Optimal k-anonymity using generalization and suppression is NP-hard

Ryan Williams (with Adam Meyerson)

03/27/03

What is k-anonymity?

- Proposed by Latanya Sweeney
- Strategy for releasing large amounts of personal data, while still protecting privacy of individuals
- Level of privacy protection depends on a constant parameter k

What is k-anonymity?

In particular, data fields are either *generalized* or *suppressed*

- *Generalized:* e.g. "age 35" becomes "age 20-40"
- *Suppressed:* e.g. "age 35" is withheld entirely

In this talk, we will deal only with optimal k-anonymity via *suppression* (generalization is a generalization of suppression)

Optimal k-anonymity: Given a list of personal *records*, **minimize** the number of *fields* suppressed, such that for each record r, there are $k-1$ other records that are *indistinguishable* from r.

Example of k-anonymity

Consider the query “Who had an x-ray at this hospital yesterday?” and the following response:

first	**last**	**age**	**race**
Harry	Stone	34	Afr-Am
John	Reyser	36	Cauc
Beatrice	Stone	34	Afr-Am
John	Delgado	22	Hisp

- Want to 2-anonymize this data (using suppression) before release

Example of k-anonymity

Consider the query "Who had an x-ray at this hospital yesterday?" and the following response:

first	last	age	race
*	Stone	34	Afr-Am
John	*	*	*
*	Stone	34	Afr-Am
John	*	*	*

- Rows 1 and 3 are indistinguishable, 2 and 4 are indistinguishable

The Goal

Optimal k-anonymity: Given a list of records, minimize the number of fields suppressed, such that for each record r, there are $k - 1$ other records that are indistinguishable from r.

We will give a reduction from k-dimensional perfect matching to the above problem

k-dimensional perfect matching is NP-hard (cf. Garey and Johnson)

Hence optimal k-anonymity is not possible to achieve efficiently, unless $P = NP$

k-dimensional perfect matching

Given a collection C of k-sets over a universe U, is there a subset $S \subseteq C$ such that:

- Every $x \in U$ is in some k-set s in S
- The sets of S are disjoint; i.e. for every $s_1, s_2 \in S$, $s_1 \cap s_2 = \emptyset$

Example:

Let $k = 3$, $U = \{1,2,3,4,5,6\}$, and

$C = \{ \{1,2,3\}, \{1,4,5\}, \{4,5,6\}, \{2,3,6\} \}$.

- $\{ \{1,2,3\}, \{4,5,6\} \}$ and $\{ \{1,2,3\}, \{4,5,6\} \}$ are perfect matchings
- $\{ \{1,2,3\}, \{1,4,5\} \}$ and $\{ \{1,2,3\}, \{4,5,6\}, \{2,3,6\} \}$ are *not*

Note: When $k = 2$, this is polynomial time solvable (but the problem is hard for $k \geq 3$)

From 3-D perfect matching to 3-anonymity

Given an instance of 3-dim. perfect matching:

$U = \{x_1, x_2, \ldots, x_n\}$, $C = \{s_1, \ldots, s_m\}$ *such that*

For all $j = 1, \ldots, m$, $s_j \subseteq U$ *and* $|s_j| = 3$,

Define a table T of records where:

- Records (rows) correspond to $x_i \in U$
- Attributes (columns) correspond to $s_j \in C$

More precisely,

$$T[i,j] := \begin{cases} 0 & \text{if } x_i \in s_j, \\ i & \text{otherwise.} \end{cases}$$

We then ask: *does the optimal 3-anonymized solution suppress at most* $n \cdot (m-1)$ *fields?*

Example of reduction in action

Consider our example from before:

$U = \{1, 2, 3, 4, 5, 6\}$ and $C = \{ \{1, 2, 3\}, \{1, 4, 5\}, \{4, 5, 6\}, \{2, 3, 6\} \}$

The reduction results in the table:

	$\{1, 2, 3\}$	$\{1, 4, 5\}$	$\{4, 5, 6\}$	$\{2, 3, 6\}$
1	0	0	1	1
2	0	2	2	0
3	0	3	3	0
4	4	0	0	4
5	5	0	0	5
6	6	6	0	0

To get an 3-anonymous table, the minimum number of fields in T that need to be suppressed is $18 = 6 \cdot 3$

Perfect Matching 1

3-D perfect matching $\{ \{1, 2, 3\}, \{4, 5, 6\} \}$ corresponds to the 3-anonymized table:

	$\{1, 2, 3\}$	$\{1, 4, 5\}$	$\{4, 5, 6\}$	$\{2, 3, 6\}$
1	0	*	*	*
2	0	*	*	*
3	0	*	*	*
4	*	*	0	*
5	*	*	0	*
6	*	*	0	*

Perfect Matching 2

3-D perfect matching $\{ \{1,4,5\}, \{2,3,6\} \}$ corresponds to:

	$\{1,2,3\}$	$\{1,4,5\}$	$\{4,5,6\}$	$\{2,3,6\}$
1	*	0	*	*
2	*	*	*	0
3	*	*	*	0
4	*	0	*	*
5	*	0	*	*
6	*	*	*	0

Some observations:

- If a set s_j doesn't appear in the perfect matching, then its column is all *'s
- If s_j does appear, then 3 entries in its column are not *'s

Why does this work?

(Recall m = number of sets in collection = number of columns in table)

- A group of 3 rows needs at least $3 \cdot (m-1)$ stars in order for the group to become indistinguishable

    Follows from $T[i,j] := i$ **if** $x_i \notin s_j$

- A group of 3 rows corresponds to the elements of a set s_j *if and only if* exactly $3 \cdot (m-1)$ stars are required

    The rows have 0 in the jth column, differ in other columns

- Thus there is a perfect matching *iff* for every group of 3 rows, exactly $3 \cdot (m-1)$ stars are necessary

    $\Longrightarrow n \cdot (m-1)$ **stars in total**

So there is a 3-D perfect matching *if and only if* **the number of entries suppressed in the optimal 3-anonymized solution is** $n \cdot (m-1)$

Some special cases

Let n be the number of records.

What if...

- **Number of attributes per record (number of columns) is at most $\log(n)$?**

  Reduction doesn't work; resulting subcase of k-dimensional perfect matching is easy

- **Number of possible field entries is constant?**

  Open – could be hard. Variant of k-anonymity where entire columns are suppressed is hard in this case

That’s it!

Any questions?