

Class-based Graph Anonymization for Social Network Data

Smriti Bhagat (Rutgers University)

**Graham Cormode, Balachander Krishnamurthy,
Divesh Srivastava (AT&T Labs-Research)**

Key Takeaways

- ◆ Anonymization of (social) network data is a challenging problem
 - Goal: useful anonymization without altering network structure
- ◆ Contributions:
 - Label list and partition approaches: trade off privacy versus utility
 - Class safety conditions provide privacy guarantees
 - Accurate ad hoc aggregate analyses on real datasets

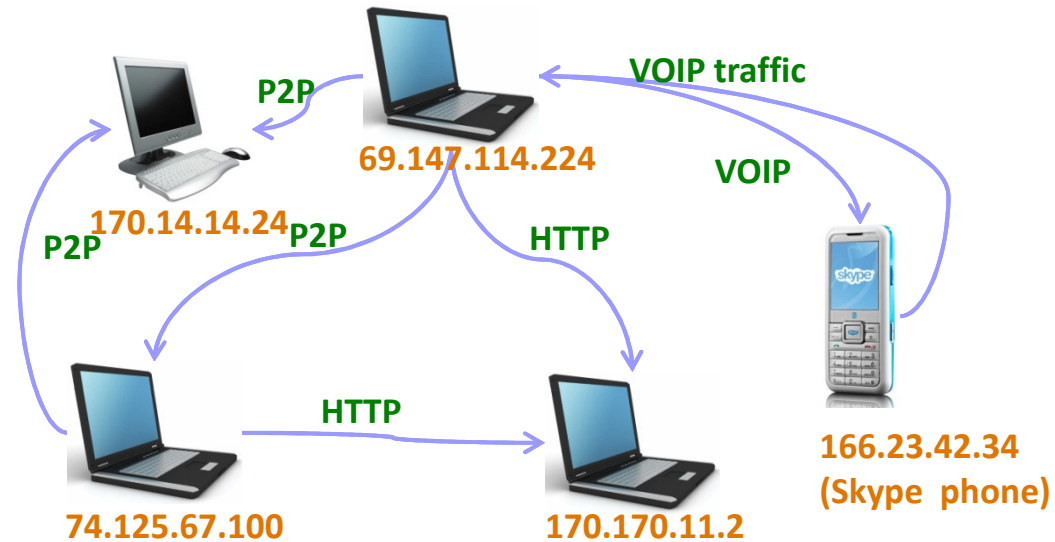
Outline

- ◆ Motivation
- ◆ Anonymization strategies
- ◆ Experiments: evaluation of utility

Anonymization Problem

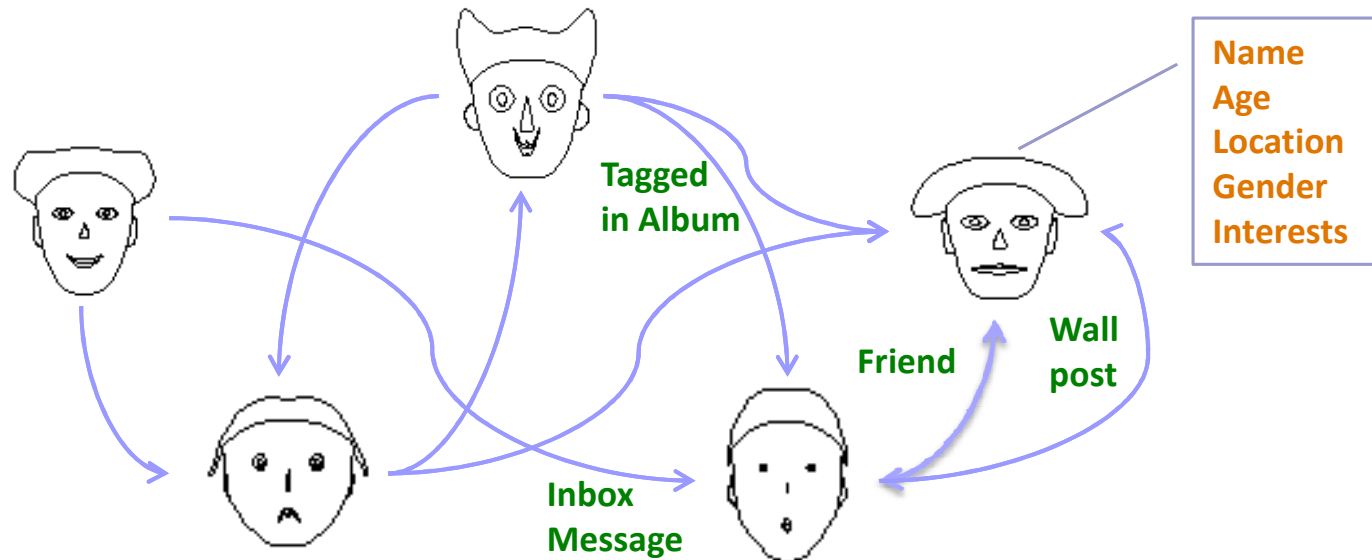
- ◆ Why anonymize?
 - Results of ad hoc data analysis have significant benefits to society
 - Privacy concerns, legal restrictions on what can be stored, shared
- ◆ Goal: useful anonymization of (social) network data
 - Cannot uniquely identify entities and their interactions
 - Accurate ad hoc aggregate analyses on the data
- ◆ Issue: simply removing unique identifiers is inadequate
 - NY Times identified specific individuals in AOL search logs

Example: IP Network Data



- ◆ Utility: analysis of specific traffic patterns such as P2P, MMORG
- ◆ Prior anonymization: prefix-preserving anonymization

Example: Social Network Data



- ◆ Utility: study interaction patterns among demographic groups
- ◆ Prior anonymization: modify graph to have structural similarity

Prior Work

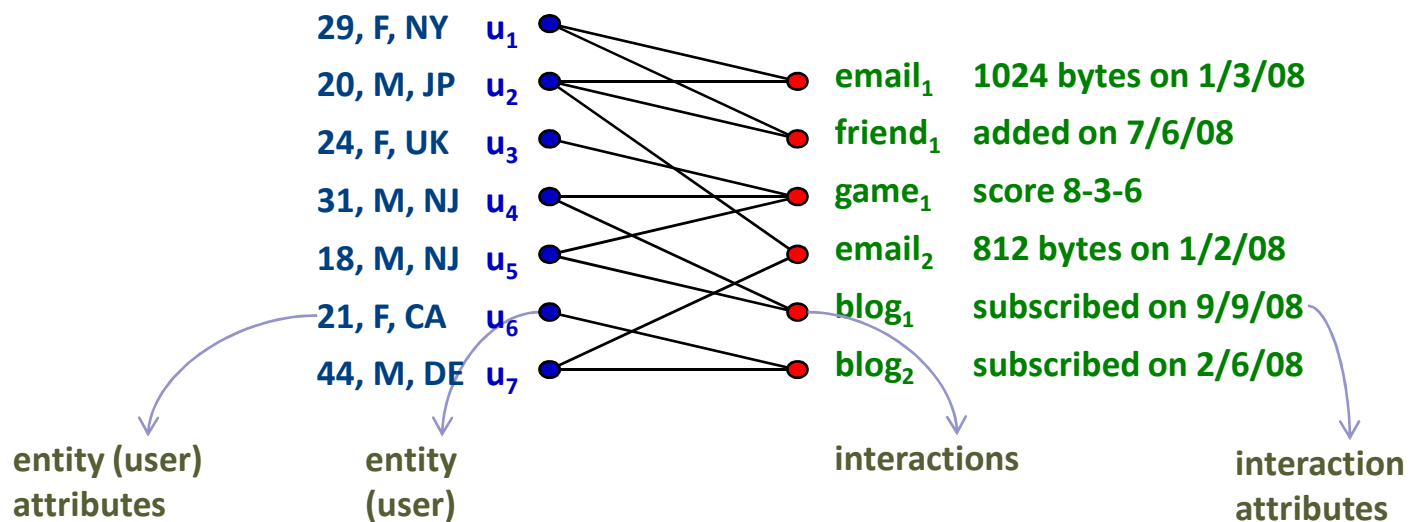
- ◆ Negative result: structural re-identification [BDK07]
 - Active and passive attacks can re-identify some nodes
- ◆ Alter graph to make neighborhoods similar [LT08, ZP08, ZCO09]
 - Graph alteration can impact graph-theoretic properties, utility
- ◆ Group nodes, hide the node → entity mapping [HJ+08, CS+08]
 - Techniques do not add or remove nodes or edges in the graph
 - Useful when active attacks are infeasible
 - Our approach builds on the techniques of [HJ+08, CS+08]

Outline

- ◆ Motivation
- ◆ Anonymization strategies
- ◆ Experiments: evaluation of utility

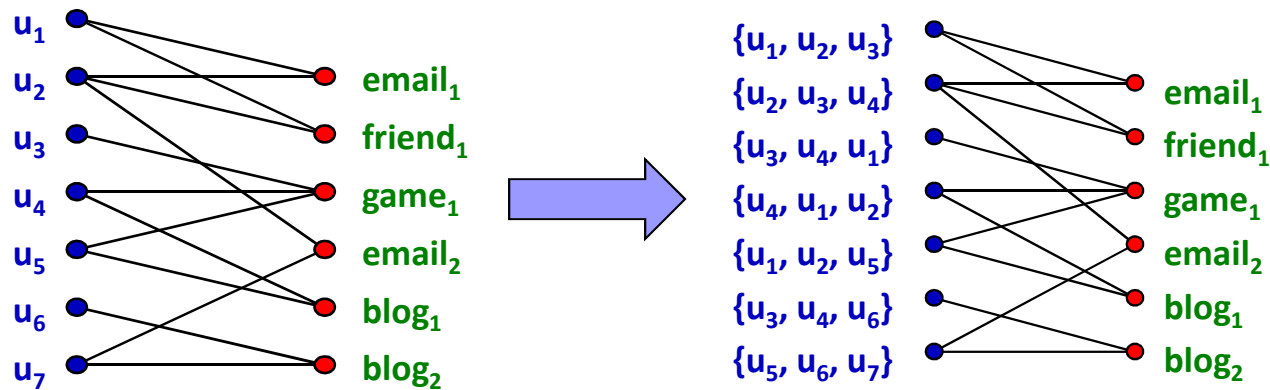
Representing Social Networks

- ◆ Social networks form rich communication hypergraphs
 - Interactions can involve sets of entities, not just pairs
 - Represented as bipartite graph, connecting entities to interactions



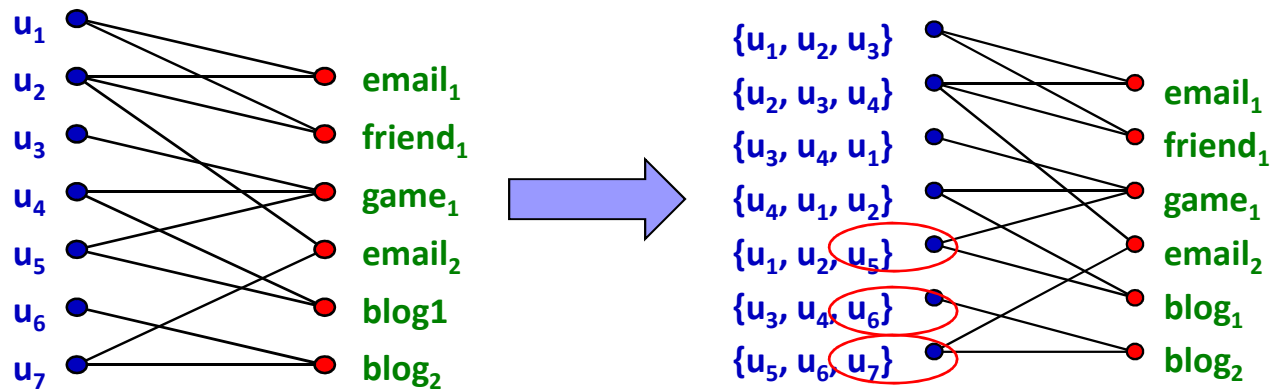
Label Lists

- ◆ Idea: replace node label with list containing true label [CS+08]
 - Safety in numbers: should not be able to tell true label of a node



Label Lists

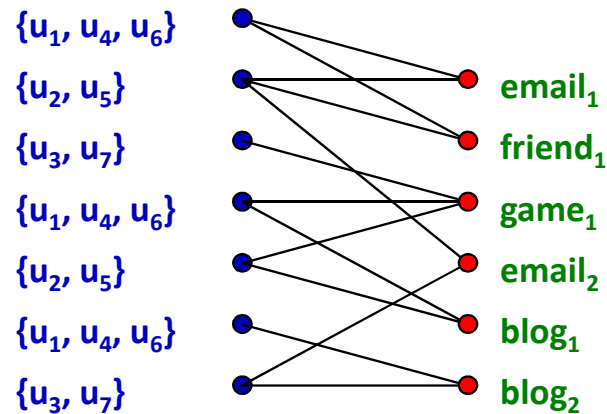
- ◆ Idea: replace node label with list containing true label [CS+08]
 - Safety in numbers: should not be able to tell true label of a node



- ◆ Picking arbitrary lists is inadequate for privacy: need structure
 - Easy to identify u_5, u_6, u_7 ; u_6 and u_7 share blog_2 interaction

Uniform Lists

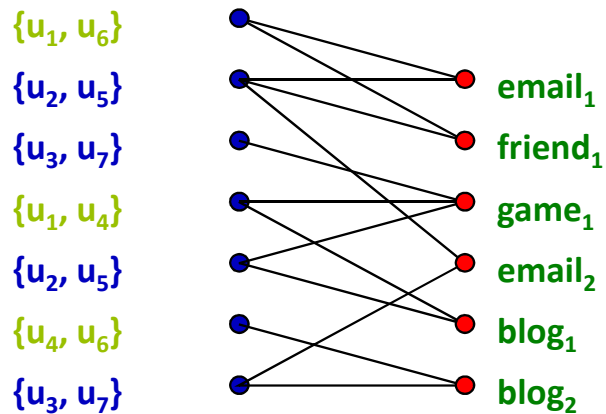
- ◆ Idea: enforce symmetry to avoid inference
 - Partition nodes into classes of size m : $\{u_1, u_2, \dots, u_m\}$
 - Create symmetric lists from each class: e.g., prefix list, full list
 - Assign lists to nodes such that node's label list includes true label



- ◆ Full label list

Uniform Lists

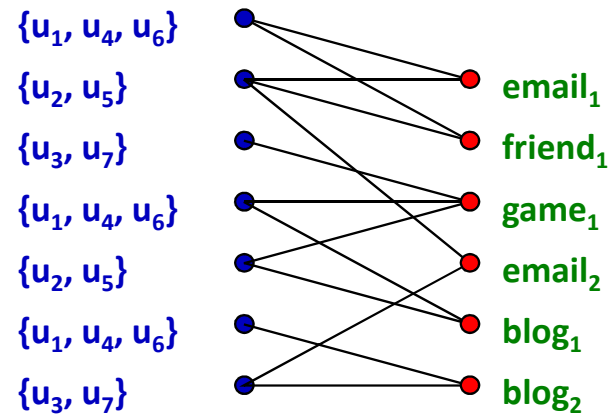
- ◆ Idea: enforce symmetry to avoid inference
 - Partition nodes into classes of size m : $\{u_1, u_2, \dots, u_m\}$
 - Create symmetric lists from each class: e.g., prefix list, full list
 - Assign lists to nodes such that node's label list includes true label



- ◆ Prefix label list

Class Safety Condition

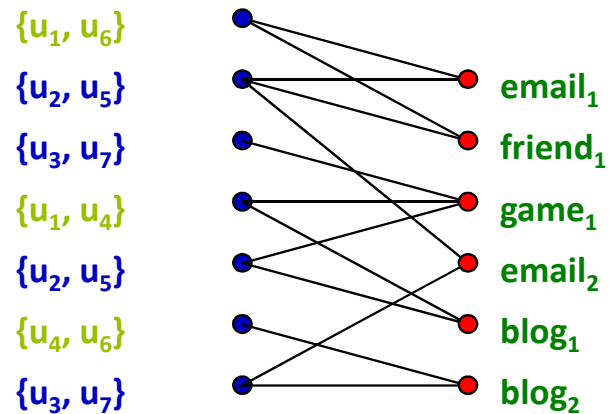
- ◆ Uniform lists vulnerable if interactions between classes is dense
- ◆ Class safety condition: to keep inter-class interactions sparse
 - Each node has interactions with at most one node in any class



- ◆ Full label list

Class Safety Condition

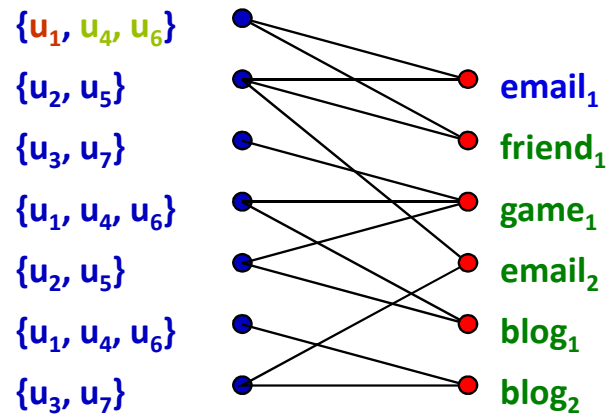
- ◆ Uniform lists vulnerable if interactions between classes is dense
- ◆ Class safety condition: to keep inter-class interactions sparse
 - Each node has interactions with at most one node in any class



- ◆ Prefix label list

Privacy Guarantees

- ◆ Theorem: for a class of size $\geq m$, and a node v in interaction i , there are at least $m-1$ possible label assignments where node v does not participate in interaction i .



Privacy Guarantees

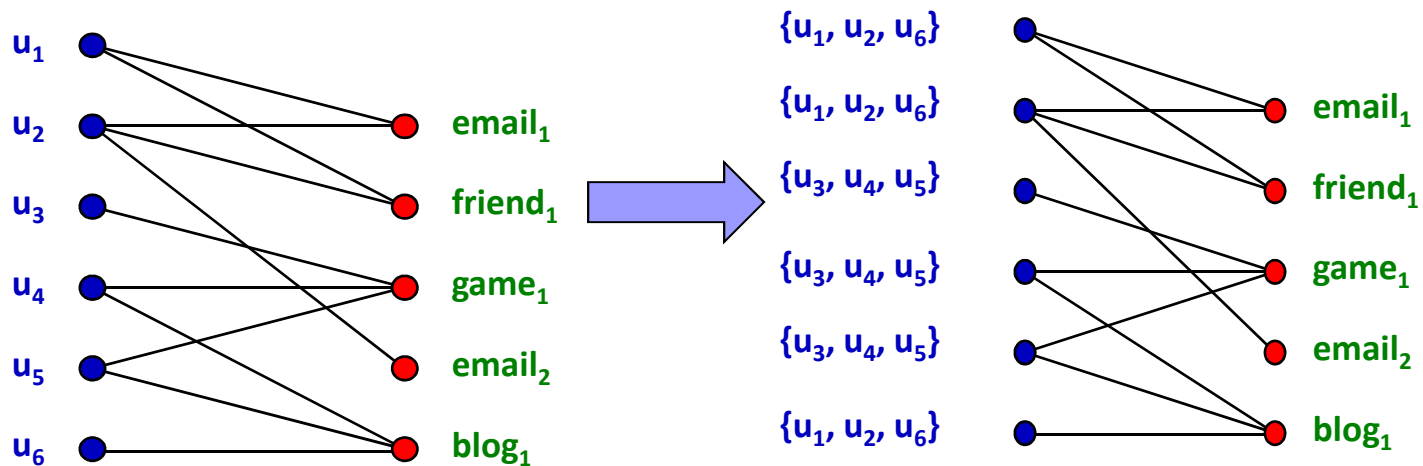
- ◆ What if an attacker knows a subset of interactions?
 - Possible to identify certain nodes associated with interactions
- ◆ What about other nodes in that class?
 - Known interactions partition class into classes of size 1 and $m-1$
 - Safety condition still holds on the subclass of size $m-1$

Anonymization Using Label Lists

- ◆ Partition nodes into classes, respecting safety condition
 - Sort nodes based on similarity of attributes to increase utility
- ◆ Create label lists for each class, assign lists to nodes
 - Requires some combinatorial exploration, in general
 - Easy for full lists: all nodes in a class get same label lists
- ◆ Publish graph with a list of at least m labels for each node, preserving all edge information

Anonymization Using Label Lists

- ◆ Partition nodes into classes ($m = 3$), respecting safety condition
 - Full lists: all nodes in a class get same label lists

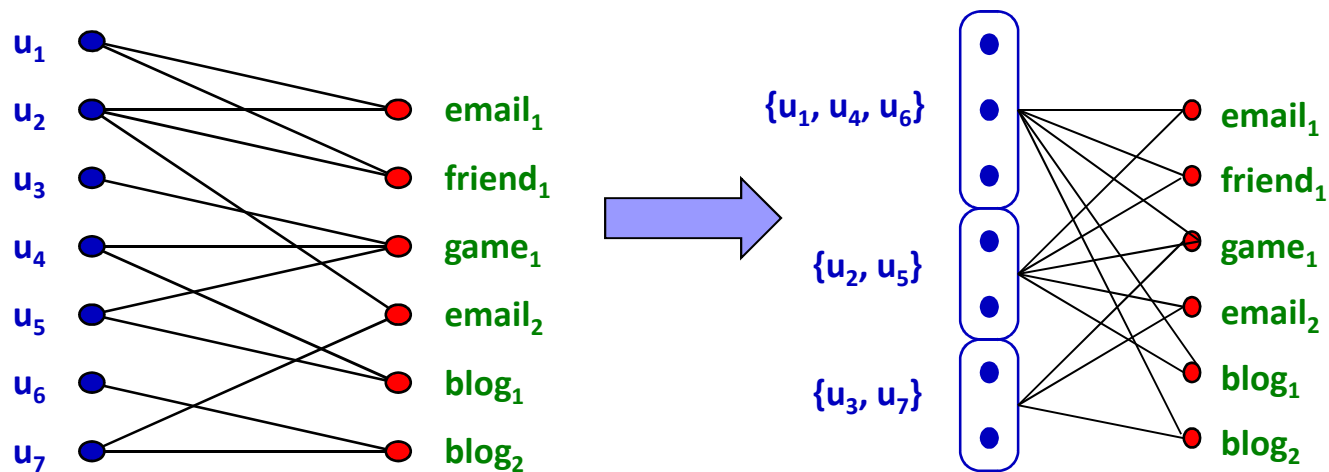


Partitioning Approach

- ◆ Guarantees of label lists inadequate if attacker has more info
- ◆ Partition approach: based on [HJ+08]
 - Partition nodes into classes of size m , respecting safety condition
 - Only reveal number of edges from each class to interaction nodes
- ◆ Better privacy guarantees than label list approach
 - Class safety condition ensures attacker cannot use edge density
 - Attacker who knows $< m$ entities cannot make further inferences

Partitioning Approach

- ◆ Partition nodes into classes ($m = 3$), respecting safety condition
 - Attacker who knows $< m$ entities cannot make further inferences

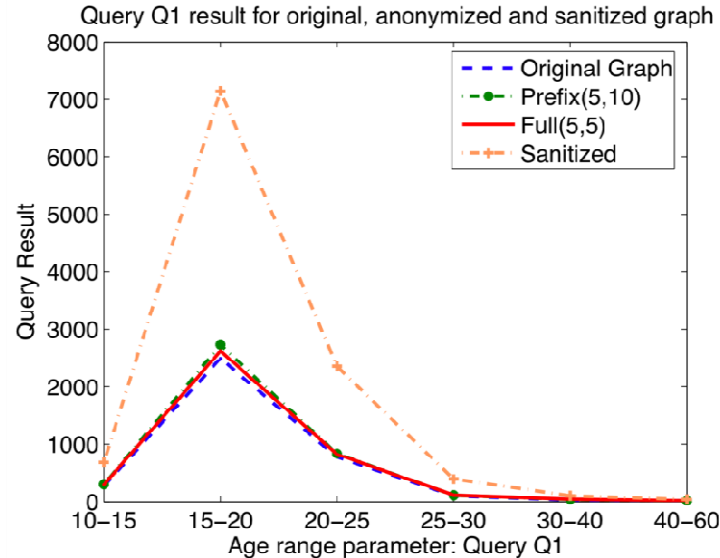


Outline

- ◆ Motivation
- ◆ Anonymization strategies
- ◆ Experiments: evaluation of utility

Experimental Analysis

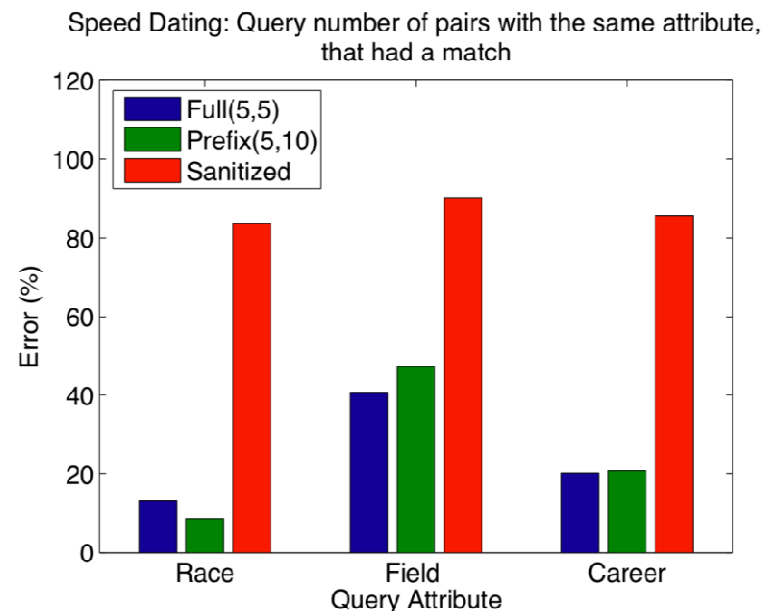
- ◆ Implemented and evaluated our approaches over two datasets
 - Blog (780K nodes, 3M edges), speed dating (530 nodes, 4K edges)
- ◆ Fixed privacy guarantee, analyzed impact on analysis accuracy



Q1 (Pair Query): How many Americans from various age groups are friends with residents of Hong Kong with age <20?

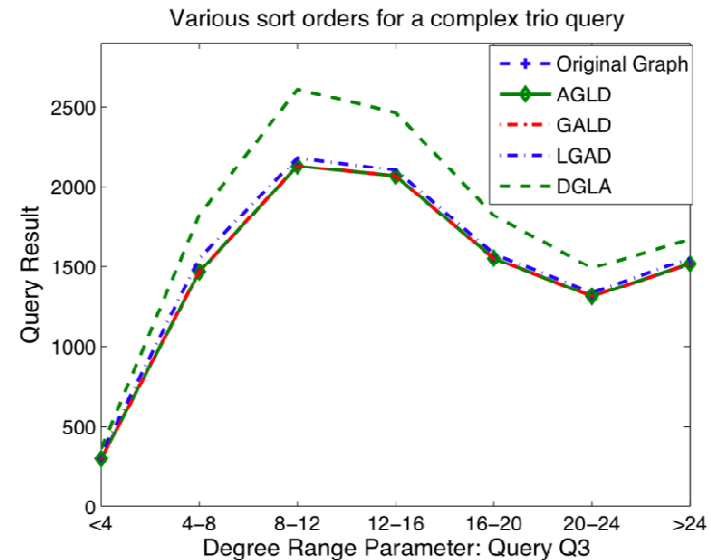
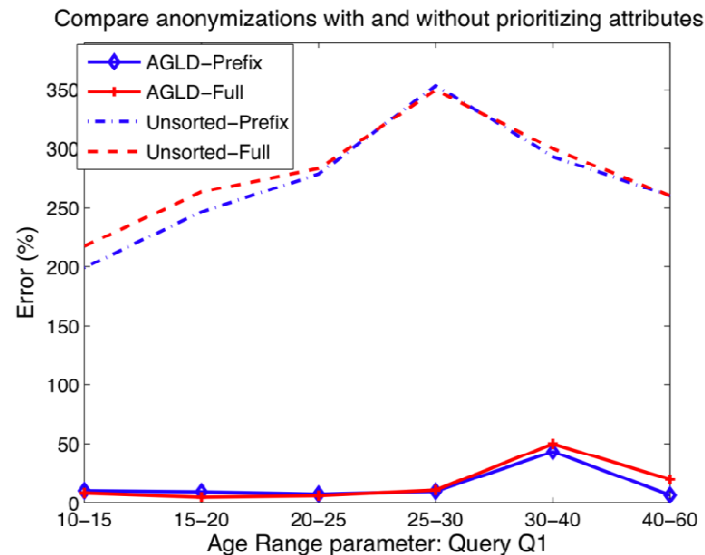
Experimental Analysis

- ◆ Implemented and evaluated our approaches over two datasets
 - Blog (780K nodes, 3M edges), speed dating (530 nodes, 4K edges)
- ◆ Fixed privacy guarantee, analyzed impact on analysis accuracy



Experimental Analysis

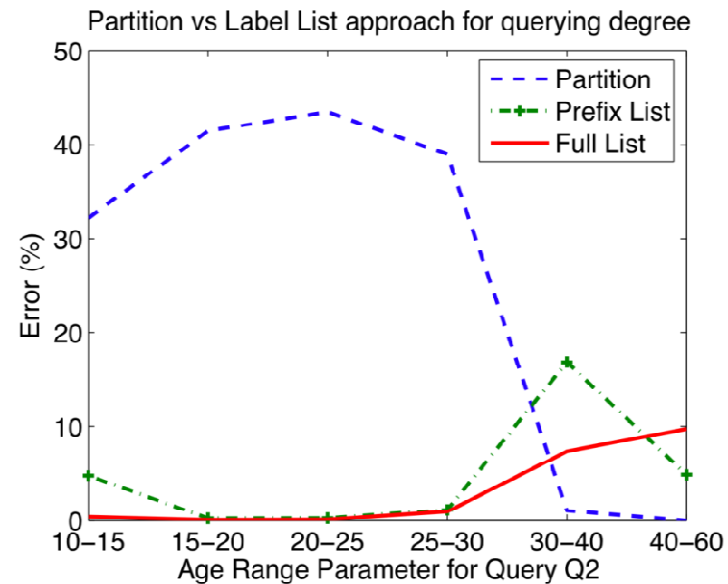
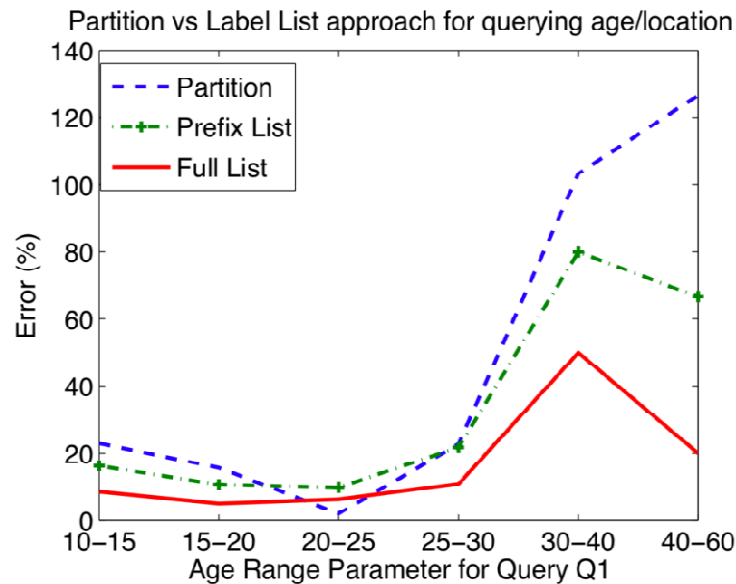
- ◆ Uniform list anonymization over Blog dataset
 - Compare sort orders on Age, Gender, Location, Degree



- ◆ Sorting improves accuracy; choice of sort order matters

Experimental Analysis

- ◆ Comparing uniform list approach with partition approach



- ◆ Errors are typically higher with the partition approach

Conclusions

- ◆ Anonymization of social network data is a challenging problem
- ◆ Contributions
 - Label list and partition approaches: trade off privacy versus utility
 - Class safety conditions provide privacy guarantees
 - Accurate ad hoc aggregate analyses on real datasets
 - Utility is enhanced by using workload-guided sort orders
- ◆ Future work
 - Defuse minimality attack when using workload-guided sort orders
 - Anonymization of time varying social network data