideally, such patterns will not be scrubbed away
- Data mining on uncertain data needs new algorithms
  - Recall, motivation for anonymization is to try new analysis
- Monte Carlo approach not always successful
  - How to combine results from multiple sampled worlds?

Association Rule Mining

- A natural mining problem on transaction data
  - Find pattern of items which imply a common consequent
  - Only want to find patterns with high support and confidence
- Publishing exact association rules can still be privacy revealing
  - E.g. If $AB \Rightarrow C$ has high confidence, and C is sensitive
  - E.g. If $A \Rightarrow C$ and $AB \Rightarrow C$ have almost same confidence, may deduce that $A \neg C \Rightarrow B$ has low support, high confidence
- Two approaches to ensure privacy:
  - Anonymize first, then run ARM on anonymized data
  - Extract exact rules, but then anonymize rules [ABGP 08]

ARM example

Uid	DOB	Sex	ZIP	Public	Private
U1	1/21/76	M	53715	{Ap, LB, In}	{ }
U2	4/13/86	F	53715	{Ap, LB}	{SV}
U3	2/28/76	M	53703	{HG, FW}	{ }
U4	1/21/76	M	53703	{LB, In}	{HC, SV}
U5	4/13/86	F	53706	{Ap, In}	{ }
U6	2/28/76	F	53706	{HG, FW}	{ }

- (k,h,p) anonymization was designed to be "ARM-friendly"
  - Some items have been suppressed so will not appear in rules
  - Support of other items unchanged, so same rules can be found
  - E.g. Can recover Ap → In with conf 2/3, support 1/2

Summary of Query Answering

- A variety of techniques for general query answering
  - Monte-Carlo, Karp-Luby
- Mining anonymized data needs new algorithms
  - Due to the additional uncertainty in the data
  - Can adapt previously known methods
- Much scope for work targeting querying anonymized data
  - No systems yet support arbitrary aggregations on such data

Outline

Part 1

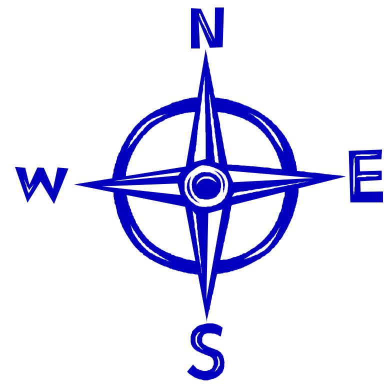
- Introduction to Anonymization and Uncertainty
- Tabular Data Anonymization

Part 2

- Set and Graph Data Anonymization
- Models of Uncertain Data
- Query Answering on Anonymized Data
- **Open Problems and Other Directions**

Open Problems and Other Directions

- This section: a variety of other ideas and directions
  - Outline only (a slide or two per idea)

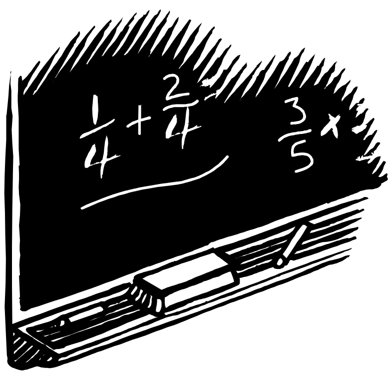


More integration into systems

- Explicit support for anonymized data in UDBMSs
  - Have tried to make the case in this tutorial
  - Some new primitives/syntactic sugar may be needed
- Motivates more attention on aggregate querying and mining
  - Analysis beyond standard SQL primitives
  - Support for top-k, mining operations
- Motivates operations that *add* uncertainty to data
  - Only MayBMS and MCDB talk about adding uncertainty
  - Places whole process (generation, modelling, usage) in DMBS

Formal Reasoning

- Formal reasoning about anonymity via uncertainty
- Can privacy requirements be translated into formal statements over uncertain data?
- Some possible goals:
  - Formulate a query to measure privacy (and utility) in a given uncertain table in some high level language
  - Run query on a certain table to output uncertain table with specified privacy guarantees



Put theory into practice

- Need to see more positive examples of anonymization
- Unfortunately, bad examples are easier to remember
  - AOL Search data still high in people's minds
  - People remember other controversies, not their resolution
  - Census data has been anonymized for years, without problem
- Still some nervousness about using anonymization
  - What if someone finds a new attack not thought of before?
- Attempts to standardize might help
  - New crypto standards are subject to intense scrutiny
  - Opportunity for new "challenges" (similar to KDD cup)

Cryptographic connections

- Conceptually cryptography connects to anonymization
  - Both concerned with privacy of individuals' data
- Cryptography feels more mature and field tested
  - Crypto methods in widespread use, foundation of e-commerce
- Additional visibility gives more confidence in security
  - Many eyes looking for flaws and weaknesses

- Can same approach be brought to anonymization?
  - Can an anonymization method be based on crypto assumptions?
  - Can break anonymization iff can break some encryption method

Differential Privacy

$X \leftrightarrow X'$

- Differential privacy gives stronger guarantee than others here
  - Take databases X, and X', which differ only in a single place
  - Differentially Private if $\Pr[\text{Output}(X)] \leq (1+\varepsilon) \Pr[\text{Output}(X')]$
- Very strong guarantee:
  - Even if attacker knows everything about X except one bit, the two possibilities look (approximately) equally likely
- Guarantee is achievable:
  - For some publishing some global aggregates
  - In some interactive querying settings
  - At great computational cost in other cases
- Merits a whole tutorial of its own [Smith 08]

Incremental Data Release

- May want to release new data as it is obtained
- Trivial approach: re-anonymize whole data set afresh
  - Vulnerable to attacks linking two versions of same data
- More complex: extend existing anonymization
  - Changes within a group may violate diversity requirements
  - Deletions from a group may reveal remaining tuples
- Example work: m-invariance [Xiao Tao 07]
  - Add counterfeit tuples so group distribution is invariant
  - Additional source of distortion in query answering

Geographic Data

- Increasing availability of location data from modern technology
  - Cell phones have cell tower, GPS information
- Current (and former) location can be very sensitive
  - Should a parent know exactly where their kids are?
  - Should someone know exactly where their partner is?
- Merits a whole tutorial of its own
  - "From data privacy to location privacy", [Liu, VLDB '07]
- Can adapt notions from tabular data (k-anonymity, l-diversity)
  - A natural generalization model replaces points with regions
- Question: how to include semantics of location privacy?
  - Locations may be distinct but close; dense or sparse regions

Temporal Data

- Time data can add an extra challenge for anonymization
  - Due to the semantics of time data as “domain knowledge”
  - E.g. an individual cannot be associated with a crime that happened prior to their date of birth
- Simple solutions: ensure that all temporal information is either identifying, or sensitive, but not both
  - Limits utility: essentially suppresses some time values
- More complex: additional constraint to prevent inference
  - More general question: how to model and prevent other inferences based on “domain knowledge”
  - E.g. individuals cannot travel 1000 miles in 10 minutes

Other structured data

- Easy to imagine other structured data needing anonymization
  - XML data, text data, image data, etc.
- In each case, need to work through a series of questions
  - For what reasons is anonymization needed?
  - What properties should be preserved by anonymization?
  - What is the form of domain and background knowledge?
  - What are limitations of applying existing anonymization methods?
  - What is a good measure of utility of resulting data?
  - What uncertainty model does this entail?
- May need deep connections to other areas
  - Text anonymization requires natural language processing

Conclusions

- Anonymized data leads to many complex questions
  - Connections to other areas, esp. uncertain data management
- Will lead to new research problems for years to come

- Full references in the slides

References

[ABGP 08] Maurizio Atzori, Francesco Bonchi, Fosca Giannotti, Dino Pedreschi: Anonymity preserving pattern discovery. VLDB J. 17(4): 703-727 (2008).

[Aggarwal 05] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB 2005.

[Aggarwal+ 05a] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu. Anonymizing tables. In ICDT 2005.

[Aggarwal+ 05b] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu. Approximation algorithms for k-anonymity. Journal of Privacy Technology, 2005.

[AGGN] Barbara Anthony, Vineet Goyal, Anupam Gupta, and Viswanath Nagarajan. A plant location guide for the unsure. In ACM-SIAM SODA 2008.

[AJKO 08] L. Antova, T. Jansen, C. Koch, and D. Olteanu. Fast and simple relational processing of uncertain data. In ICDE 2008.

References

[Backstrom+ 07] Lars Backstrom, Cynthia Dwork, Jon M. Kleinberg. Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In WWW 2007.

[Barbaro Zeller 06] Michael Barbaro, Tom Zeller Jr. A face Is exposed for AOL searcher No. 4417749. New York Times, August 9 2006. http://www.nytimes.com/2006/08/09/technology/09aol.html

[Bayardo+ 05] Roberto J. Bayardo Jr., Rakesh Agrawal. Data privacy through optimal k-anonymization. In ICDE 2005.

[BLR 08] Avrim Blum, Katrina Ligett, Aaron Roth. A learning theory approach to non-interactive database privacy. In ACM STOC 2008.

[BSHW 06] O. Benjelloun, A. D. Sarma, C. Hayworth, and J. Widom. An introduction to ULDBs and the Trio system. IEEE Data Engineering Bulletin, 29(1):5–16, Mar. 2006.

[Brickell Shmatikov 06] Justin Brickell, Vitaly Shmatikov. Efficient anonymity-preserving data collection. In ACM KDD 2006.

References

[CKP 04] Reynold Cheng, Dmitri V. Kalashnikov, Sunil Prabhakar: Querying imprecise data in moving object environments. IEEE Trans. Knowl. Data Eng. 16(9): 1112-1127 (2004).

[CLY 09] Graham Cormode, Feifei Li, Ke Yi. Semantics of ranking queries for probabilistic data and expected ranks. In ICDE 2009.

[Cormode+ 08] Graham Cormode, Divesh Srivastava, Ting Yu, and Qing Zhang. Anonymizing bipartite graph data using safe groupings. In VLDB 2008.

[Cormode Garofalakis 07] Graham Cormode, Minos N. Garofalakis. Sketching probabilistic data streams. In SIGMOD 2007.

[Cormode Garofalakis 09] Graham Cormode, Minos N. Garofalakis. Histograms and Wavelets on Probabilistic Data. In ICDE 2009.

[Cormode McGregor 08] Graham Cormode, Andrew McGregor. Approximation algorithms for clustering uncertain data. In ACM PODS 2008.

References

[Fung+ 05] B. Fung, K. Wang, P. Yu. Top-down specialization for information and privacy preservation. In ICDE 2005.

[Ghinita+ 08] Gabriel Ghinita, Yufei Tao, Panos Kalnis. On the anonymization of sparse high-dimensional data. In ICDE 2008.

[Guha Minagala 09] Sudipto Guha and Kamesh Munagala. Exceeding expectations and clustering uncertain data. In ACM PODS 2009.

[Hay+ 08] Michael Hay, Gerome Miklau, David Jensen, Donald F. Towsley, Philipp Weis. Resisting structural re-identification in anonymized social networks. In VLDB 2008.

[He Naughton 09] Anonymization of Set-Valued Data via TopDown, Local Generalization. In VLDB 2009

[HPZL 08] M. Hua, J. Pei, W. Zhang, and X. Lin. Efficiently answering probabilistic threshold top-k queries on uncertain data. In ICDE 2008.

[Imielinski Lipski 84] Tomasz Imielinski, Witold Lipski Jr. Incomplete information in relational databases. J. ACM 31(4): 761-791 (1984).

References

[Iyengar 02] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In ACM KDD 2002.

[JMMV 07] T. S. Jayram, A. McGregor, S. Muthukrishnan, and E. Vee. Estimating statistical aggregates on probabilistic data streams. In ACM PODS 2007.

[JPXJWH '08] R. Jampani, L. L. Perez, F. Xu, C. Jermaine, M. Wi, and P. Haas. MCDB: A monte carlo approach to managing uncertain data. In ACM SIGMOD 2008.

[Kifer '09] D. Kifer. Attacks on privacy and deFinetti's theorem. In SIGMOD 2009.

[KL83] R.M. Karp, M. Luby. Monte-Carlo algorithms for enumeration and reliability problems. In 24th Annual Symposium on Foundations of Computer Science, 1983.

[Korolova+ 08] Aleksandra Korolova, Rajeev Motwani, Shubha U. Nabar, Ying Xu. Link privacy in social networks. In CIKM 2008.

[LeFevre+ 05] Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan. Incognito: efficient full-domain k-anonymity. In ACM SIGMOD Conference 2005.

[LeFevre+ 06] Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In ICDE 2006.

References

[Li Deshpande 09] Jian Li, Amol Deshpande. Consensus Answers for Queries over Probabilistic Databases. In PODS 2009

[LSD 09] Jian Li, Barna Saha, Amol Deshpande. A Unified Approach to Ranking in Probabilistic Databases. In VLDB 2009.

[Li+ 07] Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian. t-closeness: privacy beyond k-anonymity and l-diversity. In ICDE 2007.

[Liu 07] Ling Liu. From data privacy to location privacy: models and algorithms. In VLDB 2007.

[Machanavajjhala+ 06] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam. l-Diversity: privacy Beyond k-anonymity. In ICDE 2006.

[Meyerson+ 04] Adam Meyerson, Ryan Williams. On the complexity of optimal k-anonymity. In ACM PODS 2004.

[Mokbel Chow Aref 07] Mohamed F. Mokbel, Chi-Yin Chow, Walid G. Aref. The new Casper: a privacy-aware location-based database server. In ICDE 2007.

References

[Narayanan+ 09] Arvind Narayanan, Vitaly Shmatikov. De-anonymizing social networks. In S&P 2009.

[Narayanan Shmatikov 08] A. Narayanan, V. Shmatikov. Robust de-anonymization of large sparse datasets (How to break anonymity of the Netflix prize dataset). In S&P 2008.

[Samarati 01] Pierangela Samarati. Protecting respondents' identities in microdata release. In IEEE Trans. Knowl. Data Eng. 13(6): 1010-1027 (2001).

[Samarati Sweeney 98] Pierangela Samarati, Latanya Sweeney. Generalizing data to provide anonymity when disclosing information. In ACM PODS 1998.

[SIC 07] Mohamed A. Soliman, Ihab F. Ilyas, and Kevin C.-C. Chang. Top-k query processing in uncertain databases. In IEEE ICDE 2007.

[Smith 08] Adam Smith. Pinning down “privacy" in statistical databases. http://www.cse.psu.edu/~asmith/talks/diff-priv-March-18-2008.pdf 2008.

[Sweeney 02] Latanya Sweeney. k-Anonymity: a model for protecting privacy. International journal of uncertainty, fuzziness, and knowledge-based systems 2002.

References

[Terrovitis+ 08] Manolis Terrovitis, Nikos Mamoulis, Panos Kalnis. Privacy-preserving anonymization of set-valued data. In VLDB 2008.

[Winkler 02] William Winkler. Using simulated annealing for k-anonymity. Technical Report, U.S. Census Bureau.

[Wong+ 07] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In VLDB 2007.

[Xiao+ 06] Xiaokui Xiao, Yufei Tao. Anatomy: simple and effective privacy preservation. In VLDB 2006.

[Xiao Tao 07] Xiaokui Xiao, Yufei Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In SIGMOD 2007.

[Xu+ 08] Yabo Xu, Ke Wang, Ada Wai-Chee Fu, Philip S. Yu. Anonymizing transaction databases for publication. In ACM KDD 2008.

[YLSK 08] Ke Yi, Feifei Li, Divesh Srivastava, and George Kollios. Efficient processing of top-k queries in uncertain databases with x-relations. IEEE TKDE, vol. 20, no. 12, pp. 1669–1682, 2008.

References

[YZW 05] Zhiqiang Yang, Sheng Zhong, Rebecca N. Wright. Anonymity-preserving data collection. In ACM KDD 2005.

[Zhang+ 07] Qing Zhang, Nick Koudas, Divesh Srivastava, Ting Yu. Aggregate query answering on anonymized tables. In ICDE 2007.

[Zheleva+ 07] Elena Zheleva, Lise Getoor. Preserving the privacy of sensitive relationships in graph data. In PinKDD 2007.

[Zhou+ 08] Bin Zhou, Jian Pei. Preserving privacy in social networks against neighborhood attacks. In ICDE 2008.