On the Complexity of Optimal K-Anonymity

Ryan Williams (with Adam Meyerson)

PODS 2004

What is k-anonymity?

- Strategy for releasing large amounts of personal data, while still protecting privacy of individuals
- Originally proposed by Latanya Sweeney
- Level of privacy protection depends on a 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 our work, we deal only with optimal k-anonymity via *suppression*

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.

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

Overview of Talk

- **NP-hardness of optimal k-anonymity**
  - For a sufficiently large alphabet, k-anonymity is hard for any $k \geq 3$
- **Approximation of k-anonymity**
  - Can find a solution that suppresses at most $O(k \log k)$ times the optimum number of fields
  - Two $O(k \log k)$-approximation algorithms: a simple one with $O(n^{2k})$ time, and a more complicated one with $O(n^3)$ time *(the latter improves the second algorithm in the paper)*

Hardness of k-anonymity

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: 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 $s_1, s_2 \in S$, $s_1 \cap s_2 = \emptyset$

Note: When $k = 2$, this is polynomial time solvable (but the problem is NP-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

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

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 j**th 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 – Sweeney has announced a polytime algorithm

- **Number of possible field entries (alphabet) is constant?**

  Recently resolved in a paper submitted to ESA 2004 – it suffices to have a ternary alphabet

$O(k \log k)$-approximation for k-anonymity

We will approximately solve a related problem, which we call k*-minimum diameter sum*

Given a collection of vectors $S \subseteq \Sigma^m$, the *diameter of* S is

$$d(S) := \max_{u,v \in S} h(u, v),$$

where h is Hamming distance

($d(S)$ is the diameter of the smallest Hamming ball enclosing S)

The k-minimum diameter sum problem: Given $V \subseteq \Sigma^m$, find a partition Π of V into sets S with $|S| \in [k, 2k-1]$, so that $\sum_{S \in \Pi} d(S)$ is minimized

Minimum diameters and k-anonymity

Theorem. Suppose partition Π of V is an α-approximation to k-minimum diameter sum. Then the following is a $3k\alpha$-approximation algorithm for optimally k-anonymizing V:

For each $S \in \Pi$ and for all $j = 1, \ldots, m$, if there are $u, v \in S$ with $u[j] \neq v[j]$, set $w[j] := *$ for all $w \in S$.

Sketch: For any partition Π and any $S \in \Pi$,

- At least $d(S)$ coordinates (out of m) need to be suppressed to make the vectors of S identical

  $\Longrightarrow$ *at least* $|S| \cdot d(S) \geq kd(S)$ stars are required to anonymize S

- Every pair $\{u, v\} \subseteq S$ has $d(u, v) \leq d(S)$, so we only need to insert at most $d(S)$ stars per pair

  $\Longrightarrow$ the algorithm uses *at most* $\binom{|S|}{2} \cdot d(S) \leq 3k^2 d(S)$ stars to anonymize S

Approximating Minimum Diameter Sum

One line summary: Reduce to Set Cover, convert cover into partition

Set Cover: Given a collection $\mathcal{C}$ of sets from a universe U and a weight function $w : \mathcal{C} \rightarrow \mathbb{N}$, find $\mathcal{S} \subseteq \mathcal{C}$ where $\sum_{S \in \mathcal{S}} w(S)$ is minimized and every $x \in U$ appears in some $S \in \mathcal{S}$

Outline of reduction

- Let $\mathcal{C}$ be collection of $S \subseteq V$ such that $k \leq |S| \leq 2k - 1$. Find a set cover $\mathcal{S}$ for $\mathcal{C}$ using the standard greedy $(1 + \ln 2k)$-approximation that repeatedly chooses the most "cost-effective" set S
- For any pair of sets $S, T \in \mathcal{S}$, both containing some $v \in V$,
  - if one of S or T is larger than k, remove v from it
  - if not, $|S| = |T| = k$, so replace S and T with $S \cup T$ in $\mathcal{S}$

Claim: The resulting partition has a diameter sum that is no more than the diameter sum of $\mathcal{S}$

Caveat!

Building the collection $\mathcal{C}$ of all subsets with cardinality in the range $[k, 2k-1]$ takes $O(n^{2k-1})$ time

- This can be skirted with a little geometric trickery
- Still get an $O(k \log k)$ approximation, but now $O(n^3)$ time

Outline of faster algorithm

Instead of using the whole collection $\mathcal{C}$, use a much smaller one, which is reconstructed at each iteration of the greedy set cover algorithm

Each iteration i of the set cover approximation algorithm adds a new set to its collection

For $j = 1, \ldots, 2k-1$ and $v \in V$, define $S_{i,j,v}$ to be the set of j nearest neighbors of v (including v) that are not yet included in the cover at iteration i; if $j < k$, also include the $k - j$ covered vectors closest to v

Let $\mathcal{C}_i$ be the collection of $S_{i,j,v}$ at iteration i

- $\mathcal{C}_i$ is "re-built" (in $O(kn^2)$ time) at each iteration of the greedy algorithm, as more vectors become covered
- Greedy algorithm runs in $O(n)$ iterations, so $O(kn^3)$ time

Claim: This gives a $2(1 + \ln 2k)$-approximation to minimum diameter sum, *i.e.* a $6k(1 + \ln 2k)$-approximation to k-anonymity

Recent improvements *(not in the paper)*

Aggarwal, Feder, Kentapadi, Motwani, Panigrahy, Thomas, and Zhu

(a.k.a. a bunch of people at Stanford) have shown:

- Still NP-hard for a ternary alphabet
- $O(k)$-approximation for k-anonymity
- 1.5-approximation for 2-anonymity, and 2-approximation for 3-anonymity

This paper may appear in ESA04; stay tuned

Interesting directions *(not in the paper)*

- The **maximum disclosure** problem: k-anonymizing, but now we want to maximize the total number of fields *not* suppressed – how well can one approximate?

  We (that is, I) conjecture there is an $O(k)$*-approximation*

- The **costly suppression** problem: Suppose you can only suppress at most F fields among all the records – what's the **maximum** k such that you can still k-anonymize the records?

  NP*-hard, but I've no idea what approximation is like*