KANIS: Preserving k-Anonymity over Distributed Data

Katerina Doka, Dimitrios Tsoumakos, Nectarios Koziris
Computing Systems Laboratory
National Technical University of Athens
{katerina, dtsouma, nkoziris}@cslab.ece.ntua.gr

Motivation

- Popularity of personalization services
  - Gathering of sensitive information
  - Privacy concerns
  - e.g., location based services
- Existing approaches (e.g., k-Anonymity)
  - Centralized storage
- Huge amounts of data ⇨ distribution
- Anonymity over distributed data?

Goals

- Preserve k-anonymity of data
  - Multidimensional
  - Hierarchical
  - Horizontally distributed
- Real time
- Maximize utility
- Minimize communication overhead

Contributions

- KANIS
  - Complete DHT-based indexing system
  - Online operation
  - Real time k-anonymization during updates
  - Distributed environment
  - Adjustment of indexing level
    - Each node monitors the privacy of local data
  - Preservation of hierarchy semantics

Presentation Outline

- Background
- KANIS System Design
- KANIS Operations
- Experimental Evaluation
- Conclusions-Future Work

What is k-Anonymity?

- Make every tuple identical to at least k-1 other w.r.t. an attribute set
  - Quasi Identifier set (QID)

No	Gender	Age	Postcode	Problem
1	male	middle	4350	flu
2	male	middle	4350	ulcer
3	male	middle	4351	ulcer
4	female	old	4353	flu
5	female	old	4353	ulcer

- Example:2-anonymity
  - QID={Gender,Age,Postcode}
  - No.3 unique

How can it be achieved?

- Domain generalization
  - global, local
  - climb up levels in the domain hierarchy

Gender | *Age* | *Postcode*

Gender	*Age*	*Postcode*
*	*	*
gender	interval	state
	value	region
		city
		suburb

No	Gender	Age	Postcode	Problem
1	male	middle	4350	flu
2	male	middle	4350	ulcer
3	male	middle	4351	ulcer
4	female	old	4353	flu
5	female	old	4353	ulcer

No	Gender	Age	Postcode	Problem
1	male	middle	435*	flu
2	male	middle	435*	ulcer
3	male	middle	435*	ulcer
4	female	old	435*	flu
5	female	old	435*	ulcer

Quality of k-anonymity

- Distortion of a table
- Weighted Hierarchical Distance WHD
  - Each hierarchy level i has a weight w_i
  - Generalizing from level p to level q

$$WHD = \frac{\sum_{p+1}^{q} w_i}{\sum_{2}^{h} w_i}$$

  - Distortion of a generalized tuple is the sum of WHD of all attributes of QID
  - Distortion of a generalized table is the sum of

K-ANonymity Indexing System

- Complete DHT-based system for preserving k-anonymity of distributed data under updates
- Initial insertion at a *pivot* level combination
- System monitors its data while updates keep coming
- Adaptive re-indexing to
  - Maintain k-anonymity
  - Minimize distortion

National Technical University of Athens
CSLab

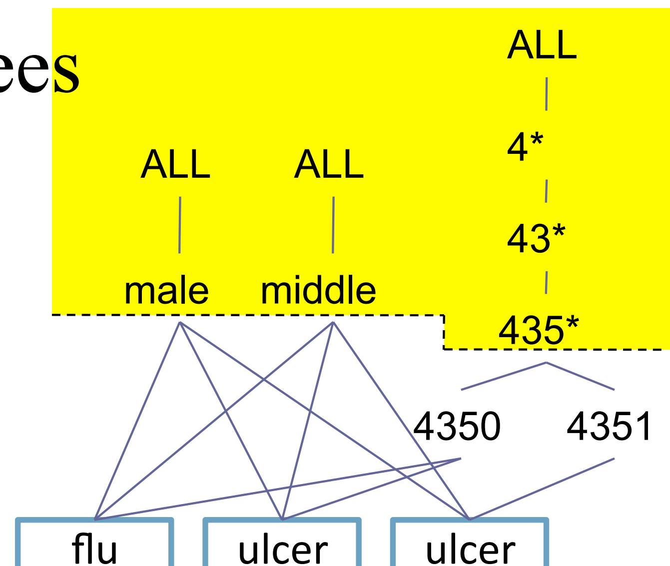
KANIS Insertion

- A *pivot* level combination that satisfies k-anonymity is globally selected
  - e.g., `<gender,interval,city>`
- The key of each tuple is the hashed pivot level value
  - E.g., key = SHA1(male, middle, 435*)
- Tuples stored in the form of trees

male	middle	4350	flu

male	middle	4350	ulcer

male	middle	4351	ulcer



KANIS Updating

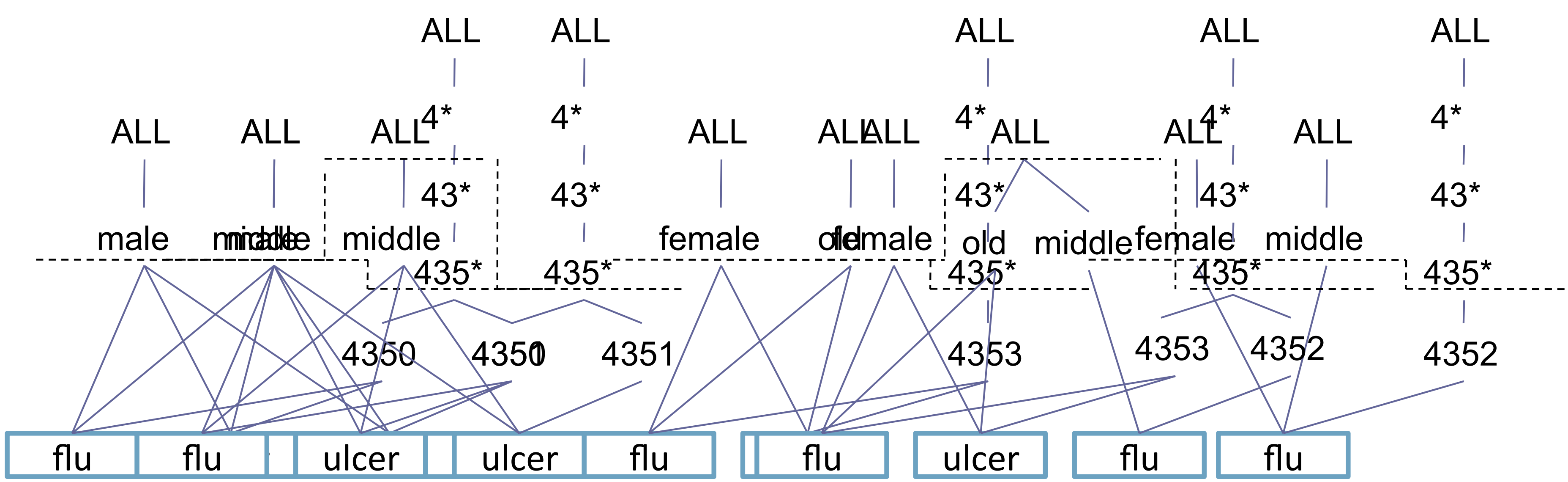
- Insertion of new tuples (read-only)
- Hash according to *pivot* and store tuple
  - if tuple unique for QID, new tree
    - jeopardizes k-anonymity
  - if not, append to existing tree
    - may overgeneralize table
- Both cases require the selection of a new *pivot*
  - Roll-up or drill-down in domain hierarchy

Roll-up Anonymization

- When k-anonymity is broken
- Information from all nodes needed
  - *collectStats* message
  - all possible combinations above *pivot*
  - nodes return frequencies of combinations
- Initiator chooses combination that
  - results in frequencies > k
  - causes minimum distortion
- Data re-distributed according to new pivot

Roll-up Anonymization example

- Insertion of <female, middle, 4352, flu>
  - 2-anonymity violation
  - new *pivot* `<gender,*,city>`

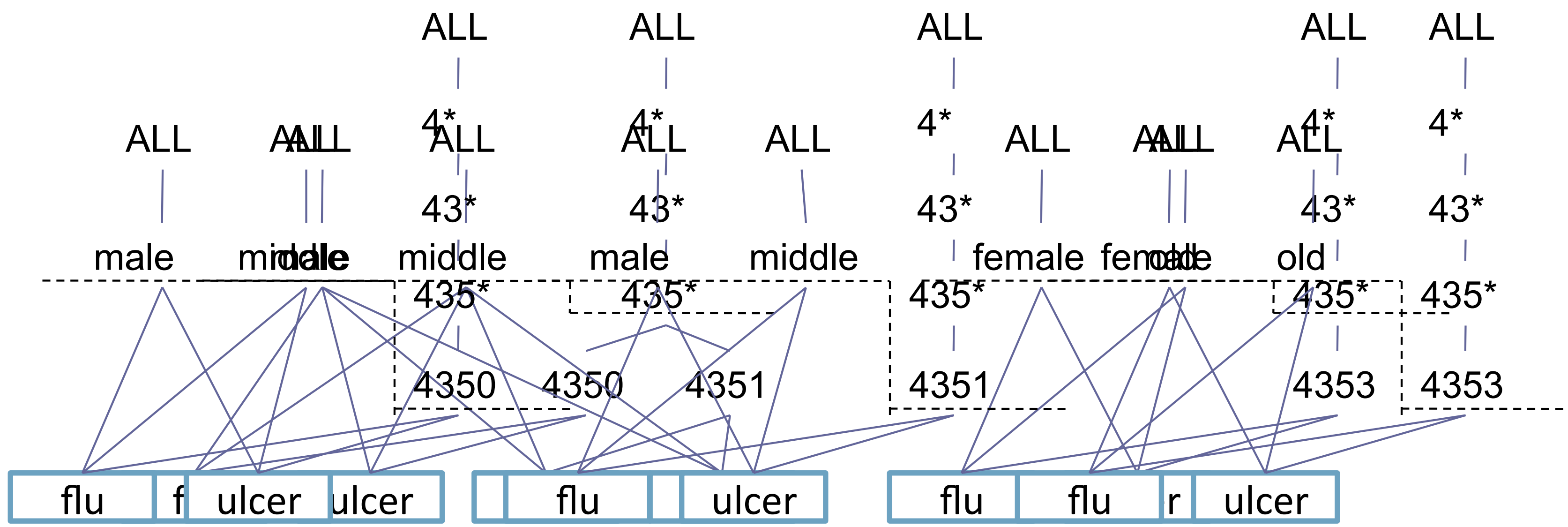


Drill-down Anonymization

- Possible overgeneralization if a specific tree contains > 2k tuples
- Local phase:
  - Find the set of possible combinations below *pivot* that satisfy the k constraint
- Global phase:
  - Set flooded to all nodes
  - Nodes return subset that satisfies k-anonymity locally
- Initiator selects combination that minimizes distortion

Drill-down Anonymization example

- Insertion of <male, middle, 4351, flu>
  - possible over-generalization
  - new *pivot* `<gender,interval,suburb>`



KANIS Reindexing

- Re-organization of data
- Flooding of a *Re-index* message
- Each receiver rehashes tuples according to new *pivot*
- Sends tuples to corresponding nodes
- Tuples grouped by recipient

Experimental Evaluation

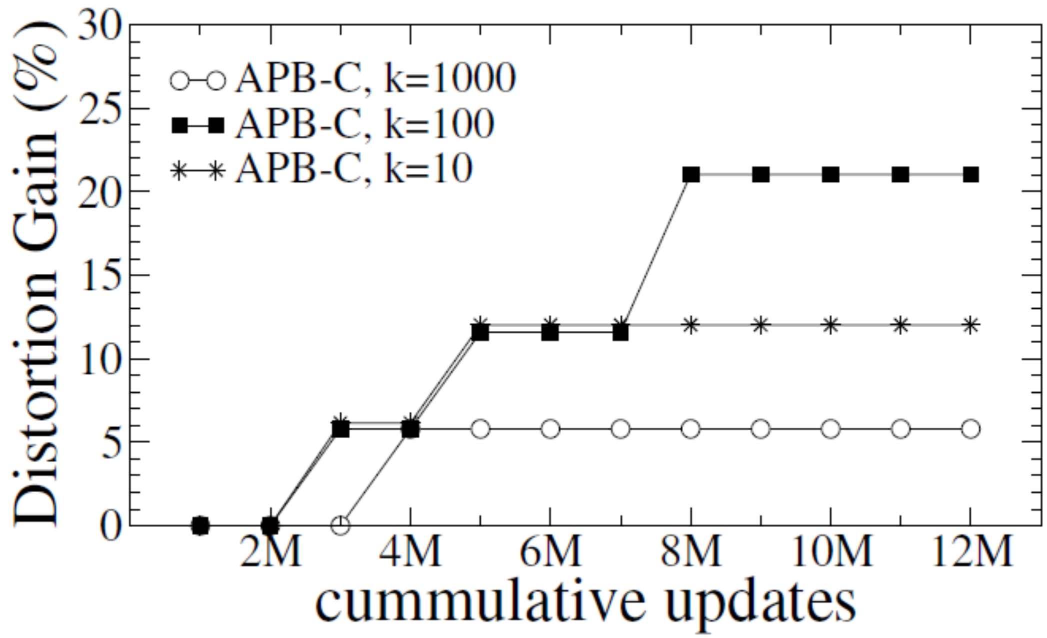
- Modified FreePastry simulator
- 16-128 nodes
- Dataset
  - Adult dataset (45k tuples, 8-d)
  - APB benchmark generator (up to 12M tuples, 4-d)
- Quality – Communication cost
- Compared to
  - Incognito – distortion gain
  - Baseline case – distortion deviation

Size of update batch

	upd size	KANIS #ReInd	KANIS msg/node	KANIS BW	distortion deviation
k=5	1k	1	5.3	1.6M	2%
	5k	1	5.1	1.7M	1%
	10k	1	5.1	1.7M	1%
k=10	1k	2	9.8	2.6M	4%
	5k	2	9.8	3.3M	4%
	10k	2	9.8	4.5M	4%
k=15	1k	2	9.8	2.5M	3%
	5k	2	9.8	3.1M	3%
	10k	2	9.8	4.4M	3%

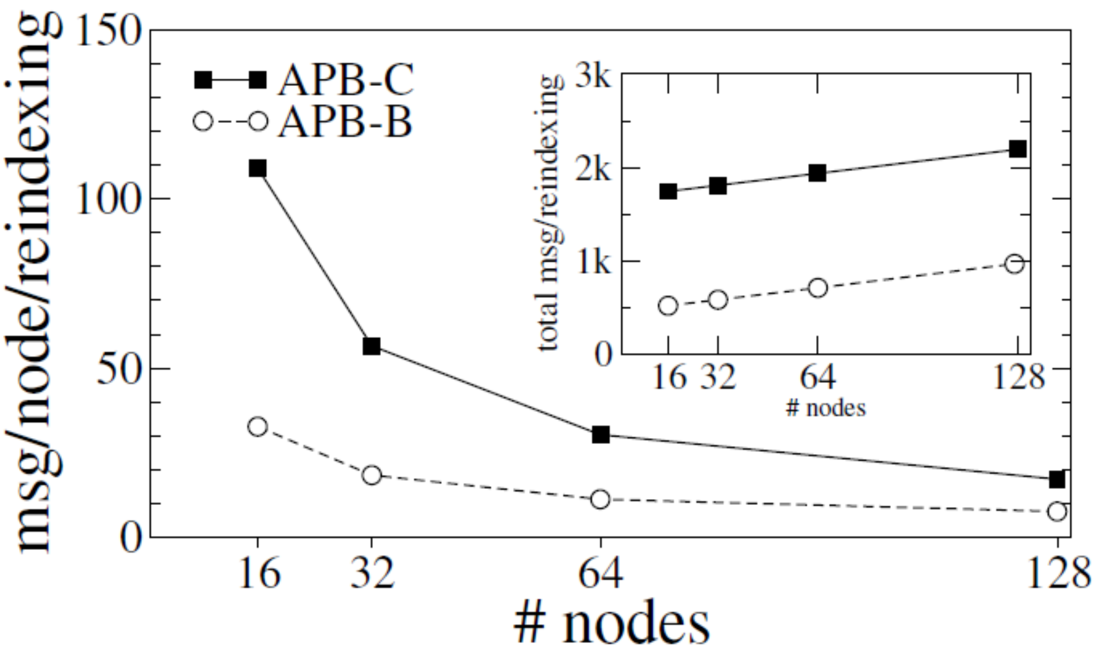
- Adult dataset, 5k initially, 1k-10k update batches
- Deviation remains less than 4%
- Less than 2 reindexings – affordable communication cost
- Smaller k values – less communication cost

Number of tuples



- Initial k-anonymized table of 1M tuples + batches of 1M updates
- Distortion for various k values
- Gain in distortion rises with the addition of new data
- 20% quality improvement compared to the centralized algorithm.
- Distortion deviation is 0 during the biggest part of the experiment

Number of nodes



- Varying #nodes from 16 to 128
- Average number of messages per reindexing increases
- Average load per node decreases
- Steady gains in distortion regardless of the network size

Conclusions – Future Work

- KANIS preserves anonymity over distributed data under continuous updates.
- Employs adaptive scheme that adjusts indexing according to privacy constraints
- Experiments show
  - Up to 22% quality improvement over centralized method
  - Near optimal distortion regardless of network or dataset size
  - Small communication overhead, scattered among nodes
- Extension of KANIS for
  - local recoding
  - other privacy principles

Thank you!


LET ME EXPLAIN THIS A BIT MORE CLEARLY. IT WAS A MAN WHOSE NAME IS "NOBODY" THAT BLINDED ME.