A Note on the Hardness of the k-Ambiguity Problem

DSG Technical Report 2002-006

Staal A. Vinterbo[a,b]

[a] *Decision Systems Group, Brigham and Women's Hospital,*
75 Francis Street, Boston, MA 02115, USA

[b] *Harvard Medical School,*
Boston, MA, USA

staal@dsg.harvard.edu

June 10, 2002

Abstract

We address the problem of minimal information loss in order to k-ambiguate data, a problem related to disclosure control in disseminated data. We show that this problem is NP-hard by considering cell suppression as the ambiguation mechanism. On the way we prove that the minimum k-union problem (aka. minimum k-coverage, aka. maximum k-intersection), which is the problem of selecting k sets from a collection of n sets such that the cardinality of their union is the minimum, is NP-hard. Shown is also that if the cardinality of the sets in the collection is bounded by a constant, this restricted problem is in APX.

Keywords: minimum k-union, maximum k-intersection, set cover, ambiguity, disclosure control, minimum k-coverage, combinatorial optimization

1 Introduction

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 authorizes the Federal government to establish a national standard for medical record privacy either by legislative or regulatory action. On December 28, 2000, the Federal Register issued regulations that require the safeguard of the integrity, confidentiality, and availability of patient data.

There exist several technical approaches to disclosure control in disseminated data, which exhibit different properties in terms of the utility of the data after application for a given purpose. We are here concerned with a particular instance of information generalization [5, 2, 3], the ultimate goal of which is to hamper inferences on the data while preserving as much of the semantics of the data as possible.

We formulate the general k-ambiguity problem question as:

- Let $U = \{x_i\}_{i=1}^{n}$ be a set of size $n \in \mathbf{N}$ of objects of interest. Given a positive integer $k \leq n-1$, how much information about x_i do we have to discard in order to make x_i *indiscernible* from k other elements x_j in U?

2 Analysis

Let $U = \{x_i\}_{i=1}^{n}$ be a finite set of size n of objects of interest. We define an *attribute* a on U to be a function $a : U \rightarrow V_a$ from U into an attribute value set V_a. A data table over U and a set of attributes $A = \{a_j\}_{j=1}^{m}$ can be described as a matrix $T[A] = \{c_{ij} | c_{ij} = a_j(x_i)\}$. The row in $T[A]$ associated with x_i is denoted $r_i[A]$. Each entry c_{ij} in T is called a *cell* and contains either a value from V_{a_j}, or a special value $\top$. The value $\top$ is by definition indiscernible from all possible values. We say that cell c_{kj} is indiscernible from cell c_{lj} if both cells contain the same value (in $V_{a_j} \cup \{\top\}$), or c_{kj} contains the value $\top$. We denote this indiscernibility by $c_{kj} \doteq c_{lj}$. In a data table $T[A]$, we say that row $r_i[A]$ is indiscernible from row $r_j[A]$, $r_i[A] \doteq r_j[A]$, if and only if the cells in $r_i[A]$ are indiscernible from the corresponding cells in $r_j[A]$, i.e., if and only if $c_{il} \doteq c_{jl}$ for all attributes $a_l \in A$. The changing of cell c_{ij} from a value in V_{a_j} to $\top$ is called the *suppression* of this cell.

In the following we equate information loss with the number of suppressed cells.

Akin to Skowron et al [4] we define a $n \times n$ *discernibility matrix* over U and A to be $M_U^A = \{m_{ij}\}_{i,j=1}^{n,n}$ where

$$m_{ij} = \{k | a_k(x_i) \not\doteq a_k(x_j)\}.$$

For convenience define M_i to be the collection of sets m_{ij} such that $j \neq i$. We associate with each $M \subseteq \{1, 2, \ldots, m\}$ a function f_M defined by $f_M(x) =$

$(v_1, v_2, \ldots, v_m)$ such that $v_j = a_j(x)$ if $j \in M$ and $v_j = \top$ otherwise. This enables us to formalize the question above as the following optimization problem:

Problem 1 *For fixed i, find a minimal set H such that $H \supseteq m_{ij}$ for at least $k \leq |M_i|$ distinct values j such that $m_{ij} \in M_i$.*

For $k = |M_i|$ and $k = 1$, this problem is solvable in polynomial time. For the $k = |M_i|$ case, any feasible solution must contain $H' = \cup_{m \in M_i} m$, hence H' is the minimal, and can be computed in polynomial time. For the $k = 1$ case, any $m_{ij} \in M_i$ such that $|m_{ij}|$ is minimal will do. Computing $|m_{ij}|$ can be done in polynomial time. Now, consider the following NP-optimization problem:

Problem 2 *(**Minimum k-union**) A minimization problem with*

- *instance: collection $[C_i]_{i \in J}$ where $C_i \subseteq U$, and positive integer $k \leq |J|$,*
- *feasible solutions: $\{I \subseteq J \mid |I| \geq k\}$, and*
- *measure: $m(I) = | \cup_{i \in I} C_i|$.*

We can immediately see that Problem 1 is an instance of Problem 2. That we can reduce Problem 2 to Problem 1 can be seen by the following. View $A(U)$ as a binary data table where cell (i, j) corresponds to $a_j(x_i)$, and for each set i in the collection C let row i be a binary string encoding of that set, where a 1 in position l denotes that l is in the set. This, in essence, shows that given a finite collection of sets, we can construct a data table such that $\{m_{ij}\}_{i \neq j}$ is that finite collection of sets. Thus, we have the wanted reduction. This means that Problem 1 and Problem 2 are essentially equivalent.

As $\cup_{i \in I} C_i = \overline{\cap_{i \in I} \overline{C_i}}$, and minimizing $|\overline{\cap_{i \in I} \overline{C_i}}|$ equals maximizing $|\cap_{i \in I} \overline{C_i}|$, we see that Problem 2 is essentially equivalent to the following problem:

Problem 3 *(**Maximum k-intersection**) A maximization problem with*

- *instance: collection $[C_i]_{i \in J}$ where $C_i \subseteq U$, and positive integer $k \leq |J|$,*
- *feasible solutions: $\{I \subseteq J \mid |I| \geq k\}$, and*
- *measure: $m(I) = | \cap_{i \in I} C_i|$.*

We now prove that Problem 3 (as well as Problems 2 and 1 by equivalence) is NP-hard. We do this via a reduction from the NP-Complete BALANCED COMPLETE BIPARTITE SUBGRAPH problem [1, GT24], which is the following problem: Given a bipartite graph $G = (V, E)$ and a positive integer $k \leq |V|$, are there two disjoint subsets $V_1, V_2 \subseteq V$ such that $|V_1| = |V_2| = k$ and $V_1 \times V_2 \subseteq E$?

We start by proving that for $V_1 \subseteq V$ such that $|V_1| = k$, there exists $V_2 \subseteq V$ such that $|V_2| = k$ and $V_1 \times V_2 \subseteq E$ if and only if $| \cap_{i \in V_1} C_i| \geq k$, where $C_i = \{j|(i,j) \in E\}$. Assume that $|V_1| = |V_2| = k$ and $V_1 \times V_2 \subseteq E$. For each $i \in V_1$ we have that $V_2 \subseteq C_i$. As $|V_2| = k$, we have that $| \cap_{i \in V_1} C_i| \geq k$. Now assume that there exists V_1, $|V_1| = k$, such that $| \cap_{i \in V_1} C_i| \geq k$. As G is bipartite, we have that V_1 and $\cap_{i \in V_1} C_i$ are disjoint. Any set $V_2 \subseteq \cap_{i \in V_1} C_i$ results in $V_1 \times V_2 \in E$ by definition of C_i. We can choose $V_2 \subseteq \cap_{i \in V_1} C_i$ such that $|V_2| = k$ by assumption.

Now, let I be a solution to the instance $(\{C_i\}_{i \in V}, k)$ of Problem 3. If $m(I) < k$, there exists no $V_1 \subseteq V$ of size k such that $| \cap_{i \in V_1} C_i| \geq k$, and, by the above, no balanced complete bipartite subgraph with $2k$ vertices. If $m(I) \geq k$, we can choose any $V_1 \subseteq I$ such that $|V_1| = k$ and any $V_2 \subseteq \cap_{i \in I} C_i$ such that $|V_2| = k$ and have that $V_1 \times V_2 \subseteq E$. As we can find C_i in polynomial time in the size of G for each $i \in V$, we can conclude:

Theorem 1 *Problem 3 is NP-hard, as are Problems 2 and 1.*

For Problem 3 consider the following greedy algorithm:

1. Set H to be the union of all integers in $\{1, 2, \ldots, m\}$ that can be found in at least k sets in the collection, set counter to 1, and set solution to be the empty set.
2. If H is empty or counter is k, stop and return solution.
3. Find set C_i in the collection that maximizes the intersection with H
4. Add i to the solution, remove C_i from the collection, change H to be the intersection of H and C_i, and increment counter by 1
5. Go to step 2.

Let $\mathcal{I}$ be an instance of Problem 3 and let m_A be the measure returned by the algorithm above. If the solution is empty, we know that $m_A = m^* = 0$. Now assume that $m_A > 0$, and that the optimal measure is m^*. We can immediately conclude that

$$\frac{m^*}{m_A} \leq \frac{s_k}{m_A} \leq s_k,$$

where s_k is the size of the k-th set in a sorted list of the $|J|$ sets in the collection sorted on descending size. This can be seen by observing that the optimal solution has to contain k sets, and that m^* is no larger than the size of the smallest one of these. A consequence of this is that if we restrict the problem to be such that only $k - 1$ sets in the collection are allowed to be larger than a fixed constant, then this restricted version of Problem 3 is in APX.

3 Conclusion

We have here offered a proof that the problem of minimizing information loss while making an object indiscernible from k others is NP-hard. We did this by choosing cell suppression as the mechanism, the number of suppressed cells as a measure of information loss, and showing that these choices lead to the minimum k-coverage problem. We then proved that the minimum k-coverage problem and its equivalent problems are NP-hard, and in APX if the set sizes are bounded by a constant.

Acknowledgements

This work was supported by grant R01-LM07273 from the National Library of Medicine.

References

[1] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman and Company, New York, NY, 1979.

[2] A. J. Hundepool and L. C. R. J. Willenborg. Mu- and tau-argus: Software for statistical disclosure control. In *Third International Seminar on Statistical Confidentiality at Bled*, Bled, 1996.

[3] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 1-3, 1998, Seattle, Washington*, page 188. ACM Press, 1998.

[4] A. Skowron and C. Rauszer. The discernibility matrices and functions in information systems. volume 11 of *Series D: System Theory, Knowledge Engineering and Problem Solving*, chapter III-2, pages 331–362. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

[5] L. Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. *Proc AMIA Annu Fall Symp*, pages 51–5, 1997.