Research Paper ■

Protecting Privacy Using k-Anonymity

Khaled El Emam, PhD, Fida Kamal Dankar, MSc

Abstract **Objective:** There is increasing pressure to share health information and even make it publicly available. However, such disclosures of personal health information raise serious privacy concerns. To alleviate such concerns, it is possible to anonymize the data before disclosure. One popular anonymization approach is k-anonymity. There have been no evaluations of the actual re-identification probability of k-anonymized data sets.

Design: Through a simulation, we evaluated the re-identification risk of k-anonymization and three different improvements on three large data sets.

Measurement: Re-identification probability is measured under two different re-identification scenarios. Information loss is measured by the commonly used discernability metric.

Results: For one of the re-identification scenarios, k-Anonymity consistently over-anonymizes data sets, with this over-anonymization being most pronounced with small sampling fractions. Over-anonymization results in excessive distortions to the data (i.e., high information loss), making the data less useful for subsequent analysis. We found that a hypothesis testing approach provided the best control over re-identification risk and reduces the extent of information loss compared to baseline k-anonymity.

Conclusion: Guidelines are provided on when to use the hypothesis testing approach instead of baseline k-anonymity.

■ **J Am Med Inform Assoc.** 2008;15:627–637. DOI 10.1197/jamia.M2716.

Introduction

The sharing of raw research data is believed to have many benefits, including making it easier for the research community to confirm published results, ensuring the availability of original data for meta-analysis, facilitating additional innovative analysis on the same data sets, getting feedback to improve data quality for on-going data collection efforts, achieving cost savings from not having to collect the same data multiple times by different research groups, minimizing the need for research participants to provide data repeatedly, facilitating linkage of research data sets with administrative records, and making data available for instruction and education.[1–14] Consequently, there are pressures to make such research data more generally available.[8,15,16] For example, in January 2004 Canada was a signatory to the OECD Declaration on Access to Research Data from Public Funding.[17] This is intended to ensure that data generated through public funds are publicly accessible for researchers as much as possible.[18] To the extent that this is implemented, potentially more personal health data about Canadians will be made available to researchers world wide. The European Commission has passed a regulation facilitating the sharing with external researchers of data collected by Community government agencies.[19] There is interest by the pharmaceutical industry and academia to share raw data from clinical trials.[16,20]

Researchers in the future may *have to* disclose their data. The Canadian Medical Association Journal has recently contemplated requiring authors to make the full data set from their published studies available publicly online.[3] Similar calls for depositing raw data with published manuscripts have been made recently.[2,5,7,20–22] The Canadian Institutes of Health Research (CIHR) has a policy, effective on 1st January 2008, that requires making some data available with publications.[23] The UK MRC policy on data sharing sets the expectation that data from their funded projects will be made publicly available.[24] The UK Economic and Social Research Council requires its funded projects to deposit data sets in the UK Data Archive (such projects generate health and lifestyle data on, for example, diet, reproduction, pain, and mental health).[25] The European Research Council considers it essential that raw data be made available preferably immediately after publication, but not later than six months after publication.[26] The NIH in the US expects investigators seeking more than $500,000 per year in funding to include a data sharing plan (or explain why that is not possible).[27] Courts, in criminal and civil cases, may compel disclosure of research data.[11,28]

Such broad disclosures of health data pose significant privacy risks.[38] The risks are real, as demonstrated by recent

Affiliations of the authors: Children's Hospital of Eastern Ontario Research Institute (KEE, FD), Ottawa, Ontario, Canada; Pediatrics, Faculty of Medicine, University of Ottawa (KEE), Ottawa, Ontario, Canada.

The author(s) declare that they have no competing interests.

The authors thank Bradley Malin, Vanderbilt University, and Jean-Louis Tambay, Statistics Canada, for their detailed feedback and suggestions on earlier versions of this paper. The authors also thank the anonymous reviewers for many helpful suggestions.

Correspondence: Khaled El Emam, Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada; e-mail: <kelemam@uottawa.ca>.

Received for review: 01/09/08; accepted for publication: 05/21/08.

Table 1 ■ Some Examples of Re-identification Attempts in General Data Sets and of Health Data Sets*

General Examples of Re-identification	
AOL search data[29–31]	AOL put anonymized Internet search data (including health-related searches) on its web site. New York Times reporters were able to re-identify an individual from her search records within a few days.
Chicago homicide database[32]	Students were able to re-identify a significant percentage of individuals in the Chicago homicide database by linking with the social security death index.
Netflix movie recommendations[33]	Individuals in an anonymized publicly available database of customer movie recommendations from Netflix are re-identified by linking their ratings with ratings in a publicly available Internet movie rating web site.
Health-specific Examples of Re-identification	
Re-identification of the medical record of the governor of Massachusetts[34]	Data from the Group Insurance Commission, which purchases health insurance for state employees, was matched against the voter list for Cambridge, re-identifying the governor's record.
Southern Illinoisan vs. The Department of Public Health[35, 36]	An expert witness was able to re-identify with certainty 18 out of 20 individuals in a neuroblastoma data set from the Illinois cancer registry, and was able to suggest one of two alternative names for the remaining two individuals.
Canadian Adverse Event Database[37]	A national broadcaster aired a report on the death of a 26 year-old student taking a particular drug who was re-identified from the adverse drug reaction database released by Health Canada.

*The former type of data can contain health information (as in the case of the individual re-identified in the AOL example), and life style and sexual orientation information (as in the case of one of the individuals re-identified in the Netflix example).

successful re-identifications of individuals in publicly disclosed data sets (see the examples in Table 1). One approach for protecting the identity of individuals when releasing or sharing sensitive health data is to anonymize it.[19]

A popular approach for data anonymization is k-anonymity.[39–42] With k-anonymity an original data set containing personal health information can be transformed so that it is difficult for an intruder to determine the identity of the individuals in that data set. A k-anonymized data set has the property that each record is similar to at least another k-1 other records on the potentially identifying variables. For example, if $k = 5$ and the potentially identifying variables are age and gender, then a k-anonymized data set has at least 5 records for each value combination of age and gender. The most common implementations of k-anonymity use transformation techniques such as generalization, global recoding, and suppression.[39,40,42–45]

Any record in a k-anonymized data set has a *maximum* probability $1/k$ of being re-identified.[44] In practice, a data custodian would select a value of k commensurate with the re-identification probability they are willing to tolerate—a *threshold risk*. Higher values of k imply a lower probability of re-identification, but also more distortion to the data, and hence greater information loss due to k-anonymization. In general, excessive anonymization can make the disclosed data less useful to the recipients because some analysis becomes impossible or the analysis produces biased and incorrect results.[46–51]

Thus far there has been no empirical examination of how close the *actual* re-identification probability is to this maximum. Ideally, the actual re-identification probability of a k-anonymized data set would be close to $1/k$ since that balances the data custodian's risk tolerance with the extent of distortion that is introduced due to k-anonymization. However, if the actual probability is much lower than $1/k$ then k-anonymity may be over-protective, and hence results in unnecessarily excessive distortions to the data.

In this paper we make explicit the two re-identification scenarios that k-anonymity protects against, and show that the actual probability of re-identification with k-anonymity is much lower than $1/k$ for one of these scenarios, resulting in excessive information loss. To address that problem, we evaluate three different modifications to k-anonymity and identify one that ensures that the actual risk is close to the threshold risk and that also reduces information loss considerably. The paper concludes with guidelines for deciding when to use the baseline versus the modified k-anonymity procedure. Following these guidelines will ensure that re-identification risk is controlled with minimal information loss when using k-anonymity.

Background

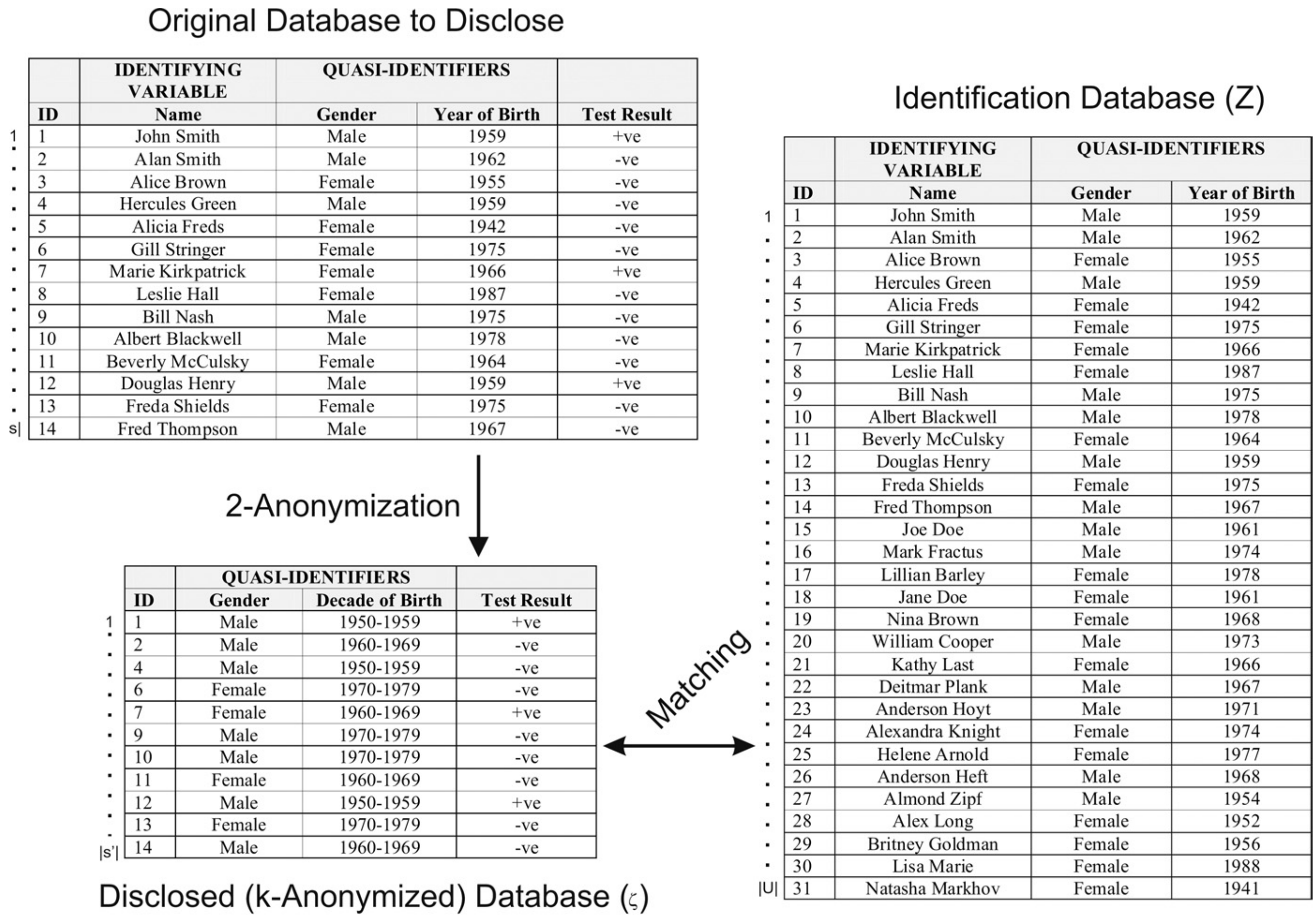
The Two Re-identification Scenarios for a k-Anonymized Data Set

The concern of k-anonymity is with the re-identification of a *single individual* in an anonymized data set.[44] There are two re-identification scenarios for a single individual:[52–54]

1. *Re-identify a specific individual (known as the prosecutor re-identification scenario).* The intruder (e.g., a prosecutor) would know that a particular individual (e.g., a defendant) exists in an anonymized database and wishes to find out which record belongs to that individual.
2. *Re-identify an arbitrary individual (known as the journalist re-identification scenario).* The intruder does not care which individual is being re-identified, but is only interested in being able to claim that it can be done. In this case the intruder wishes to re-identify a single individual to discredit the organization disclosing the data.

Re-identification Risk under the Prosecutor Scenario

The set of patients in the file to be disclosed is denoted by s. Before the file about s can be disclosed, it must be anonymized. Some of the records in the file will be suppressed during anonymization, therefore a different subset of patients, s', will be represented in the anonymized version of this file. Let the anonymized file be denoted by ζ. There is a one-to-one mapping between the records in ζ and the individuals in s'.


2-Anonymization
Matching
1
s|
|s'|
|U|

Original Database to Disclose

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
ID	Name	Gender	Year of Birth	Test Result
1	John Smith	Male	1959	+ve
2	Alan Smith	Male	1962	-ve
3	Alice Brown	Female	1955	-ve
4	Hercules Green	Male	1959	-ve
5	Alicia Freds	Female	1942	-ve
6	Gill Stringer	Female	1975	-ve
7	Marie Kirkpatrick	Female	1966	+ve
8	Leslie Hall	Female	1987	-ve
9	Bill Nash	Male	1975	-ve
10	Albert Blackwell	Male	1978	-ve
11	Beverly McCulsky	Female	1964	-ve
12	Douglas Henry	Male	1959	+ve
13	Freda Shields	Female	1975	-ve
14	Fred Thompson	Male	1967	-ve

Disclosed (k-Anonymized) Database (ζ)

	QUASI-IDENTIFIERS		
ID	Gender	Decade of Birth	Test Result
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Female	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
14	Male	1960-1969	-ve

Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

Figure 1. A hypothetical example of the three databases assumed in the k-anonymity privacy model under the journalist re-identification scenario. The intruder performs the matching while the data custodian performs the 2-anonymization.

Under the prosecutor scenario, a specific individual is being re-identified, say, a VIP. The intruder will match the VIP with the records in ζ on *quasi-identifiers*. Variables such as gender, date of birth, postal code, and race are commonly used quasi-identifiers. Records in ζ that have the same values on the quasi-identifiers are called an *equivalence class*.[55]

Let the number of records in ζ that have exactly the same quasi-identifier values as the VIP be f. The re-identification risk for the VIP is then $1/f$. For example, if the individual being re-identified is a 50 year old male, then f is the number of records on 50 year old males in ζ. The intruder has a probability $1/f$ of getting a correct match.

Since the data custodian does not know, a priori, which equivalence class a VIP will match against, one can assume the worse case scenario. Under the worse case scenario, the intruder will have a VIP who matches with the smallest equivalence class in ζ, which in a k-anonymized data set will have a size of at least k. Hence the re-identification probability will be at most $1/k$.

Therefore, under the prosecutor re-identification scenario k-anonymity can ensure that the re-identification risk is approximately equal to the threshold risk, as intended by the data custodian. This, however, is not the case under the journalist re-identification scenario.

Re-identification Risk under the Journalist Scenario

We assume that there exists a large finite population of patients denoted by the set U. We then have $s' \subseteq s \subseteq U$. An intruder would have access to an *identification database* about the population U, and uses this identification database to match against the patients in ζ. The identification database is denoted by Z, and the records in Z have a one-to-one mapping to the individuals in U.

In the example of Figure 1 we have a data set about 14 individuals that needs to be disclosed. This data set is 2-anonymized to produce the anonymized data set, ζ. After 2-anonymization, there are only 11 records left in ζ since three had to be suppressed.

An intruder gets hold of an identification database with 31 records. This is the Z database. The intruder then attempts re-identification by matching an arbitrary record against the records in ζ on year of birth and gender. In our example, once an arbitrary individual is re-identified, the intruder will have that individual's test result.

The discrete variable formed by cross-classifying all values on the quasi-identifiers in ζ can take on J distinct values. Let $X_{\zeta,i}$ denote the value of a record i in the ζ data set. For example, if we have two quasi-identifiers, such as gender and age, then we may have $X_{\zeta,1} = "MALE, 50"$, $X_{\zeta,2} = "MALE, 53"$, and so on. Similarly let $X_{Z,i}$ denote the value of record i in the Z data set.

The sizes of the different equivalence classes are given by $f_j = \sum_{i \in s'} I\,(X_{\zeta,i} = j), \quad j = 1, \ldots, J$, where f_j is the size of a ζ equivalence class and $I(\cdot)$ is the indicator function. Similarly we have $F_j = \sum_{i \in U} I(X_{Z,i} = j), \quad j = 1, \ldots, J$, where F_j is the size of an equivalence class in Z.

Under the journalist re-identification scenario, the probability of re-identification of a record in an equivalence class j is $1/F_j$.[56,57] However, a smart intruder would focus on the records in equivalence classes with the highest probability of re-identification. Equivalence classes with the smallest value for F_j have the highest probability of being re-identified, and therefore we assume that a smart intruder will focus on these. The probability of re-identification of an arbitrary individual by a smart intruder is then given by:

$$\theta_{\max} = 1 \Big/ \min_j\,(F_j)$$

If we consider Figure 1 again, the 2-anonymized file had the age converted into 10 year intervals. In that example we can see that $\theta_{\max} = 0.25$ because the smallest equivalence class in Z has 4 records (ID numbers 1, 4, 12, and 27). With 2-anonymization the data custodian was using a threshold risk of 0.5, but the actual risk of re-identification, $\theta_{\max}$, was half of that. This conservatism may seem like a good idea, but in fact it has a large negative impact on data quality. In our example, 2-anonymization resulted in converting age into ten year intervals and the suppression of more than one fifth of the records that had to be disclosed (3 of 14 records had to be suppressed). By most standards, losing one fifth of a data set due to anonymization would be considered extensive information loss.

Now consider another approach: k-map. With k-map it is assumed that the data custodian can k-anonymize the identification database itself (and hence directly control the F_j values). Let's say that the Z identification database is k-anonymized to produce Z'. The k-map property states that each record in ζ is similar to at least k records in Z'.[34,41] This is illustrated in Figure 2. Here, the data custodian 2-anonymizes the identification database directly, and then implements the transformations to the data set to be disclosed. In this example $\theta_{\max} = 0.5$ because the smallest equivalence classes in Z' for records 1 to 14 have two records. Also, the extent of information loss is reduced significantly: there are no records suppressed in the disclosed data set and the age is converted into 5 year intervals rather than 10 year intervals. By using the k-map property we have ensured that the actual re-identification risk is what the data custodian intended and we have simultaneously reduced information loss.

In practice, the k-map model is not used because it is assumed that the data custodian does not have access to an identification database, but that an intruder does.[34,41] Therefore, the k-anonymity model is used instead.

There are good reasons why the data custodian would not have an identification database. Often, a population database is expensive to get hold of. Plus, it is likely that the data custodian will have to protect multiple populations, hence multiplying the expense. For example, the construction of a single profession-specific database using semi-public registries that can be used for re-identification attacks in Canada costs between $150,000 to $188,000.[58] Commercial databases can be comparatively costly. Furthermore, an intruder may commit illegal acts to get access to population registries. For example, privacy legislation and the Elections Act in Canada restrict the use of voter lists to running and supporting election activities.[58] There is at least one known case where a charity allegedly supporting a terrorist group has been able to obtain Canadian voter lists for fund raising.[59–61] A legitimate data custodian would not engage in such acts.

However, a number of methods have been developed in the statistical disclosure control literature to estimate the size of the equivalence classes in Z from a sample. If these estimates are accurate, then they can be used to approximate k-map. Approximating k-map will ensure that the actual risk is close to the threshold risk, and consequently that there will be less information loss. Three such methods are considered below.

Proposed Improvements to k-Anonymity under the Journalist Re-identification Scenario

We consider three alternative approaches to reduce the extent of over-anonymization under the journalist re-identification scenario. These three approaches extend k-anonymity to approximate k-map. The details of the proposed approaches are provided in Appendix A (Risk Estimates) (available as a JAMIA online-only data supplement at www.jamia.org).

Baseline (D1)

Baseline k-Anonymization algorithms apply transformations, such as generalization, global recoding, and suppression until all equivalence classes in ζ are of size k or more.

Individual Risk Estimation (D2)

The actual re-identification risk for each equivalence class in ζ, $1/\hat{F}_j$, can be directly estimated.

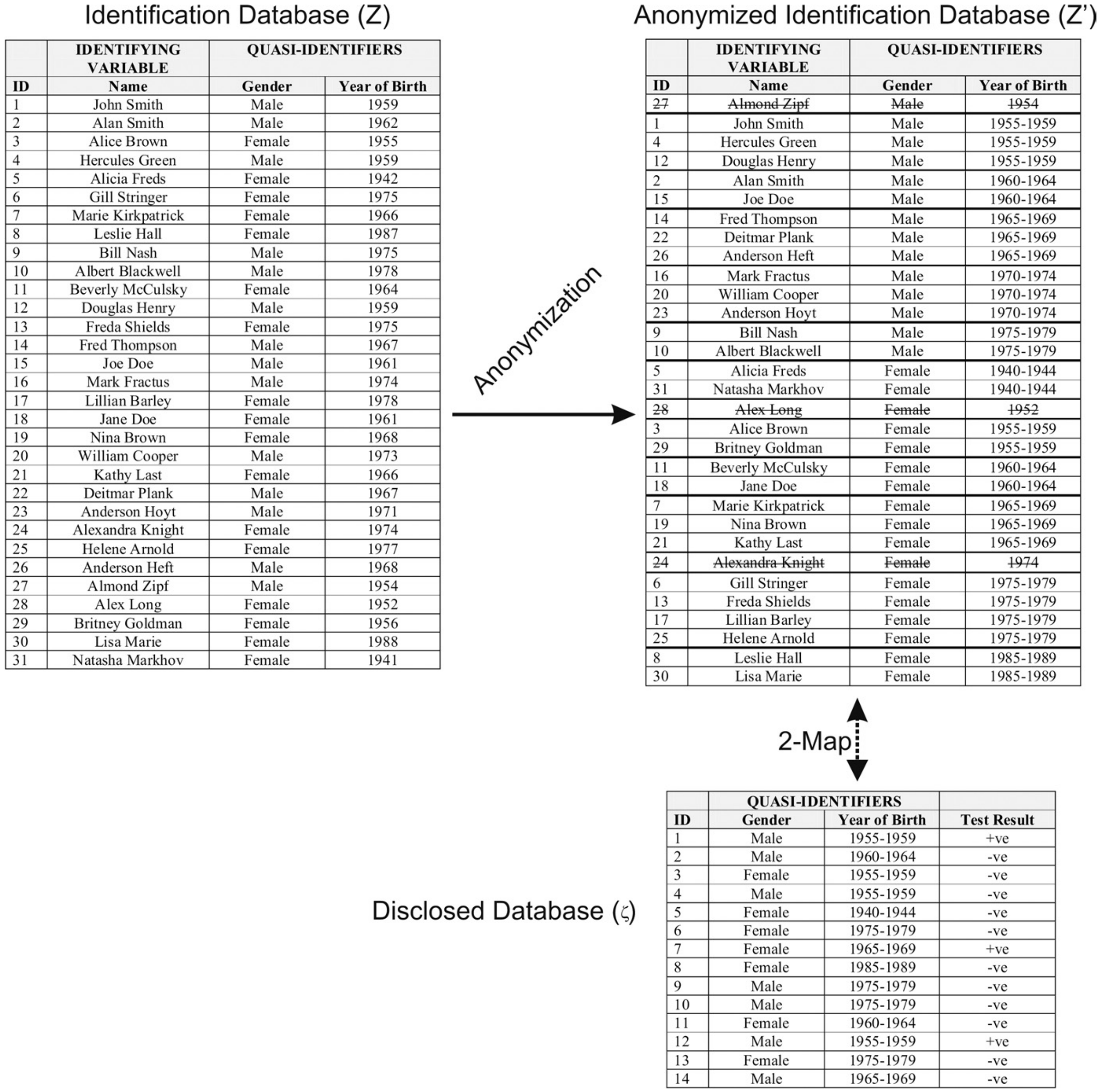
Subsequently, the k-anonymization algorithm should ensure that all equivalence classes meet the following condition $1/\hat{F}_j \leq 1/k$. One estimator for $1/\hat{F}_j$ has been studied extensively[57,62–66] and was also incorporated in the mu-argus tool (which was developed by the Netherlands national statistical agency and used by many other national statistical agencies for disclosure control purposes),[67–70] but it has never been evaluated in the context of k-anonymity. To the extent that this individual risk estimator is accurate, it can ensure that the actual risk is as close as possible to the threshold risk.

Hypothesis Testing Using a Poisson Distribution (D3)

One can use a hypothesis testing approach for determining if $F_j > k$.[56,71] If we assume that the size of the sample equivalence classes f_j follow a Poisson distribution, we can construct the null Poisson distribution for $H_0: F_j < k$ and determine which observed value of f_j will reject the null hypothesis at $\alpha = 0.1$. Let's denote this value as k'. Then the k-anonymity algorithm should ensure that the following condition $f_j \geq \min(k,k')$ is met for all equivalence classes.

Hypothesis Testing Using a Truncated-at-zero Poisson Distribution (D4)

In practice, we ignore equivalence classes that do not appear in the sample, therefore, the value of f_j cannot be equal to zero. An improvement in the hypothesis testing approach above would then be to use a truncated-at-zero Poisson



Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

Anonymized Identification Database (Z')

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
~~27~~	~~Almond Zipf~~	~~Male~~	~~1954~~
1	John Smith	Male	1955-1959
4	Hercules Green	Male	1955-1959
12	Douglas Henry	Male	1955-1959
2	Alan Smith	Male	1960-1964
15	Joe Doe	Male	1960-1964
14	Fred Thompson	Male	1965-1969
22	Deitmar Plank	Male	1965-1969
26	Anderson Heft	Male	1965-1969
16	Mark Fractus	Male	1970-1974
20	William Cooper	Male	1970-1974
23	Anderson Hoyt	Male	1970-1974
9	Bill Nash	Male	1975-1979
10	Albert Blackwell	Male	1975-1979
5	Alicia Freds	Female	1940-1944
31	Natasha Markhov	Female	1940-1944
~~28~~	~~Alex Long~~	~~Female~~	~~1952~~
3	Alice Brown	Female	1955-1959
29	Britney Goldman	Female	1955-1959
11	Beverly McCulsky	Female	1960-1964
18	Jane Doe	Female	1960-1964
7	Marie Kirkpatrick	Female	1965-1969
19	Nina Brown	Female	1965-1969
21	Kathy Last	Female	1965-1969
~~24~~	~~Alexandra Knight~~	~~Female~~	~~1974~~
6	Gill Stringer	Female	1975-1979
13	Freda Shields	Female	1975-1979
17	Lillian Barley	Female	1975-1979
25	Helene Arnold	Female	1975-1979
8	Leslie Hall	Female	1985-1989
30	Lisa Marie	Female	1985-1989

Disclosed Database (ζ)

	QUASI-IDENTIFIERS		
ID	Gender	Year of Birth	Test Result
1	Male	1955-1959	+ve
2	Male	1960-1964	-ve
3	Female	1955-1959	-ve
4	Male	1955-1959	-ve
5	Female	1940-1944	-ve
6	Female	1975-1979	-ve
7	Female	1965-1969	+ve
8	Female	1985-1989	-ve
9	Male	1975-1979	-ve
10	Male	1975-1979	-ve
11	Female	1960-1964	-ve
12	Male	1955-1959	+ve
13	Female	1975-1979	-ve
14	Male	1965-1969	-ve

Figure 2. An illustration of the k-map approach whereby the data custodian does have access to the identification database. The crossed out values are suppressed records.

distribution[72,73] to determine the value of *k*′. The k-anonymity algorithm should ensure that the condition $f_j \geq \min(k,k')$ is met for all equivalence classes.

Methods

Our objective was to evaluate the three methods described above, and compare their performance to the baseline k-anonymity approach. We performed a simulation study to evaluate (a) the actual re-identification probability for k-anonymized data sets under the journalist re-identification scenario, and (b) the information loss due to this k-anonymization. We use values of $k = 5$, 10, and 15. Even though a minimum *k* value of 3 is often suggested,[54,74] a common recommendation in practice is to ensure that there are at least five similar observations ($k = 5$).[75–80] It is uncommon for data custodians to use values of *k* above 5, and quite rare that values of *k* greater than 15 are used in practice.

Data Sets

For our simulation we used 3 data sets which served as our populations. The first is the list of physicians and their basic demographics from the College of Physicians and Surgeons of Ontario with 23,590 observations.[81] The quasi-identifiers we used were: postal code (5,349 unique values), graduation year (70 unique values), and gender (2 unique values). The second was a data set from the Paralyzed Veterans Association on veterans with spinal cord injuries or disease with 95,412 observations.[82] The quasi-identifiers we used were: zip code (10,909 unique values), date of birth (901 unique values), and gender (2 values). The third data set is the fatal crash information database from the department of transportation with 101,034 observations.[83] The quasi-identifiers used were age (98 unique values), gender (2 unique values), race (19 unique values), and date of death (386 unique values).

The quasi-identifiers we used in our three data sets are ones known to make it easy to link with publicly available information in Canada and the US.[32,58,84,85]

k-Anonymization

One thousand simple random samples were drawn from each data set at nine different sampling fractions (0.1 to 0.9 in increments of 0.1). Any identifying variables were removed and each sample was k-anonymized.

An existing global optimization algorithm[44] was implemented to k-anonymize the samples. This algorithm uses a cost function to guide the k-anonymization process (the objective is to minimize this cost). A commonly used cost function to achieve baseline k-anonymity is the discernability metric.[44,86–91] In Appendix A (Risk Estimates) we describe how this cost function is adjusted to implement the approaches D2, D3, and D4 within the same global optimization algorithm.

Note that records with missing values on the quasi-identifiers were removed from our analysis.

Evaluation

For each k-anonymized data set the actual risk is measured as $\theta_{\max}$ and the information loss is measured in terms of the discernability metric. Averages were calculated for each sampling fraction across the 1000 samples.

The results are presented in the form of three sets of graphs:

Risk. This shows the value of $\theta_{\max}$ against sampling fraction for each of the four approaches.

Information Loss. Because the discernability metric is affected by the sample size (and hence makes it difficult to compare across differing sampling fractions), we normalize it for D2, D3, and D4 by the baseline value. For example, a value of 0.8 (or 80%) for D2 means that the information loss for D2 is 80% of that for the baseline k-anonymity approach. The graph shows the normalized discernability metric for these three approaches against the sampling fraction. The value for D1 will by definition always be 1 (or 100%).

Suppression. Because the extent of suppression is an important indicator of data quality by itself, we show graphs of the percentage of suppressed records against sampling fraction for the four approaches.

Results

We will only present the results for $k = 5$ (i.e., a risk threshold = 0.2) here, with the remaining graphs for $k = 10$ and $k = 15$ provided in Appendix B (Results for $k = 10$ and $k = 15$) (available as a JAMIA online-only data supplement at www.jamia.org). The conclusions for $k = 10$ and $k = 15$ support the $k = 5$ results.

The actual re-identification risk $\theta_{\max}$ is shown using the four approaches in Figure 3. The baseline approach (D1), which is current practice, is quite low and exhibits a wide gap between the actual risk and the 0.2 risk threshold at $k = 5$. This gap is quite marked for small sampling fractions and disappears for large sampling fractions. At higher sampling fractions there is no difference among the baseline approach and the other three in terms of actual risk (they all converge to 0.2 as the sampling fraction approaches 1).

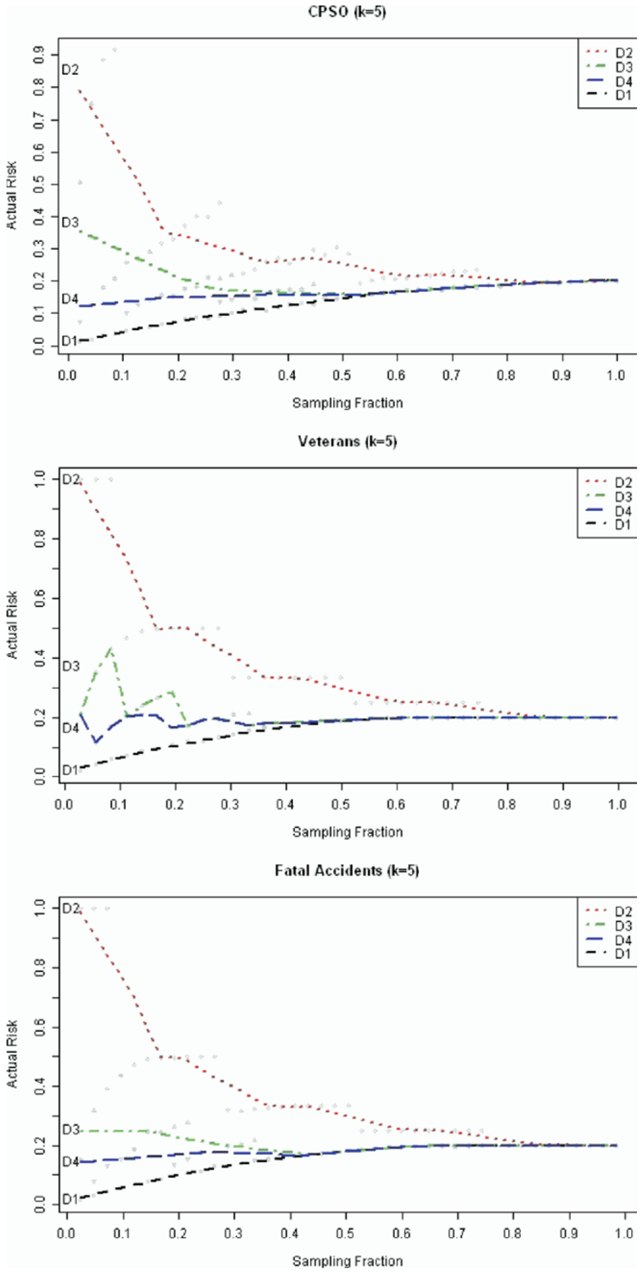


Figure 3. Plots of the actual risk ($\theta_{\max}$) against the sampling fraction for our data sets at $k = 5$. We use locally weighted scatterplot smoothing (lowess) to plot the curves.

The individual risk estimation approach (D2) results in particularly large actual risk values for sampling fractions as high as 0.6. Even though individual risk estimates may be relatively accurate *on average* across many equivalence classes, their use will not result in a reasonable level of protection against a smart intruder who will focus only on the smallest equivalence class.

Approach D3 is better, but still results in actual risk values above the threshold quite often and by a wide gap for sampling fractions of 0.3 and less. But the fact that D3 cannot maintain risk below the threshold for sampling fractions as high as 0.3 make it unsuitable for practical use.

The best approach is D4, whereby it does maintain the actual risk closest to and below the threshold risk of 0.2. Compared to D1, its actual risk is higher. But because it is below or very close to the threshold, its behaviour is consistent with what a data custodian would expect.

Figure 4 shows the normalized information loss in terms of the discernability metric. As expected, the D1 approach has the largest information loss among all four approaches, especially at lower sampling fractions (this is evidenced by the normalized discernability metric always having values below 100%). At higher sampling fractions all approaches tend to converge.

The hypothesis testing approach D4 has higher information loss than D3, but in many cases that difference is not very pronounced. But D4 is a significant improvement on D1, especially for low sampling fractions. For example, for the CPSO data set, D4 has 45% of the information loss of D1 at a sampling fraction of 0.1. Approach D2 has the lowest information loss. This is to be expected since the actual re-identification risk for D2 is often very high (as we saw in Figure 3).

While suppression is accounted for within the discernability metric, it is informative to consider the proportion of records suppressed by itself under each approach (Figure 5). The baseline approach results in a significant amount of suppression for small samples; in some cases as much as 50% of the records are suppressed. The D4 approach does reduce that percentage quite considerably, especially for small sampling fractions.

Discussion

We made explicit the two re-identification scenarios that k-anonymity was designed to protect against, known as prosecuter and journalist scenarios. The baseline k-anonymity model, which represents current practice, would work well for protecting against the prosecutor re-identification scenario. However, our empirical results show that the baseline k-anonymity model is very conservative in terms of re-identification risk under the journalist re-identification scenario. This conservatism results in extensive information loss. The information loss is exacerbated for small sampling fractions.

The reason for these results is that the appropriate disclosure control criterion for the journalist scenario is k-map, not k-anonymity. We then evaluated three methods that extend k-anonymity to approximate k-map. These can potentially ensure that the actual risk is close to the threshold risk. A hypothesis testing method based on the truncated-at-zero Poisson distribution ensures that the actual risk is quite close to the threshold risk, even for small sampling fractions, and therefore is a good approximation of k-map. It is a considerable improvement over the baseline k-anonymity approach because it provides good control of risk consistent with the expectations of a data custodian. Furthermore, this hypothesis testing approach always results in significantly less information loss than the baseline k-anonymity approach. This is an important benefit because we have shown

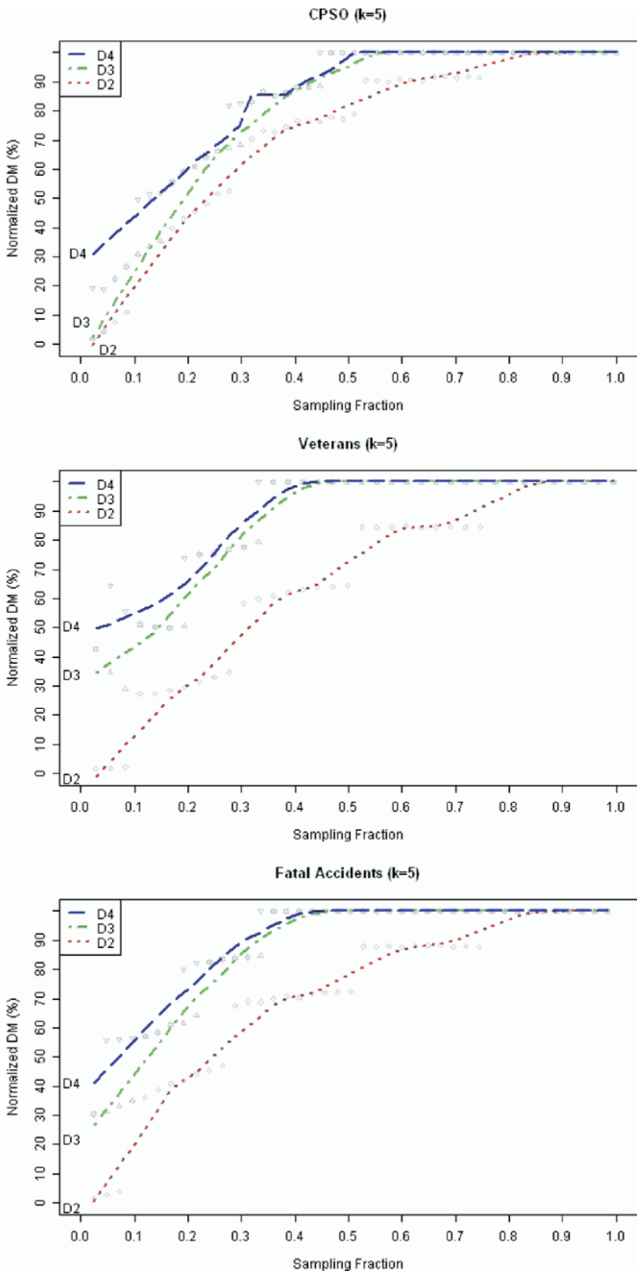


Figure 4. Plots of the normalized discernability metric (expressed as a percentage) against the sampling fraction. The values on the y-axis are relative to the baseline k-anonymity approach. For example, a value of 80% indicates the the information loss is 80% that of the baseline k-anonymity approach. We use locally weighted scatterplot smoothing (lowess) to plot the curves.

that a significant percentage of records would be suppressed using the baseline approach.

Suppression results in discarding data that was costly to collect and potentially result in a considerable loss of statistical power in any subsequent analysis. Furthermore, unless record suppression is completely random, it will bias analysis results.[92] If we take a simple example of a single quasi-identifier, records will be suppressed for the rare and extreme values on that variable. Therefore, by definition, the pattern of suppression will not be completely random.

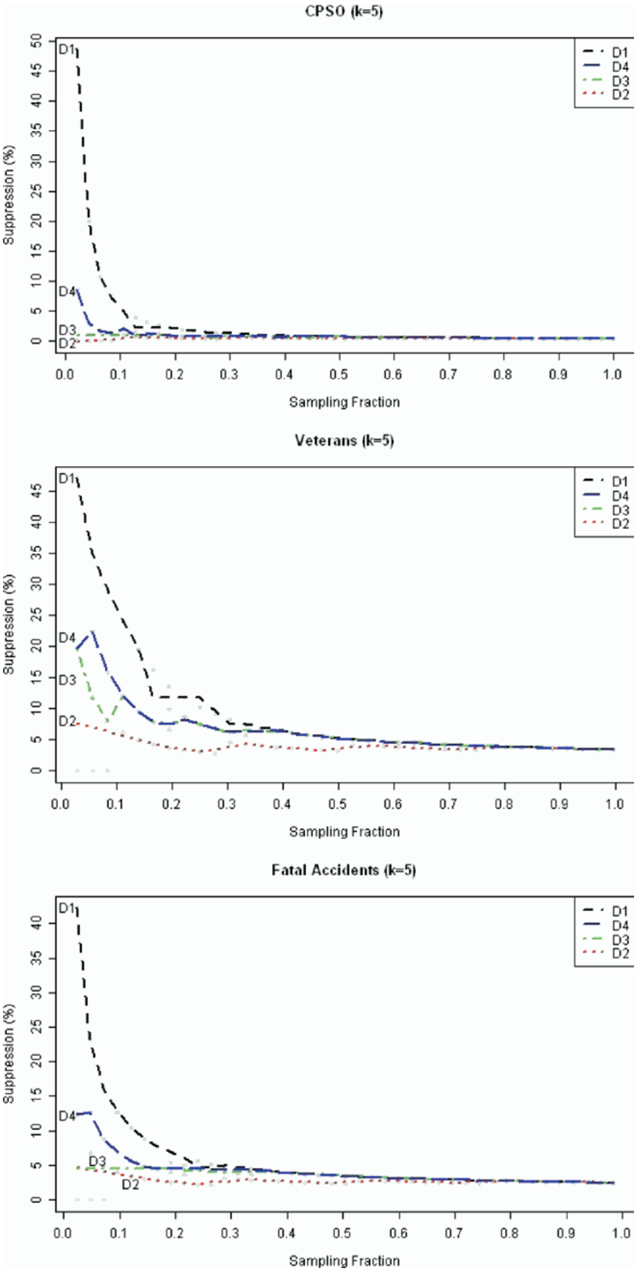


Figure 5. Plots of the percentage of suppressed records in a data set against the sampling fraction for the four different approaches. We use locally weighted scatterplot smoothing (lowess) to plot the curves.

Some k-anonymization algorithms suppress individual cells rather than full records. In practice, this may not have as much of a positive impact on the ability to do data analysis as one would hope. One common approach to deal with suppressed cells is complete case analysis (CCA), whereby only records without suppressed values are included in an analysis. Deletion of full records with any suppressed values is the default approach in most statistical packages[92] and is common practice in epidemiologic analysis.[93] It is known that CCA can result in discarding large proportions of a data set. For example, with only 2% of the values missing at random in each of 10 variables, one would lose 18.3% of the observations on average using CCA, and with 5 variables having 10% of their values missing at random, 41% of the observations would be lost with CCA, on average.[94] Another popular approach is available case analysis (ACA), whereby the records with complete values on the variables used in a particular analysis are used. For example, in constructing a correlation matrix different records are used for each pair of variables depending on the availability of both values. This, however can produce nonsense results.[94] Both CCA and ACA are only appropriate under the strong assumption that suppression is completely at random,[92,93] and we have noted above that with k-anonymity this will not be the case by definition. Therefore, full record or individual cell suppression are both detrimental to the quality of a data set.

Guidelines for Applying k-Anonymity

The way in which k-anonymity would be applied depends on the re-identification scenario one is protecting against. To protect against the prosecutor re-identification scenario, then k-anonymity should be used. If the prosecutor scenario is not applicable, then k-anonymity is not recommended, and k-map should be used instead (or our approximations of it using the hypothesis testing approach D4). If both scenarios are plausible, then k-anonymity should be used because this is the most protective. Therefore, being able to make a decision on whether the prosecutor scenario is applicable is important.

If we assume a threshold risk of 0.2, then under the prosecutor scenario the data custodian would just k-anonymize with $k = 5$. Under the journalist scenario the data custodian would determine k' using the hypothesis testing approach (D4) and then k-anonymize with $k = \min(k',5)$.

An intruder would only pursue a prosecutor re-identification scenario if s/he has certainty that the VIP has a record in ζ. There are three ways in which an intruder can have such certainty:[79,95]

1. The disclosed data set represents the whole population (e.g., a population registry) or has a large sampling fraction. If the whole population is being disclosed then the intruder would have certainty that the VIP is in the disclosed data set. Also, a large sampling fraction means that the VIP is very likely to be in the disclosed data set.
2. If it can be easily determined who is in the disclosed sample. For example, the sample may be a data set from an interview survey conducted in a company and it is generally known who participated in these interviews because the participants missed half a day of work. In such a case it is known within the company, and to an internal intruder, who is in the disclosed data set.
3. The individuals in the disclosed data set self-reveal that they are part of the sample. For example, subjects in clinical trials do generally inform their family, friends, and even acquaintances that they are participating in a trial. One of the acquaintances may attempt to re-identify one of these self-revealing subjects. However, it is not always the case that individuals do know that their data is in a data set. For example, for studies were consent has been waived or where patients provide broad authorization for their data or tissue samples to be used in research, the patients may not know that their data is in a specific data set, providing no opportunity for self-revealing their inclusion.

If any of the above conditions apply, then protecting against the prosecutor scenario is required. However, many epidemiologic and health services research studies, including secondary use studies, would not meet the criteria set out above. In such a case, protection against the journalist scenario with the D4 approach is recommended.

Relationship to Other Work

Re-identification risk is sometimes measured or estimated as the proportion of records that are unique in the population. Such uniqueness is then used as a proxy for re-identification risk. One approach for estimating population uniqueness from a sample uses the Poisson–gamma model with the α and β parameters estimated by the method of moments,[96,97] but it over-estimates with small sampling fractions and under-estimates as the sampling fraction increases.[98] Another method that uses sub-sampling performs well for larger sampling fractions.[99–101] More recent work developed probability models and estimators for two attack-based re-identification risk measures.[102] However, uniqueness measures of risk will by definition give an answer of zero for any k-anonymized data set, and therefore are inappropriate in this context.

Limitations

We limited our simulations to quasi-identifiers that have been demonstrated to be useful for re-identification attacks using public and semi-public data sources. There is evidence that information loss becomes unacceptably large as the number of quasi-identifiers increases, even for small values of k.[103] Therefore, had we used more quasi-identifiers, the information loss effects that we have shown would have been more pronounced.

There are other approaches that have been proposed for achieving k-anonymity that we did not consider, for example, local recoding.[104-107] With local recoding, observations may have different and overlapping response intervals. For instance, one observation may have an age of 27 recoded to the interval 20–29, and another observation may have an age of 27 recoded to the interval 25–35. This makes any data analysis of the k-anonymized data set more complex than having the same recoding intervals for all observations, and precludes the use of common and generally accepted statistical modeling techniques. Our implementation of k-anonymity used global recoding instead, and this ensures that response intervals are the same across all observations.

Conclusions

There is increasing pressure to disclose health research data, and this is especially true when the data has been collected using public funds. However, the disclosure of such data raises serious privacy concerns. For example, consider an individual who participated in a clinical trial having all of their clinical and lab data published in a journal web site accompanying the article on the trial. If it was possible to re-identify the records of that individual from this public data it would be a breach of privacy. Such an incident could result in fewer people participating in research studies because of privacy concerns, and if it happened in Canada, would be breaking privacy laws.

It is therefore important to understand precisely the types of re-identification attacks that can be launched on a data set and the different ways to properly anonymize the data before it is disclosed.

Anonymization techniques result in distortions to the data. Excessive anonymization may reduce the quality of the data making it unsuitable for some analysis, and possibly result in incorrect or biased results. Therefore, it is important to balance the amount of anonymization being performed against the amount of information loss.

In this paper we focused on k-anonymity, which is a popular approach for protecting privacy. We considered the two re-identification scenarios that k-anonymity is intended to protect against. For one of the scenarios, we showed that actual re-identification risk under the baseline k-anonymity is much lower than the threshold risk that the data custodian assumes, and that this results in an excessive amount of information loss, especially at small sampling fractions. We then evaluated three alternative approaches and found that one of them consistently ensures that the re-identification risk is quite close to the actual risk, and always has lower information loss than the baseline approach.

It is recommended that data custodians determine which re-identification scenarios apply on a case-by-case basis, and anonymize the data before disclosure using the baseline k-anonymity model or our modified k-anonymity model accordingly.

References ■

1. Fienberg S, Martin M, Straf M. Sharing Research Data. Committee on National Statistics, National Research Council 1985.
2. Hutchon D. Publishing raw data and real time statistical analysis on e-journals. Br Med J. 2001;322(3):530.
3. Are journals doing enough to prevent fraudulent publication? Can Med Assoc J 2006;174(4):431. Available at http://www.cmaj.ca/cgi/content/full/174/4/431/. Accessed July 24, 2008.
4. Abraham K. Microdata access and labor market research: The US experience. Allegmeines Statistisches Archiv. 2005;89:121–39.
5. Vickers A. Whose data set is it anyway? Sharing raw data from randomized trials. Trials 2006;7(15)
6. Altman D, Cates C. Authors should make their data available. BMJ. 2001;323:1069.
7. Delamothe T. Whose data are they anyway ? BMJ. 1996;312: 1241–2.
8. Smith GD. Increasing the accessibility of data. BMJ. 1994;308: 1519–20.
9. Commission of the European Communities. On scientific information in the digital age: Access, dissemination and preservation 2007.
10. Lowrance W. Access to collections of data and materials for health research: A report to the Medical Research Council and the Wellcome Trust. Medical Research Council and the Wellcome Trust 2006.
11. Yolles B, Connors J, Grufferman S. Obtaining access to data from government-sponsored medical research. NEJM. 1986; 315(26):1669–72.
12. Hogue C. Ethical issues in sharing epidemiologic data. J Clin Epidemiol. 1991;44(Suppl. I):103S–107S.
13. Hedrick T. Jutsifications for the sharing of social science data. Law and Human Behavior. 1988;12(2):163–71.
14. Mackie C, Bradburn N. Improving access to and confidentiality of research data: Report of a workshop, The National Academies Press; 2000.
15. Arzberger P, Schroeder P, Bealieu A, et al. Promoting access to public research data for scientific, economic, and social development. Data Science Journal. 2004;3(29):135–52.
16. Wager L, Krieza-Jeric K. Report of public reporting of clinical trial outcomes and results (PROCTOR) meeting. Canadian Institutes of Health Research 2008.

17. Organisation for Economic Co-operation and Development. Science, Technology and Innovation for the 21st Century 2004.
18. Organisation for Economic Co-operation and Development. Promoting Access to Public Research Data for Scientific, Economic, and Social Development: OECD Follow Up Group on Issues of Access to Publicly Funded Research Data 2003.
19. Commission regulation (EC) No 831/2002 o1 17 May 2002 on implementing council regulation (EC) No 322/97 on community statistics, concerning access to confidential data for scientific purposes. Official Journal of the European Communities 2002.
20. Kirwan J. Making original data from clinical studies available for alternative analysis. The Journal of Rheumatology. 1997; 24(5):822–5.
21. Chalmers I, Altman D. How can medical journals help prevent poor medical research? Some opportunities presented by electronic publishing The Lancet. 1999;353:490–3.
22. Eysenbach G, Sa E-R. Code of conduct is needed for publishing raw data. BMJ. 2001;323:166.
23. Canadian Institutes of Health Research. Policy on access to research outputs. 2007. Available at: http://www.cihr-irsc.gc.ca/e/34846.html. Archived at: http://www.webcitation.org/5XgxgoBzj.
24. Medical Research Council. MRC Policy on Data Sharing and Preservation 2006.
25. Economic and Social Research Council. ESRC Research Funding Guide 2008.
26. ERC Scientific Council Guidelines for Open Access. European Research Council 2007.
27. National Institutes of Health. Final NIH statement on sharing research data National Institutes of Health. Available at: http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html. Accessed July 24, 2008.
28. Cecil J, Boruch R. Compelled disclosure of research data: An early warning and suggestions for phsycologists. Law and Human Behavior. 1988;12(2):181–9.
29. Hansell S. AOL Removes Search Data on Group of Web Users New York Times. 8 August, 2006.
30. Barbaro M, Zeller Jr. T. A Face Is Exposed for AOL Searcher No. 4417749 New York Times. 9 August, 2006.
31. Zeller Jr. T. AOL Moves to Increase Privacy on Search Queries. New York Times. August 22, 2006.
32. Ochoa S, Rasmussen J, Robson C, Salib M. Reidentification of individuals in Chicago's homicide database: A technical and legal study. Massachusetts Institute of Technology 2001.
33. Narayanan A, Shmatikov V. Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset). University of Texas at Austin 2008.
34. Sweeney L. Computational disclosure control: A primer on data privacy protection: Massachusetts Institute of Technology, 2001.
35. The Southern Illinoisan v. Department of Public Health. Appelate Court of Illinois, Fifth District, 2004(No. 5-02-0836).
36. The Supreme Court of the State of Illionois. Southern Illinoisan vs. The Illinois Department of Public Health 2006.
37. Federal Court (Canada). Mike Gordon vs. The Minister of Health: Affidavit of Bill Wilson 2006.
38. Cavoukian A. Privacy concerns in preventing fraudulent publication. CMAJ. 2006;175(1):61–2.
39. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalisation and suppression. SRI International 1998.
40. Samarati P. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering. 2001;13(6):1010–27.
41. Sweeney L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002;10(5):557–70.
42. V. Ciriani, De Capitani di Vimercati SSF, Samarati P. k-Anonymity. Secure Data Management in Decentralized Systems: Springer, 2007.
43. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(5): 571–88.
44. Bayardo R, Agrawal R. Data Privacy through Optimal k-Anonymization Proceedings of the 21st International Conference on Data Engineering, 2005;217–28.
45. Iyengar V. Transforming data to satisfy privacy constraints. Proceedings of the ACM SIGKDD Int Conf Data Mining Knowledge Discov 2002;279–88.
46. Purdham K, Elliot M. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. Env Planning. 2007;39:1101–18.
47. Fefferman N, O'Neil E, Naumova E. Confidentiality and confidence: Is data aggregation a means to achieve both? J Public Health Pol. 2005;16:430–49.
48. Ohno-Machado L, Vinterbo S, Dreiseitl S. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance. J Am Med Inform Assoc. 2002;9(6):s115–s119.
49. Kamlet MS, Klepper S, Frank R. Mixing micro and macro data: Statistical issues and implication for data collection and reporting. Proceedings of the 1985 Public Health Conference on Records and Statistics, 1985.
50. Clause S, Triller D, Bornhorst C, Hamilton R, Cosler L. Conforming to HIPAA regulations and compilation of research data. Am J Health-Sys Pharm. 2004;61:1025–31.
51. Abrahamowicz M, du Berger R, Krewski D, Burnett R, Bartlett G, Tamblyn R, Leffondre K. Bias due to aggregation of individual covariates in the Cox regression model. Am J Epidemiol. 2004;160(7):696–706.
52. Marsh C, Skinner C, Arber S, Penhale B, Openshaw S, Hobcraft J, Lievesley D, Walford N. The case for samples of anonymized records from the 1991 census. J Royal Statist Soc, Ser A (Statistics in Society). 1991;154(2):305–40.
53. Elliot M, Dale A. Scenarios of attack: the data intruders perspective on statistical disclosure risk. Netherlands Official Statistics. 1999;14(Spring):6–10.
54. de Waal A, Willenborg L. A view on statistical disclosure control for microdata. Surv Methodol. 1996;22(1):95–103.
55. Greenberg B, Voshell L. Relating risk of disclosure for microdata and geographic area size. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1990;450–5.
56. Willenborg L, de Waal T. Elements of Statistical Disclosure Control, Springer-Verlag; 2001.
57. Benedetti R, Franconi L. Statistical and technological solutions for controlled data dissemination, Proceedings of New Techniques and Technologies for Statistics (vol. 1), 1998;225–32.
58. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. J Med Internet Res. 2006;8(4):e28.
59. Bell S. Alleged LTTE front had voter lists. National Post. July 22, 2006.
60. Bell S. Privacy chief probes how group got voter lists. National Post. July 25, 2006.
61. Freeze C, Clark C. Voters lists 'most disturbing' items seized in Tamil raids, documents say. Globe and Mail. May 7, 2008. Available at: http://www.theglobeandmail.com/servlet/story/RTGAM.20080507.wxtamilssb07/BNStory/National/home. Archived at: http://www.webcitation.org/5Xe4UWJKP.
62. Benedetti R, Capobianchi A, Franconi L. Individual risk of disclosure using sampling design information. Istituto nazionale di statistica (Italy) 2003.

63. Polettini S. Some remarks on the individual risk methodology. Joint ECE/Eurostat working session on statistical data confidentiality, 2003.
64. Polettini S, Stander J. A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. Privacy in Statistical Databases, 2004;247–61.
65. Di Consiglio L, Franconi L. Assessing individual risk of disclosure: An experiment. Joint ECE/Eurostat working session on statistical data confidentiality, 2003.
66. Rinott Y. On models for statistical disclosure risk estimation. Joint ECE/Eurostat Working Session on Statistical Data Confidentilaity 2003.
67. Capobianchi A, Polettini S, Lucarelli M. Strategy for the implementation of individual risk methodology in mu-argus: Independent units. Computational Aspects of Statistical Confidentiality (CASC) Project 2001.
68. Hundepool A, Wetering A, Ramaswamy R, Franconi L, Polettini S, Capobianchi A, Wolf P, Dominigo J, Torra V, R B, Giessing S. mu-Argus User Manual: Version 4.0.
69. Hundepool A. The ARGUS software. Joint ECE/Eurostat Working Sessionon Statistical Data Confidentiality 2003.
70. Franconi L, Polettini S. Individual risk estimation in mu-Argus: A review. Privacy in Statistical Databases, 2004;262–72.
71. Pannekoek J. Statistical methods for some simple disclosure limitation rules. Statistica Neerlandica. 1999;53(1):55–67.
72. Cameron A, Trivedi P. Regression Analysis of Count Data, Cambridge University Press; 1998.
73. Long S. Regression Models for Categorical and Limited Dependent Variables, Sage Publications; 1997.
74. Duncan G, Jabine T, de Wolf S. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, National Academies Press; 1993.
75. Cancer Care Ontario Data Use and Disclosure Policy. Cancer Care Ontario 2005.
76. Security and confidentiality policies and procedures. Health Quality Council 2004.
77. Privacy code. Health Quality Council 2004.
78. Privacy code. Manitoba Center for Health Policy 2002.
79. Subcommittee on Disclosure Limitation Methodology - Federal Committee on Statistical Methodology. Working paper 22: Report on statistical disclosure control. Office of Management and Budget 1994.
80. Statistics Canada. Therapeutic abortion survey. 2007. Archived at: http://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3209&lang=en&db=IMDB&dbg=f&adm=8&dis=2#b9 Accessed July 24, 2008.
81. College of Physicians and Surgeons of Ontario. Doctor Search. Available at: http://www.cpso.on.ca/Doctor_Search/dr_srch_hm.htm. Accessed July 24, 2008.
82. Paralyzed Veterans Association. Available at: http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html. Accessed July 24, 2008.
83. Department of Transportation. Fatal crash information. Available at: ftp://ftp.nhtsa.dot.gov/FARS/. Accessed July 24, 2008.
84. Sweeney L. Uniqueness of Simple Demographics in the US Population. Carnegie Mellon University, Laboratory for International Data Privacy 2000.
85. El Emam K, Jonker E, Sams S, Neri E, Neisa A, Gao T, Chowdhury S, Pan-Canadian De-Identification Guidelines for Personal Health Information. Report prepared for the Office of the Privacy Commissioner of Canada, 2007. Available at: http://www.ehealthinformation.ca/documents/OPCReportv11.pdf. Archived at: http://www.webcitation.org/5Ow1Nko5C.
86. LeFevre K, DeWitt D, Ramakrishnan R. Mondrian multidimensional k-anonymity. Proceedings of the 22nd International Conference on Data Engineering, 2006.
87. Hore B, Jammalamadaka R, Mehrotra S. Flexible anonymization for privacy preserving data publishing: A systematic search based approach. Proceedings of SIAM International Conference on Data Mining, 2007.
88. Xu J, Wang W, Pei J, Wang X, Shi B, Fu A. Utility-based anonymization for privacy preservation with less information loss. ACM SIGKDD Explorations Newsletter. 2006;8(2):21–30.
89. Nergiz M, Clifton C. Thoughts on k-anonymization. Second Intenational Workshop on Privacy Data Management, 2006.
90. Polettini S. A note on the individual risk of disclosure. Istituto nazionale di statistica (Italy) 2003.
91. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-Diversity: Privacy Beyond k-Anonymity. International Conference on Data Engineering, 2006.
92. Little R, Rubin D. Statistical Analysis With Missing Data, John Wiley & Sons; 1987.
93. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol. 1991;134(8):895–907.
94. Kim J, Curry J. The treatment of missing data in multivariate analysis. Soc Methods Res. 1977;6:215–40.
95. Willenborg L, de Waal T. Statistical Disclosure Control in Practice, Springer-Verlag; 1996.
96. Bethlehem J, Keller W, Pannekoek J. Disclosure control of microdata. J Am Stat Assoc. 1990;85(409):38–45.
97. Skinner C, Holmes D. Estimating the re-identification risk per record in microdata. J Off Stat. 1998;14(4):361–72.
98. Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata. J Off Stat. 1998;14(1):79–95.
99. Zayatz L. Estimation of the percent of unique population elements on a microdata file using the sample. US Bureau of the Census 1991.
100. Greenberg B, Voshell L. The geographic component of disclosure risk for microdata. Bureau of the Census 1990.
101. Willenborg L, Mokken R, Pannekoek J, Microdata and disclosure risks. Proceedings of the Annual Research Conference of US Bureau of the Census, 1990;167–180.
102. Skinner G, Elliot M. A measure of disclosure risk for microdata. J Royal Stat Soc (Ser B). 2002;64(Part 4):855–67.
103. Aggarwal C. On k-anonymity and the curse of dimensionality. Proceedings of the 31st VLDB Conference, 2005.
104. Du Y, Xia T, Tao Y, Zhang D, Zhu F. On Multidimensional k-Anonymity with Local Recoding Generalization. IEEE 23rd International Conference on Data Engineering, 2007;1422–4.
105. Xu J, Wang W, Pei J, Wang X, Shi B, Fu A. Utility-based anonymization using local recoding. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
106. Wong R, Li J, Fu A, Wang K. (alpa,k)-Anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
107. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A. Approximation algorithms for k-anonymity. J Priv Technol 2005.