Anonymization and Uncertainty in Social Network Data

Graham Cormode

graham@research.att.com

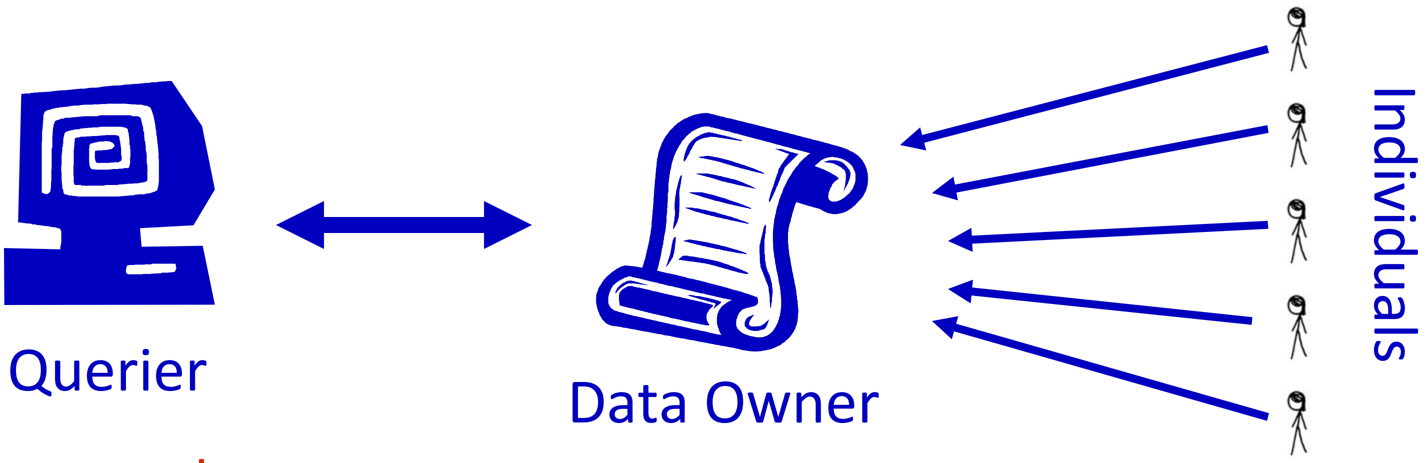
Joint work with Smriti Bhagat,
Balachander Krishnamurthy, Divesh Srivastava

at&t

Introduction to Privacy

- People have an inherent right and expectation of privacy
- "Privacy" is a complex concept (beyond the scope of this talk)
  - What exactly does "privacy" mean? When does it apply?
- 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
  - US: HIPAA, Video Privacy Protection
  - European: Data Protection Acts

Approaches to Privacy

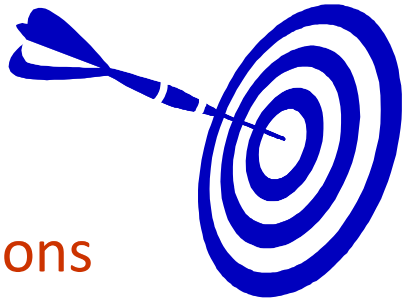


- Cryptography
  - Data hidden from anyone without the private key
- Private Information Retrieval
  - Querier gets accurate data, server learns nothing about query
- Statistical Databases
  - Queries are answered with noise to protect individuals
- **Anonymization**
  - Dataset is “published” with some information masked

Why Anonymize?

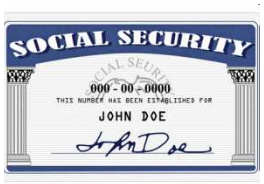
- For Data Sharing
  - Give real(istic) data to others to study without compromising privacy of individuals in the data
  - Allows third-parties to try new analysis and mining techniques not thought of by the data owner
- For Data Retention and Usage
  - Various requirements prevent companies from retaining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

Objectives for Anonymization



- Prevent (high confidence) inference of associations
  - Prevent inference of salary for an individual in census data
  - Prevent inference of individual’s activity in social network data
- 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)

Trivial Anonymization


SOCIAL SECURITY

000-00-0000

THIS NUMBER HAS BEEN ESTABLISHED FOR

JOHN DOE

- Trivial anonymization strips out identifying fields
  - Social Security Numbers (SSNs) and other unique identifiers
  - I.e. remove the keys to making “joining” impossible
- Unfortunately, this is not enough to prevent reidentification
  - Additional fields in the data can still identify individuals
  - DOB+Sex+ZIP unique for most US Residents [Sweeney 02]
- Trivial anonymization criticized from legal perspective [Ohm 09]

Examples of Anonymization

- US Census: information about every US household
  - Who, where; age, gender, racial, income and educational data
  - aggregated to regions (Zip code), released in full after 72 years
- Netflix: 100M ratings from 480K users to 18K movies
  - All direct customer information removed
  - Only subset of all data; dates modified; some ratings deleted
- AOL: 20M search queries for 650K users from 2006
  - Searches from same user linked by an arbitrary identifier
  - Many successful attacks identified individual users

Database Research and Anonymization

- [SIGMOD 2010] “Research papers will be judged … through double-blind reviewing”
- [TODS] Authors need only apply 6 simple steps to blind their submission:
  1. Anonymize the title page
  2. Remove mention of funding sources and personal acknowledgments
  3. Anonymize references found in running prose that cite your papers
  4. Anonymize citations of submitted work in the bibliography
  5. Ambiguate statements on systems that identify an author
  6. Name your files with care, document properties are also anonymized
- How can this anonymization method be attacked?

Attacking Paper Anonymization

- Identify a previously published conference paper with the same title
- Search for unusual sentences in the text/abstract to see if they appear in other papers or talk abstracts
- Build a database of bibliographies from papers, and try to match typos/misspellings

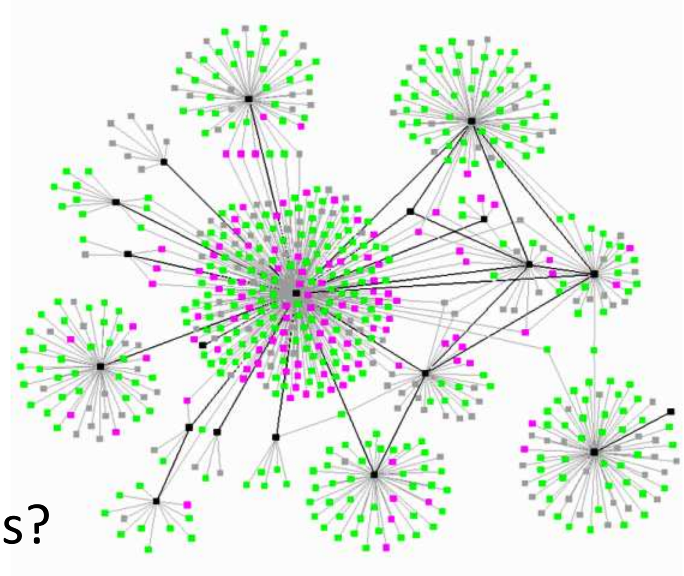
- **BONUS QUESTION**: How can authors de-anonymize their reviews to find out who killed their paper?

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…
  - – Measure “information loss” as a function of the masking used
  - – For a fixed query set, can look at max, average distortion
- ♦ Opinion: broad empirical evaluation gives a good baseline
  - – Defining measures of information loss are unconvincing (to me)

Graph Data

- Much data is best represented as a *graph* of interactions
  - Phone call data, network traffic etc.
  - Social network data



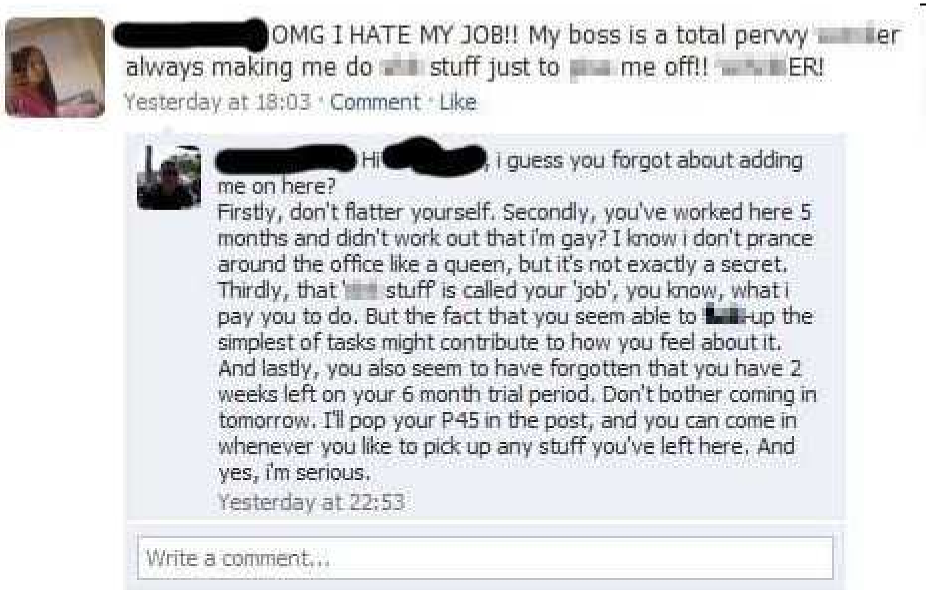
- Many natural queries on graphs:
  - How many nodes in subpopulations? (age range, location, interest groups)
  - What subpopulations are interacting?
  - Can graph be partitioned with few cuts?
  - How do these patterns change over time?
- Treating graph as a table *fails* to retain graph properties

Social Network Data

- Social Networks store much detailed personal information
  - Hundreds of millions use Facebook, LinkedIn, MySpace etc.
  - Demographic information about individuals, their likes, dislikes
  - Link information: friendship links, “wall posts”, comments etc.
  - Great source of data for research currently held by few firms
- …but isn’t all data on social networks public already?
  - No! Most social networks have privacy settings
  - Many other “private” social networks e.g. email data etc.
  - Many examples of unwanted revelation of personal data

Woman 'sacked' on Facebook for complaining about her boss after forgetting she had added him as a friend

By JULIE MOULT
Last updated at 12:26 AM on 15th August 2009



[redacted] OMG I HATE MY JOB!! My boss is a total pervvy [redacted]er always making me do [redacted] stuff just to [redacted] me off!! [redacted]ER!

Yesterday at 18:03 · Comment · Like

[redacted] Hi [redacted], i guess you forgot about adding me on here?
Firstly, don't flatter yourself. Secondly, you've worked here 5 months and didn't work out that i'm gay? I know i don't prance around the office like a queen, but it's not exactly a secret. Thirdly, that '[redacted] stuff' is called your 'job', you know, what i pay you to do. But the fact that you seem able to [redacted]-up the simplest of tasks might contribute to how you feel about it. And lastly, you also seem to have forgotten that you have 2 weeks left on your 6 month trial period. Don't bother coming in tomorrow. I'll pop your P45 in the post, and you can come in whenever you like to pick up any stuff you've left here. And yes, i'm serious.

Yesterday at 22:53

News 〉 Technology 〉 Facebook

Fugitive caught after updating his status on Facebook

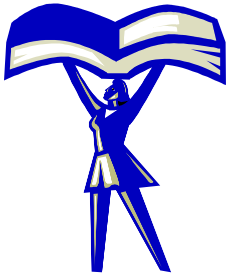
Maxi Sopo told his Facebook friends, including a former justice department official, he was living in paradise in Mexico

Although Sopo's profile was set to private, his list of friends was not. Scoville started combing through it and was surprised to see that one friend listed an affiliation with the justice department. He sent a message requesting a phone call.

Privacy and Utility Goals

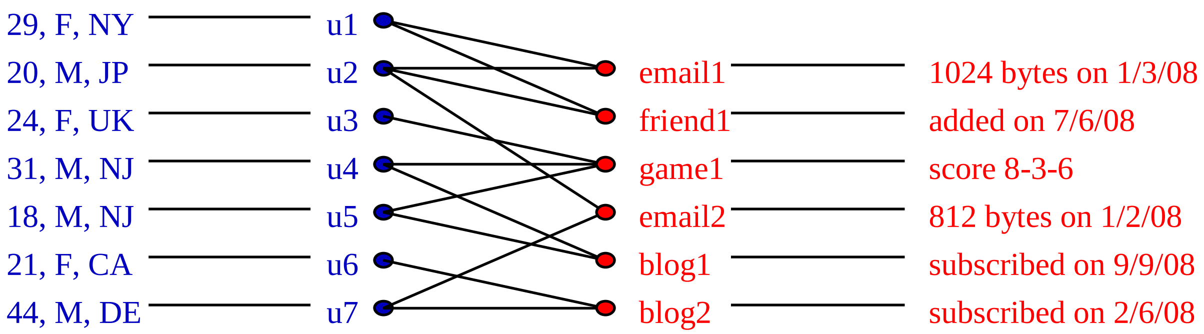
- Examples show information that users did not want disclosed
  - In those cases, it was leaked by the users themselves
- In publishing, want to avoid similar disclosures
  - Prevent knowledge of links between individuals
  - E.g. prevent user finding connections between their “friends”
- Still want researchers to be able to use the published data
  - Find global trends in the data, common patterns in the graph
  - Aggregate queries about neighborhoods

Recent Work



- Graph anonymization now a "hot topic"
- Some negative results:
  - Powerful attacker with much knowledge can reidentify some nodes [Backstrom, Dwork, Kleinberg 07; Narayanan, Shmatikov 09]
- Modification methods: add and remove edges
  - Make neighborhoods similar [Zhou, Pei 08; Liu, Terzi 08; Zou et al 09]
- Grouping methods: mask by grouping
  - Model attacker as Machine Learning alg [Zheleva, Getoor 07]
  - Group nodes and hide mapping from nodes to entities [Hay et al 08; Cormode et al 08, 09]

Interaction Graphs

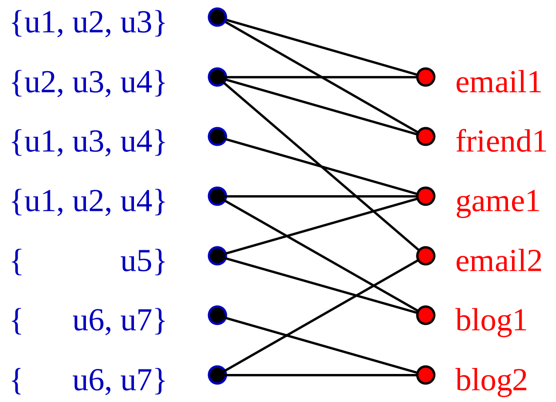


29, F, NY — u1
20, M, JP — u2
24, F, UK — u3
31, M, NJ — u4
18, M, NJ — u5
21, F, CA — u6
44, M, DE — u7

email1 — 1024 bytes on 1/3/08
friend1 — added on 7/6/08
game1 — score 8-3-6
email2 — 812 bytes on 1/2/08
blog1 — subscribed on 9/9/08
blog2 — subscribed on 2/6/08

- Represent social networks with an *interaction graph*
  - Entity nodes (with demographic properties) connect to Interaction nodes (with relevant properties)
  - Can represent (directed) pairwise and group interactions
- Anonymization requirement:
  should not be able to learn of the existence of any interaction
  - Quantify how much "background knowledge" needed to break

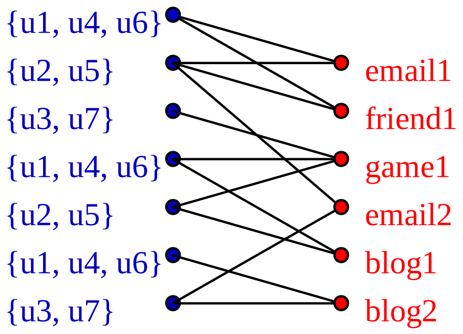
Label Lists



- Safety in numbers:
  Replace node ids with lists
- Cannot tell which is
  the true label of any node
- Aims: Preserve graph structure
  Still answer queries that touch many nodes
- Picking arbitrary lists is insufficient for security
  - As each node must appear exactly once, can eliminate options
  - In example, u1, u2, u3, u4 must be first four nodes
  - Reveals identity of some other nodes
  - Shows u6 and u7 share blog2 interaction

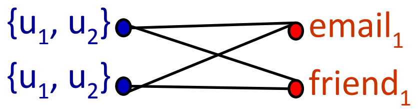
Uniform Lists

- Need more structure to the lists
  - Enforce symmetry so such deductions are not possible
- Divide nodes into classes of size *m*
- *Uniform lists* create symmetric lists in each class
  - *Full pattern*: all lists are $\{u_1, u_2 \ldots u_m\}$
  - Other symmetric patterns are possible
- Assign lists so each node's list includes true label

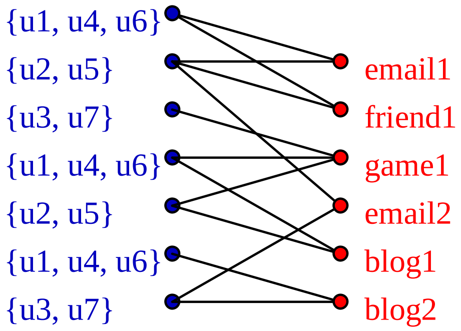


Security of published data

- Uniform lists still vulnerable if there are many interactions between nodes in same class

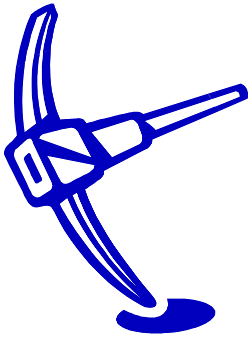


- Define a class safety condition
  - Each node v participates in interactions with at most one node in any other class
  - Keeps the inter-class interactions sparse



- Gives a provable guarantee of security
  - For classes of size ≥ m, for every way in which a node v is in an interaction, there are m-1 consistent ways where it is not
  - Based on considering structure of possible label assignments
  - Conclude attacker's belief in any possibility is at most 1/m

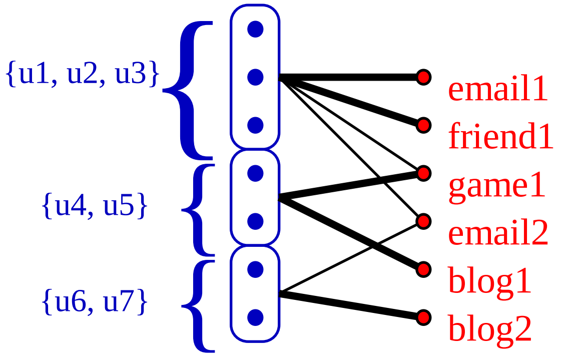
Background Knowledge Attacks



- What if an attacker knows some interactions?
  - E.g. knows Graham sent email to Torsten
- May be possible to identify certain nodes with those for which some partial knowledge exists
- Can prove that other nodes are not badly affected
  - Knowledge partitions some classes into pieces of size m-1 and 1
  - Safety condition still holds on the new classes
  - So previous guarantee holds with m-1
  - For r pieces of knowledge, inductively give guarantee with ≥ m-r
- For attacker with much information this may not be enough

Partitioning Approach

- Guarantees of the label list approach may not be enough
  e.g. lots of node degree information
- Can increase the amount of security, at cost of utility
- Partition approach:
  - Divide nodes into partitions of size ≥ m
  - Only reveal the number of edges from each partition to each interaction

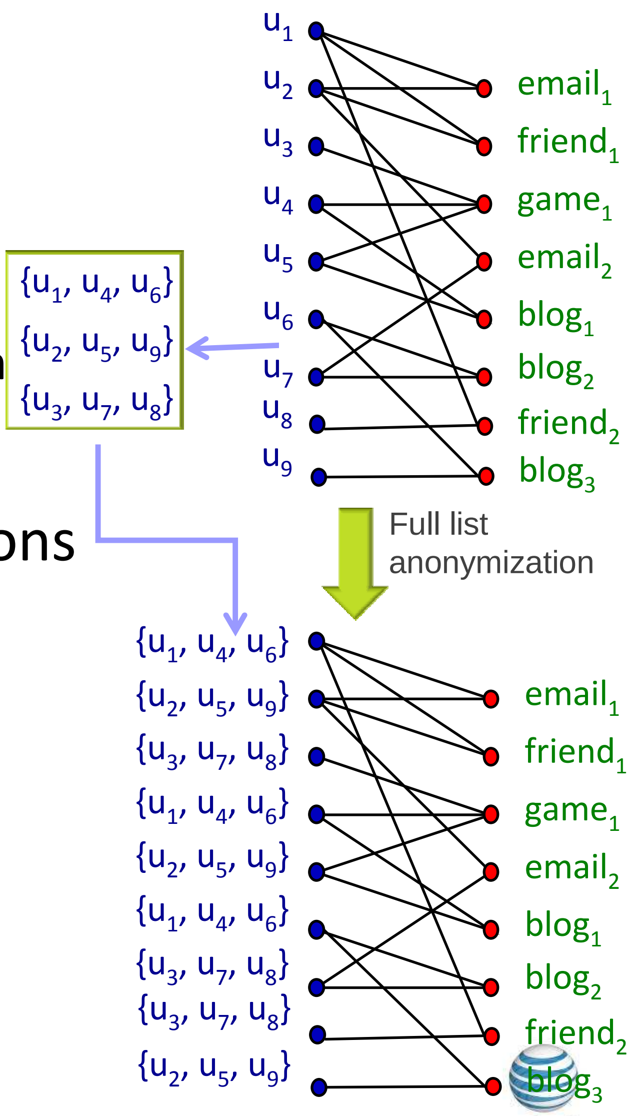


- Still need some conditions
  - Require same "safety condition" to ensure that attacker cannot use edge density to infer participation in interactions
  - Now can prove attacker with knowledge about < m entities cannot make any further inferences

Algorithm Overview

Multi-step process for label list anonymization:

1. Divide nodes into classes, respecting safety
   - Optimize: try to group similar nodes in same class
   - Sort or cluster, and try to add to each class in turn
2. Create and assign lists to nodes
3. Publish the graph using either lists or partitions
   - List: *k* labels at each node, preserving all edges
   - Partition: k labels at each node, aggregate edges



at&t

Query Answering

- A generic problem in data anonymization:
how to answer queries on anonymized data?
- Output data does not have same schema as input!
  - Have introduced uncertainties, replaced values with ranges
- Can give ad hoc solutions for particular queries
  - Unsatisfying and unscalable
- Look for a way to represent and query anonymized data
  - Anonymization adds uncertainty, so try uncertain DBMS?

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 working models

Possible Worlds Interpretation

- Uncertain anonymized data represents multiple possible worlds
  - Each possible world corresponds to a database (or graph, or…)
  - The original input data is known to be one of these worlds
  - Best approach to query answering: range over all possible worlds
- Possibilistic interpretations
  - Is a given fact possible (∃ a world W where it is true) ?
  - Is a given fact certain (∀ worlds W it is true) ?
- Probabilistic interpretations given distribution over worlds
  - 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?

Query Answering over Anonymized Data

- Represent anonymization in working model of UDBMs
  - More compact than a complete model
  - Ensure that there is a good match so result is compact
- Without other information, assume each world equally likely
  - With more information, can apply different prior beliefs
- Answer queries based on space of possible worlds
- A general Monte Carlo approach:
  - uniformly sample a possible world and evaluate query
  - Take the mean, max, min of multiple samples
- More sophisticated methods also possible

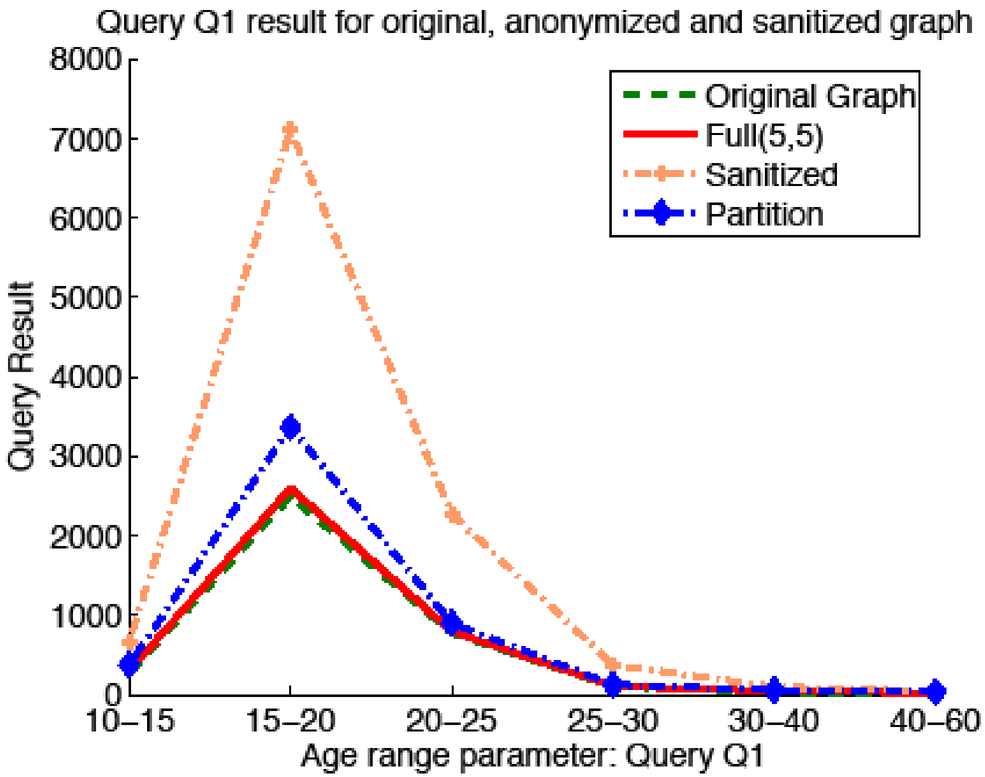
Experimental Analysis

- Data: evaluated our approaches over two social net datasets
  - Blog (780K nodes, 3M edges) drawn from Xanga network
  - Speed dating (530 nodes, 4K edges) due to Columbia study

- Utility Evaluation: on varied query workload
  - Pair: Single-hop, e.g., how many Americans befriend Germans
  - Trio: Two-hop, e.g., how many Americans are friends with Germans who have French friends
  - Triangle: Clustering co-efficient

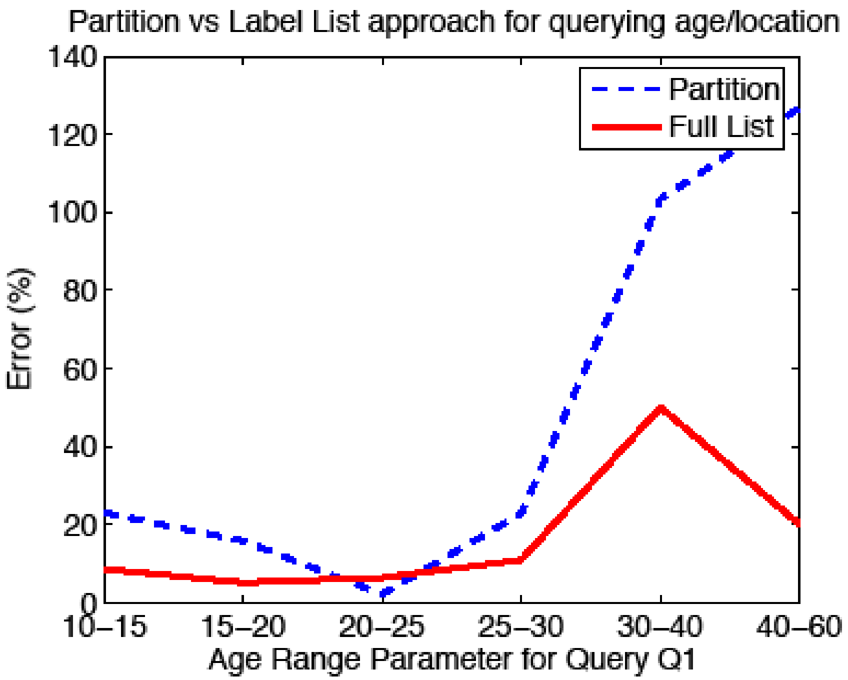
Experimental Analysis

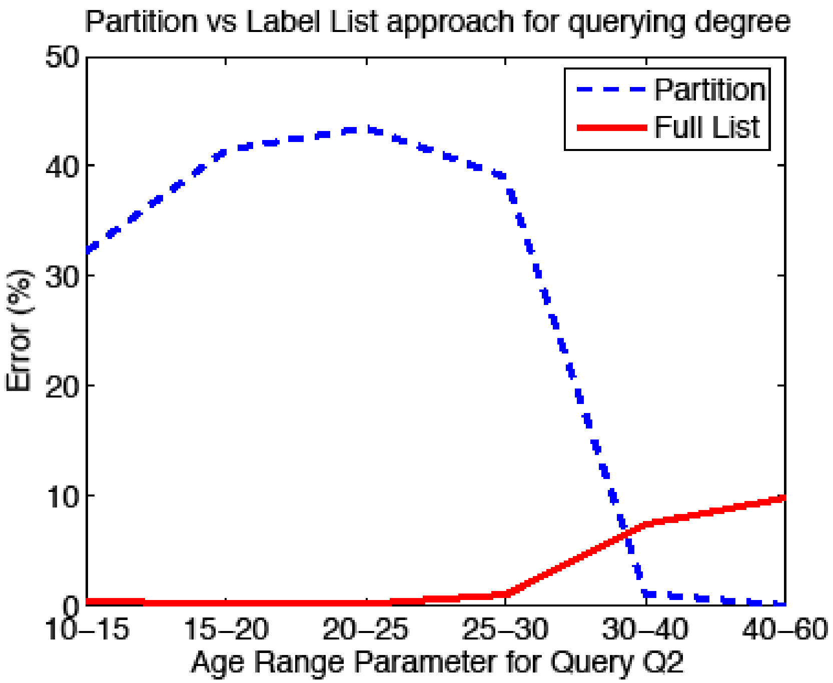
- Fixed privacy guarantee, analyzed impact on accuracy



- Pair Query (Q1): How many Americans of different ages are friends with Hong Kong residents with age <20?
- Errors are typically higher for partition approach

Experimental Analysis

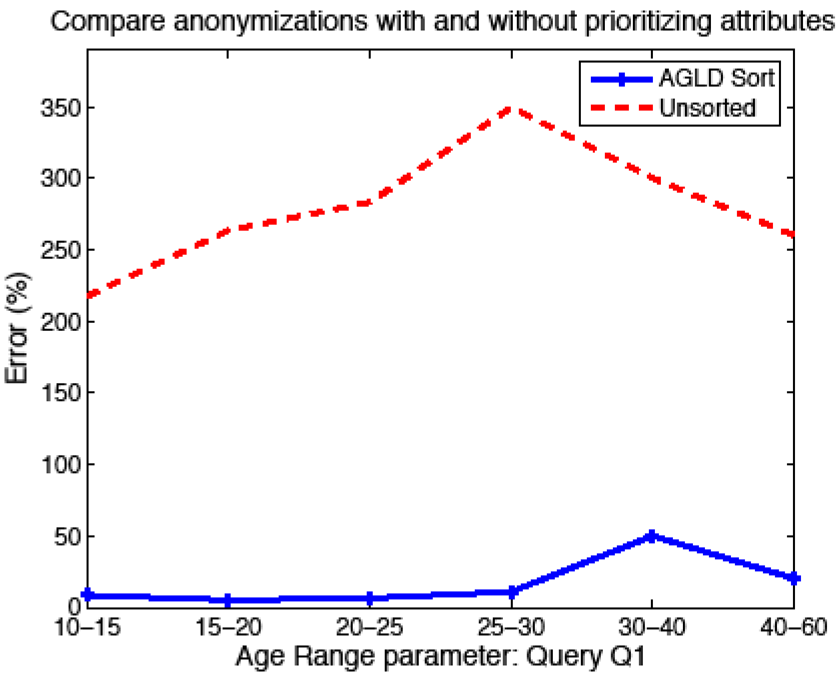

Partition vs Label List approach for querying age/location

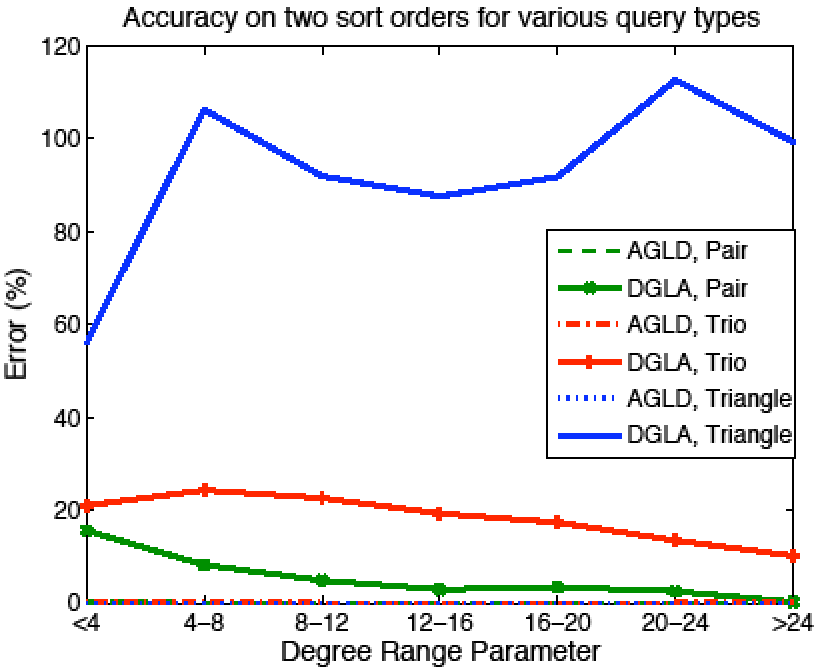

Partition vs Label List approach for querying degree

- Analysis on neighborhood size: higher accuracy on younger ages (more prevalent in data) than older
- Different group sizes vary privacy/accuracy tradeoff

Experimental Analysis

- Compare sort orders on Degree, Age, Gender, Location

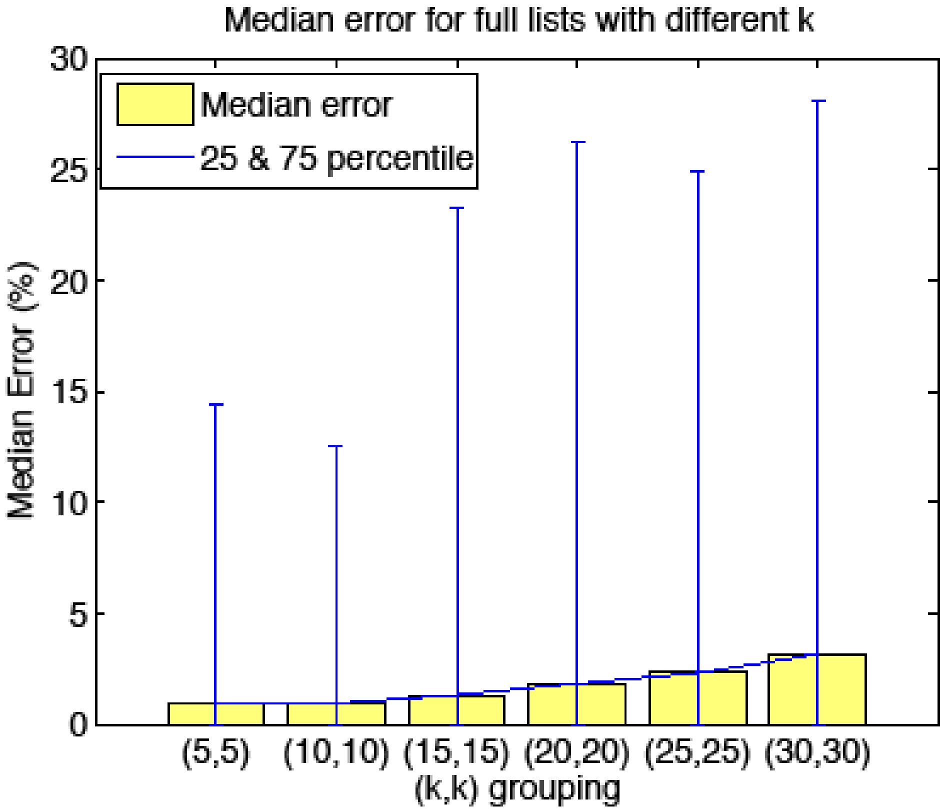




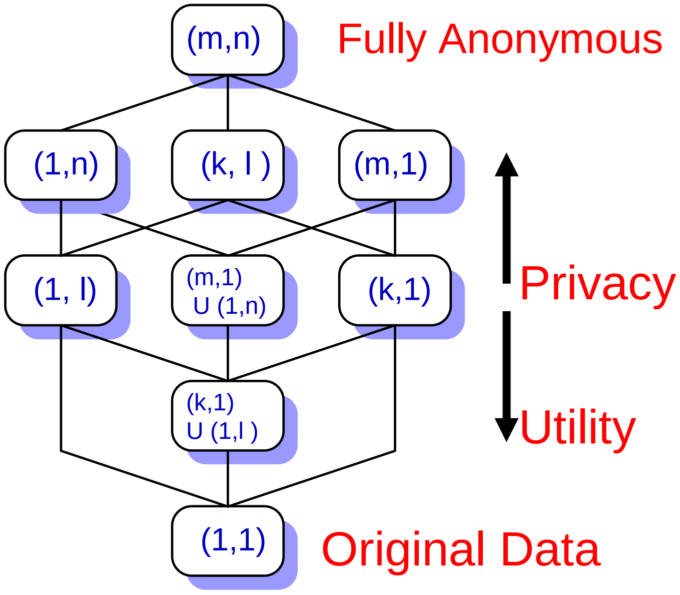
- Using a sort order clearly improves accuracy
- A sort order that matches workload helps more

Experimental Analysis

- Varied group sizes, analyzed impact on accuracy over a 100-query workload with all 3 types of queries



Future Work



- Work with time-varying graphs
  - How to "reanonymize" with new data?
- Allow different privacy requirements
  - Some association types are not private, some very sensitive
- Better understand link between uncertainty & anonymization
- Anonymize other structured data and represent in UDBMS:
  - Set valued data, free text data with links (hypertext)
  - Analysis-based view: association rules, clusters etc.

Conclusions

- Anonymization remains a challenging problem
  - Need to carefully study, what is threat model?
- Have proposed approaches for social network data
  - Offer different tradeoffs between utility and privacy
- Anonymization is an important source of uncertain data
  - Seems to have received only limited attention thus far
- Exact (aggregate) querying possible, but often expensive
  - Approximation needed to avoid exponential blow-ups

References: "Class-based graph anonymization for social network data" VLDB 09, "Anonymized data: Generation, models, usage" Tutorial in SIGMOD 09