Distribution-based Microdata Anonymization

Nick Koudas (University of Toronto)
Divesh Srivastava (AT&T Labs – Research)
Ting Yu (NC State University)
Qing Zhang (Teradata)

Microdata Anonymization

- Share collection of individual records to support flexible ad hoc data analysis
- Microdata need to be properly transformed
  - Protect privacy of individuals
  - Maximize data utility

Ideas and Contributions

- Key observation
  - Existing approaches tightly couple tuple grouping with achieving privacy objectives
  - Dilemma between desirable grouping and privacy protection
- Our approach
  - De-couple grouping from privacy goals
  - Privacy protection through distribution transformation
  - Permutation with generalization as enabling technique
  - Adding fake values to improve data utility

Motivation

	Quasi-identifiers		SA
tuple ID	zipcode	gender	salary ($)
1	91110	F	30K
2	91210	M	50K
3	91210	M	60K
4	91330	F	30K
5	52210	F	40K
6	52220	F	40K
7	52240	F	60K
8	52210	M	50K

Existing Approaches

- Form groups of tuples to *naturally* satisfy some privacy goal (e.g., ℓ-diversity, t-closeness)
  - Through QI-generalization, clustering or permutation
- Limitation
  - Grouping and protecting privacy are tightly coupled
  - Non-trivial grouping may not exist
  - Resulting grouping may not be desirable

Limitation of Existing Approaches

	Quasi-identifiers		SA
tuple ID	zipcode	gender	salary ($)
1	91110	F	30K
2	91210	M	50K
3	91210	M	60K
4	91330	F	30K
5	52210	F	40K
6	52220	F	40K
7	52240	F	60K
8	52210	M	50K

Data owner desires (1) groups with nearby tuples; (2) 4-diversity

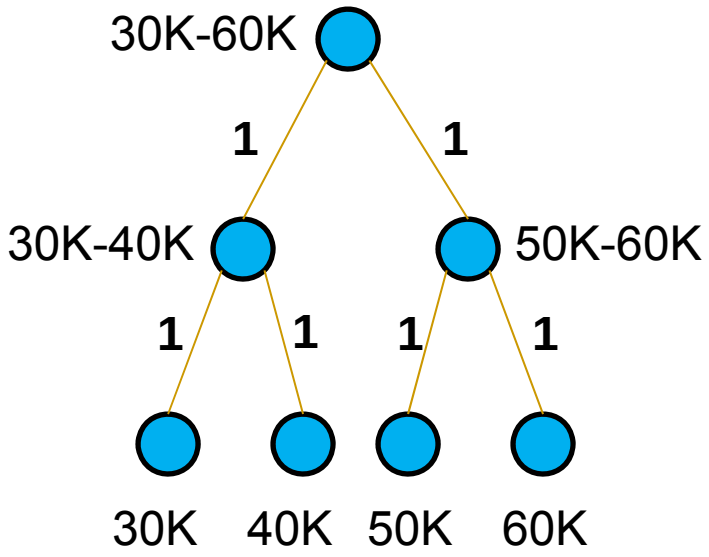
- Through QI generalization?
- Through permutation/bucketization?

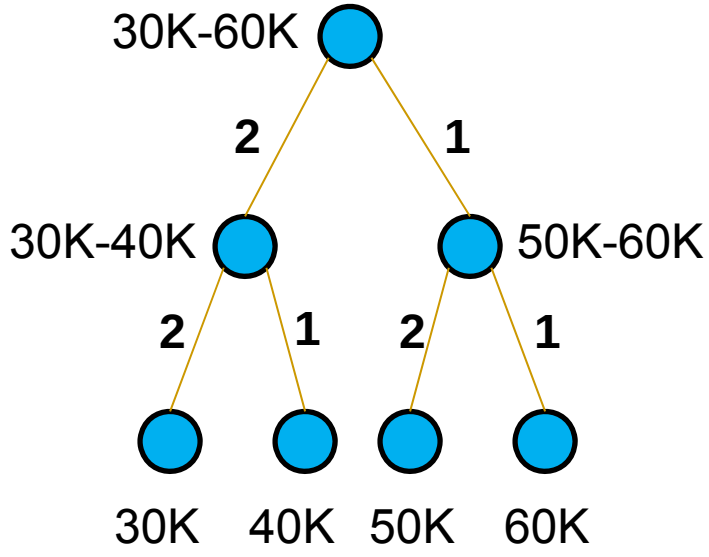
Distribution Transformation

- Privacy objectives essentially impose constraints on SA distribution in a group
  - E.g., ℓ-diversity, t-closeness
- A new methodology for anonymization
  - Decouple grouping from privacy protection
  - Given a group of tuples, transform its SA distribution to a target distribution that satisfies constraints of a privacy goal

Target Distribution

- Specified through a privacy-annotated hierarchy



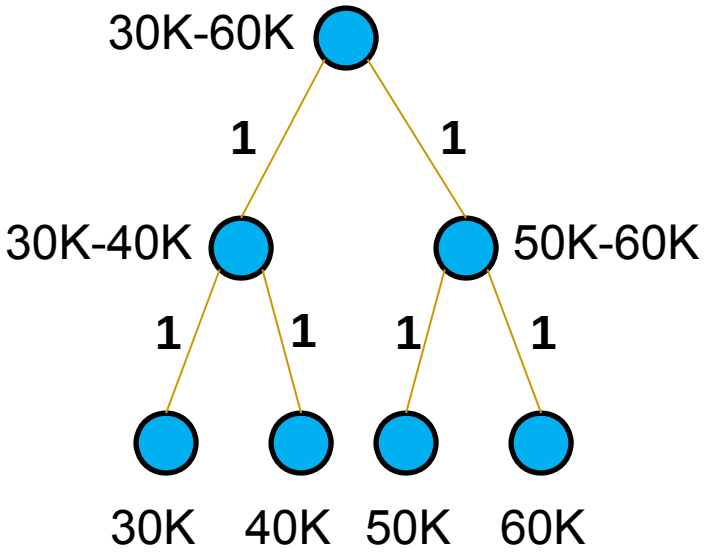


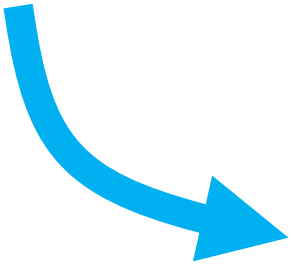
SA Generalization with Permutation

- Generalize SA of each tuple
  - As defined by the privacy-annotated hierarchy
- Permute generalized SAs inside a group
  - To ensure the SA distribution in the group to follow the target distribution

SA Generalization with Permutation

	Quasi-identifiers		SA
tuple ID	zipcode	gender	salary ($)
1	91110	F	30K
2	91210	M	50K
3	91210	M	60K
4	91330	F	30K





	Quasi-identifiers		SA
tuple ID	zipcode	gender	salary ($)
1	91110	F	[30K-40K]
2	91210	M	[30K-40K]
3	91210	M	50K
4	91330	F	60K

Optimization for Data Utility

Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	[30K-40K]
91210	M	[30K-40K]
91210	M	50K
91330	F	60K

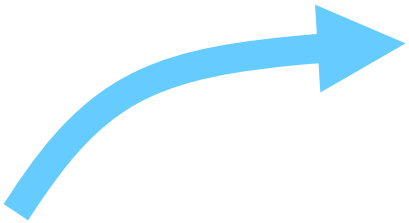
Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	[30K-40K]
91210	M	[30K-40K]
91210	M	[50K-60k]
91330	F	[50K-60k]

- Same privacy, different data utility
- Optimization: find generalizations with minimum sum of ranges
- Designed a linear time optimal algorithm

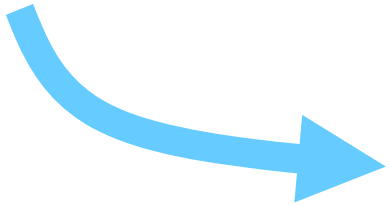
Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	[30K-60K]
91210	M	[30K-60K]
91210	M	[30K-60K]
91330	F	[30K-60K]

Adding Fake Values to Improve Utility

Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	30K
91210	M	50K
91210	M	60K



Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	[30K-40K]
91210	M	[50K-60K]
91210	M	[30K-60K]



Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	30K
91210	M	40K
91210	M	50K
		60K

Adding Fake Values to Improve Utility (cont'd)

- What is the average salary of male employees?

Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	[30K-40K]
91210	M	[50K-60K]
91210	M	[30K-60K]

30K – 60K (trivial)

Quasi-identifiers		SA
zipcode	gender	salary ($)
91110	F	30K
91210	M	40K
91210	M	50K
		60K

35K – 55K (more accurate)

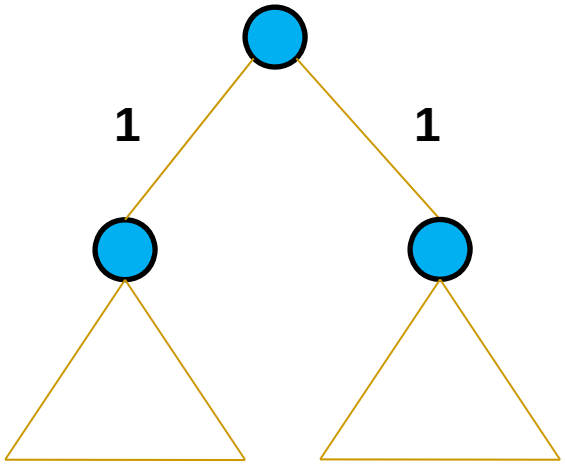
Integration of Fake Values

- Allow arbitrary number of fake values?
  - Can always reduce sum of ranges to 0
  - But not always good for utility
  - Need a bound on number of fake values
- Problem statement
  - Given a group of tuples, adding no more than *t* fake values, such that the sum of ranges after generalization is minimized

Challenges

- Adding more fake values is not always better
  - Should not only consider the case of adding *t* values
- Local optimality ≠ global optimality
  - Local optimality: adding f values is better than adding f-1 or f+1 values
- Where to add fake values has to be a global decision

Challenges



- *n* tuples in the left tree, *n*-1 tuples in the right tree
- Add one fake value
- Surprisingly, adding a fake value to the right tree is not always optimal

An Example

- Domain=[1-16]
- Uniform target distribution
- SAs = {*1, 1, 1, 2, 3, 4, 5, 6, 7, 9, 9, 11, 11, 13, 13, 15, 15*}
  - 9 values in [1-8], 8 values in [9-16]
- Add one fake value to [9-16]
  - Optimal sum of range: 46
- Add one fake value to [1-8]
  - Optimal sum of range: 38

Optimal Algorithm for Adding Fake Values

- Consider all cases from adding exactly 0, 1, to *t* fake values
- Consider all possible allocations of fake values to subtrees
- Algorithm complexity
  - $O(t^2(|g| \log |D| + t\,|D|))$

Query Answering

- Ad hoc aggregation queries
  - Arbitrary constraints on QIs and then aggregate over SA
  - E.g., select sum(salary) from EMP where gender="M" or zipcode="912**"
- Return deterministic upper/lower bounds
  - Examine each group and aggregate over all the groups
- Accuracy
  - (Ubound – Lbound)/ actual_answer

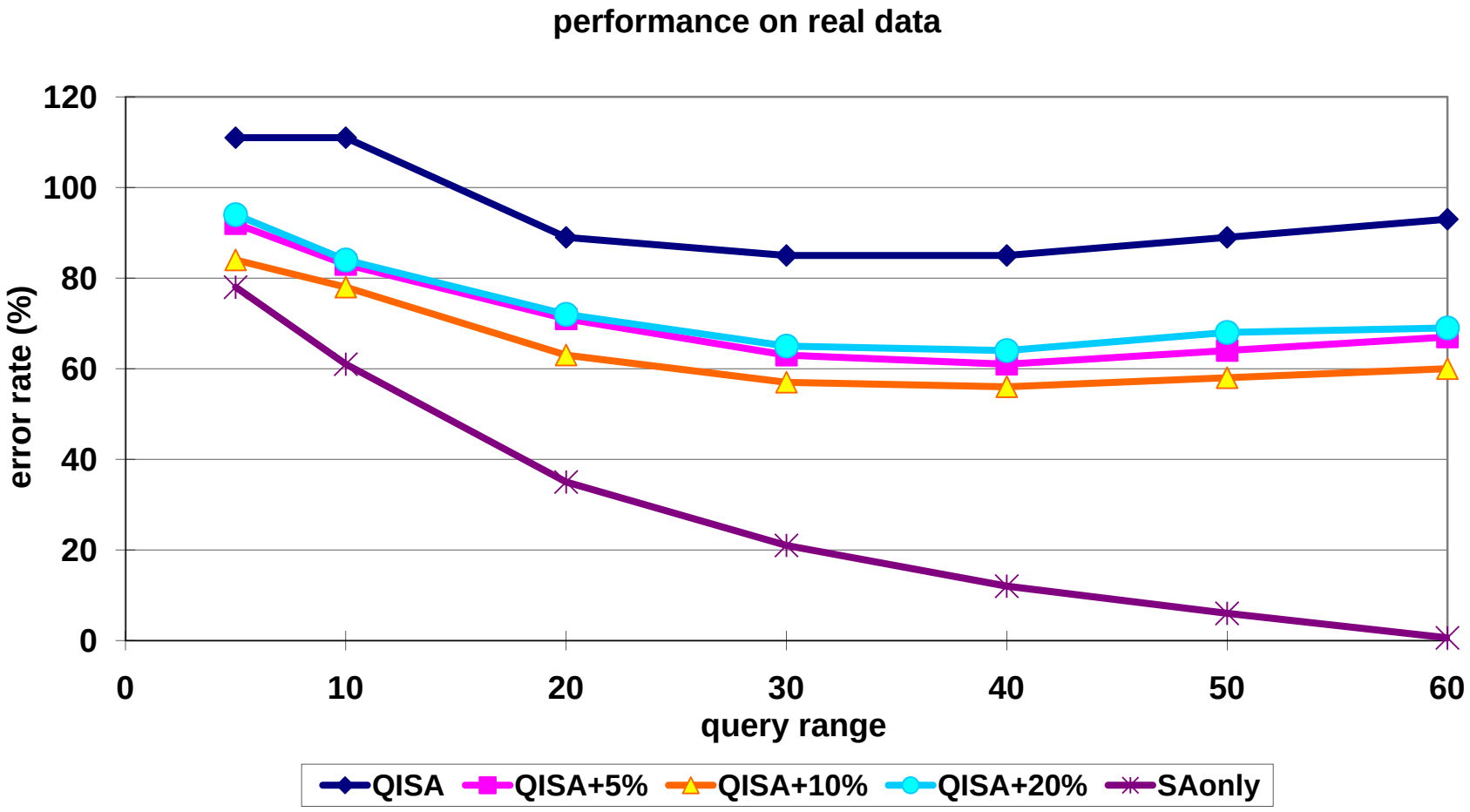
Experimental Evaluation

- Data utility
  - Generalization with permutation
  - Integrating fake values
  - Impact of target distribution
- Algorithm running time
- Data set
  - UCI Adult database
  - “capital loss” as SA
  - Synthetic dataset with the same schema

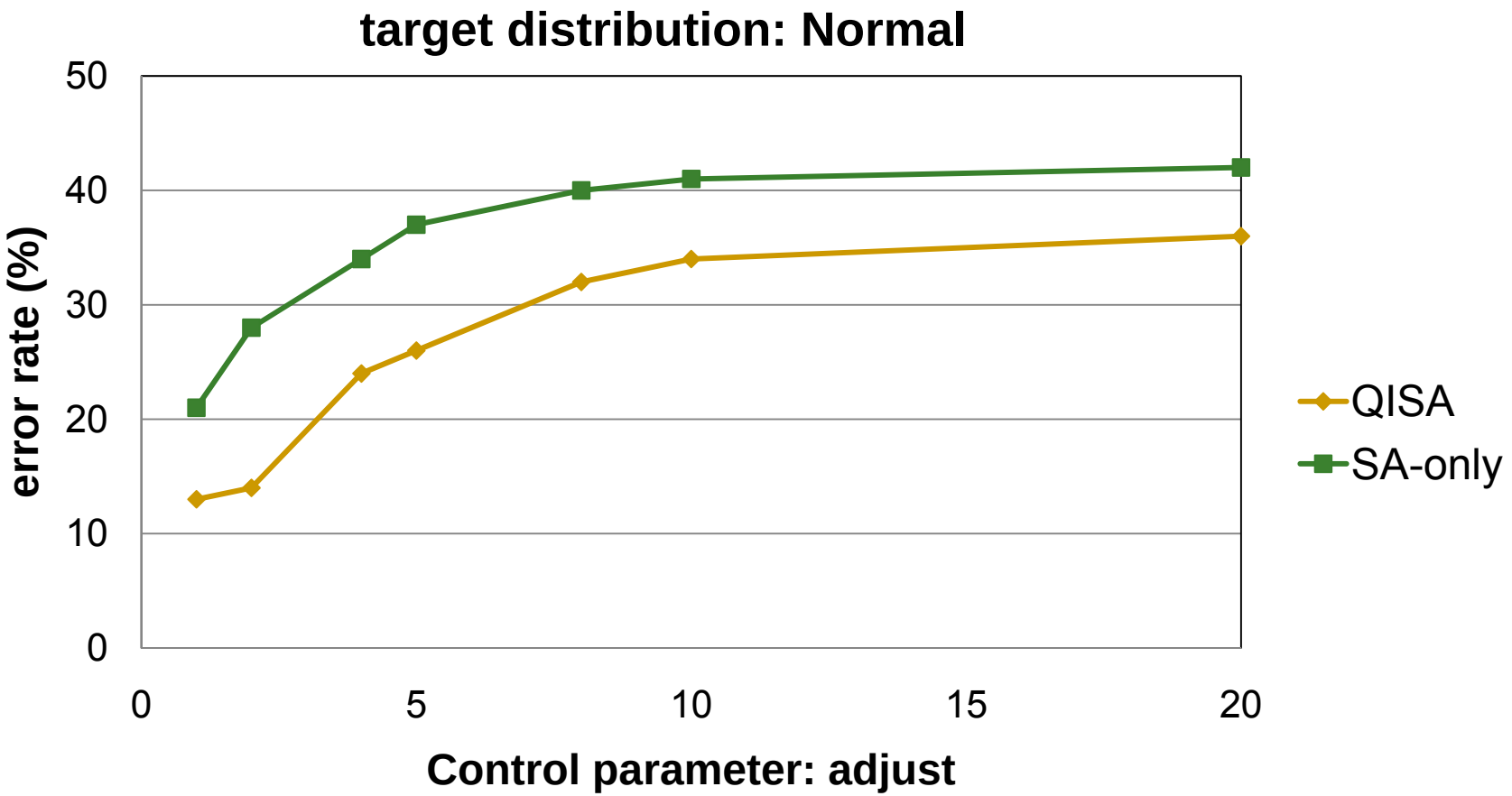
Observations of Key Factors

- Selectivity of queries
- Difference between target distribution and actual distribution
- Correlation between QI and SA

Experiment Results (Fake Values)



Experiment Results (Target Distributions)



Conclusion

- Distribution transformation as a new methodology for microdata anonymization
  - Improve flexibility and feasibility
- SA generalization with permutation to transform distribution
  - Adding fake values to improve utility
- Future work
  - Support data with rich structures
  - Techniques for performing complex data analysis