

A Method for Evaluating Marketer Re-identification Risk

Fida Kamal Dankar
 CHEO Research Institute
 401 Smyth Rd
 Ottawa, ON, Canada
 +1-613-737 7600 ext 4147
 fdankar@ehealthinformation.ca

Khaled El Emam
 CHEO Research Institute & University of Ottawa
 401 Smyth Rd
 Ottawa, ON, Canada
 +1-613-737 7600 ext 4181
 kelemam@uottawa.ca

ABSTRACT

Disclosures of health databases for secondary purposes is increasing rapidly. In this paper, we develop and evaluate a re-identification risk metric for the case where an intruder wishes to re-identify as many records as possible in a disclosed database. In this case, the intruder is concerned about the overall matching success rate. The metric is evaluated on public and health datasets and recommendations for its use are provided.

Categories and Subject Descriptors

K.4.1 [Computers and Society]: Public Policy Issues – Privacy.

Keywords

Identity disclosure, privacy, re-identification risk, disclosure control.

1. INTRODUCTION

As more ostensibly de-identified health data sets are disclosed for secondary purposes [1, 2], it is becoming important to measure the risk of patient re-identification (i.e., identity disclosure) objectively, and manage that risk. Previous risk measures focused mostly on the case where a *single* patient is being re-identified [3]. With these previous measures, the patient with the highest re-identification risk represented the risk for the whole data set.

In practice, an intruder may re-identify more than one patient. The potential harm to the patients and the custodian would be much higher if many patients are re-identified as opposed to a single one. Therefore, there will be scenarios where the data custodian is interested in assessing the number (or proportion) of records that could be correctly re-identified. There is a dearth of generally accepted re-identification risk measures for the case where an intruder attempts to re-identify *all* patients (or as many patients as possible) in a data set.

The variables that can potentially re-identify patient records in a disclosed data set are called the *quasi-identifiers* (qids) [4]. Examples of common quasi-identifiers are [5-9]: dates (such as, birth, death, admission, discharge, visit, and specimen collection),

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PAIS'10, March 22-26, 2010, Lausanne, Switzerland.
 Copyright 2010 ACM 978-1-60558-990-9/10/03...\$10.00.

race, ethnicity, languages spoken, aboriginal status, and gender. An intruder would attempt to re-identify all patients in a disclosed data set by matching against an *identification database*. An identification database would contain the qids as well as directly identifying information about the patients (e.g., their names and full addresses). There are two scenarios where this could plausibly occur.

1.1 Public Registries

In the US it is possible to obtain voter lists for free or for a modest fee in most states [10]. A voter list contains voter names and addresses, as well as their basic demographics, such as their date of birth, and gender. Some states also include race and political affiliation information. A voter list is a good example of an identification database.

Consider the example in Figure 1 of prescription records. Retail pharmacies in the US and Canada sell these records to commercial data brokers [11, 12]. These records include the basic patient demographics. An intruder can obtain voter lists for the specific county where a pharmacy resides and match them with the prescription records to potentially re-identify many patients.

In Canada voter lists are not (legally) readily available. However, other public registries exist which contain the basic demographics on large segments of the population [7], and can serve as suitable identification databases.

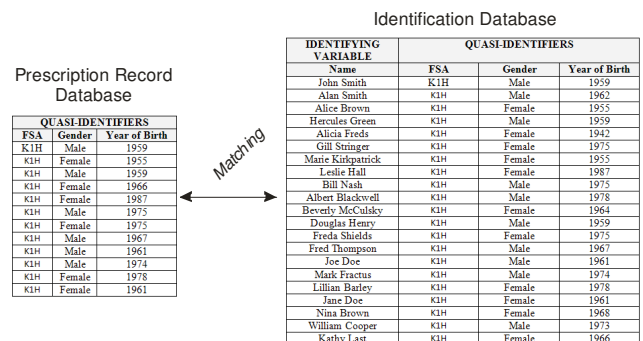


Figure 1: Example of a prescription record database being disclosed containing patient demographics being matched against a population registry (identification database) which an intruder has access to, e.g., a voter list. The prescription database is a sample of the population registry.

1.2 Insider Attacks

In many healthcare provider organizations the identity of patients and their basic demographics are broadly accessible to staff, but the clinical information is under stricter access controls. A staff member with access to such demographics would potentially be able to re-identify all patients in disclosed data sets quite easily by matching records on the demographics. For example, a recent study described the disclosure of hospital pharmacy records to a commercial data broker [13]. This data set included information about when the patient was admitted, their gender, age, and residential geographic information. A staff member with access to *only* the Admissions, Discharge, and Transfer database for the hospital would have those same variables and can potentially use them to re-identify the patients in the prescription records database, and hence their diagnosis and prescriptions.

In the US, genotype and phenotype data collected from healthcare provider institutions for genome wide association studies are being made more broadly available [14]. A recent study has demonstrated how an insider from one of the source institutions can re-identify patients in the DNA database with certainty [15].

Privacy and security breaches by insiders are relatively common [16-20]. In a recent survey a quarter of healthcare IT professionals identified a data breach by an internal person within the past year [21]. Internal breaches come from people who have (or recently had) legitimate access to the data, but whose behavior jeopardizes the organization's security and privacy. One survey found that 27% of breaches were due to insiders [22]. In a 2007 Computer Security Institute survey, 59% of the respondents claim that they witnessed abuse of the network resources by an insider [23], and the 2009 survey shows that a significant percentage of monetary losses from breaches were due to malicious and non-malicious insiders [24]. These could involve employees who have access to critical data and engage in unauthorized explorations; they could be individuals who enter an organization with the intent to commit some kind of fraud, or some naïve employees who lose their laptop after they put sensitive information on it. There are also increasing cases of ex-employees taking or keeping personal information about clients/patients and other employees after they leave their employer (because they were laid-off, resigned, or fired) [25]. Recent analyses of breach data indicate that 20% are internal breaches [26], 34% of security and IT leadership in organizations worldwide estimated that the likely source of security incidents in the last year was an employee and 16% responded that it was an ex-employee [27], and in the US and Canada approximately 24% of data breaches in medical establishments (including hospitals and health insurance claims processors) are due to insiders and 32% of breaches involving medical records are due to insiders (between January 2007 and June 2009) [28]. Overall, investigators reported a rapid increase in offences committed by insiders [29].

The type of attacks from the inside can have a huge impact on the organization. It was reported that an outside attack costs the organization on average \$56,000 while an insider breach costs on average \$2.7 million [29].

1.3 Marketer Risk

In this paper, we develop and evaluate a re-identification risk metric for the case where an intruder wishes to re-identify as many

records as possible in the disclosed database. We assume that the intruder lacks any additional information apart from the matching quasi-identifiers.

The intruder is not interested in knowing which records from the disclosed data set were correctly re-identified. Instead, the more important concern is that the *proportion* of records in the disclosed data set that are correctly re-identified is sufficiently high.

We will call the (expected) proportion of records that are correctly re-identified the *marketer risk* metric. This term is used to represent the archetypical scenario where the intruder is matching the two databases for the purposes of marketing to the individuals in the disclosed database.

There are two cases where the marketer risk needs to be computed. The first is when the disclosed database has the same individuals as the identification database. For example, if the disclosed prescription database in Figure 1 contains *all* voters, and is matched with the voter list. The second, and more likely, case is when the disclosed database is a subset/sample from the identification database. For example, as in Figure 1, the disclosed prescription records database is a subset of the population covered by the voter list.

We assume that the data custodian wishes to compute marketer risk to decide whether to release a database as is or if disclosure control actions are needed. It is quite unlikely that the custodian will have the identification database. For example, a pharmacy chain that is selling its prescription records will not purchase all voter lists across the states it operates in to create a population identification database to determine whether the marketer risk is too high or not. Therefore, the risk measure has to be computed only from the disclosed database.

In the first case above the disclosed database is the same as the identification database, therefore access to an identification database is not an issue. However, in the second case where the custodian only has a sample, there are a number of reasons why the data custodian would not have an identification database. Often, a population database is expensive to get hold of. 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 [6]. 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 [6]. There is at least one known incident where a charity allegedly supporting a terrorist group has been able to obtain Canadian voter lists for fund raising [30-32]. A legitimate data custodian would not engage in such acts.

1.4 Our Contributions

In this paper we first consolidate different metrics from previous research and show that they all describe the same marketer risk measure. Then we formulate an estimator for the second case above where the disclosed database is a sample, and we extensively evaluate it on five data sets. In addition, we provide guidance on when and how to apply the marketer risk metric for disclosure control purposes.

2. METHODS

2.1 Notation

We denote the set of the records in the disclosed patient database as U and the set of records in the identification database as D , and $U \subseteq D$. Let $|U| = n$, and $|D| = N$, which gives the total number of records in each database. Each record pertains to a unique patient. The set of qids is denoted by $Z = \{z_1, \dots, z_p\}$, and let $|z_i|$ be the number of unique values that the specific qid, z_i , takes in the actual data set.

The discrete variable formed by cross-classifying all possible values on the qids is denoted by X , with the values denoted by $1, \dots, J$. Each of these values corresponds to a possible

combination of values of the qids (note that $\prod_{i=1}^p |z_i| = J$). We

call the records with the value $j \in \{1, \dots, J\}$ an *equivalence class*. For example, all records in a data set about 17 year old males admitted on 1st January 2008 are an equivalence class.

In practice, however, not all possible equivalence classes will appear in the data set. We therefore denote by \tilde{J} the number of *actual* different values that appear in the data. Let X_i denote the value of X for patient i . The frequencies for different values of \tilde{J} are given by $F_j = \sum_{i \in D} I(X_i = j)$, where

$j \in \{1, \dots, \tilde{J}\}$ and $I(\cdot)$ is the indicator function. Similarly, we define $f_j = \sum_{i \in U} I(X_i = j)$ where $j \in \{1, \dots, \tilde{J}\}$.

We define the set of records in an equivalence class in U by g_j , and the set of records in an equivalence class in D by G_j .

This also means that $|g_j| = f_j$ and $|G_j| = F_j$ for $j \in \{1, \dots, \tilde{J}\}$.

2.2 Measuring Re-identification Risk

An intruder tries to match the two databases one equivalence class at a time. In other words, for every $j \in \{1, \dots, \tilde{J}\}$, the intruder matches the records in g_j to the records in G_j . Lacking any additional information apart from the matching qids, the intruder can match any two records from the two corresponding equivalence classes at random with equal probability. The intruder has the option to consider one-to-one mappings (i.e., no two records in g_j can be mapped to the same record in G_j) or not.

In what follows, we will prove that in both cases (i.e., when considering only one-to-one mappings or not) the expected

number of records that can be correctly matched is $\frac{f_j}{F_j}$ per equivalence class, and the expected proportion of records that can be re-identified from the disclosed database is $\frac{1}{n} \times \sum_{j=1}^{\tilde{J}} \frac{f_j}{F_j}$.

Theorem 1. The expected proportion of U records that can be disclosed in a random mapping from U to D is.

$$\lambda = \sum_{j=1}^{\tilde{J}} \frac{f_j / F_j}{n} \dots\dots\dots(1)$$

Note that if $n = N$ then $\lambda = \frac{\tilde{J}}{N}$.

Proof. We consider two cases:

1. the first case is when only one to one random mappings are used, and
2. the second case is when any random mapping is used.

A. One to one mappings:

We prove first that the expected number of records that can be re-identified from any equivalence class g_j is $\frac{f_j}{F_j}$:

Assume that m records in g_j have been matched to m different records in G_j for some $m \in \{1, \dots, f_j - 1\}$, then the probability that the $m + 1^{\text{th}}$ record in g_j (which we denote by r) will be correctly matched to its corresponding record in G_j (the corresponding match is denoted by s), or P_{rs} can be calculated as follows:

$P_{rs} = P(\text{record } s \text{ is not matched to any of the previously matched } m \text{ records}) P(r \text{ is assigned to } s)$

$$= \frac{\binom{F_j - 1}{m}}{\binom{F_j}{m}} \frac{1}{F_j - m} = \frac{F_j - m}{F_j} \frac{1}{F_j - m} = \frac{1}{F_j}$$

Hence the expected number of records that would be disclosed

from any equivalence class g_j is $\sum_1^{f_j} \frac{1}{F_j} = \frac{f_j}{F_j}$.

Now, the expected total number of records correctly matched

becomes: $\sum_{j=1}^{\tilde{J}} \frac{f_j}{F_j}$, and the proportion of records correctly

matched is $\sum_{j=1}^{\tilde{J}} \frac{f_j/F_j}{n}$.

B. Random Mappings:

We prove first that the expected number of records that can be

disclosed from any equivalence class g_j is $\frac{f_j}{F_j}$:

Let a be any record in g_j , the probability that a is correctly

matched in a random mapping from g_j to G_j is $\frac{1}{F_j}$ (because

a could be matched to any record in F_j)

Now the expected number of records that would be disclosed from

any equivalence class g_j is $\sum_{j=1}^{\tilde{J}} \frac{1}{F_j} = \frac{f_j}{F_j}$

Hence the proportion of records that can be disclosed is again

$$\sum_{j=1}^{\tilde{J}} \frac{f_j/F_j}{n}$$

2.3 Relationship to Previous Work

In a recent study, the authors consider this matching problem from the record linkage perspective [33]. They discuss the case where the linking procedure for the records in g_j and G_j is random (in other words, they assume that the intruder has no background information), they only consider one to one mappings from g_j to G_j , and they only consider the case where $n = N$, i.e. when $f_j = F_j$ for all j . In that context, they prove that the probability of re-identifying exactly R individuals from G_j is:

$$\sum_{v=0}^{F_j-R} \frac{(-1)^v}{R!} \frac{1}{v!}. \text{ The expected number of re-identified records}$$

$$\text{from an equivalence class } G_j \text{ is then: } \sum_{R=0}^{F_j} R \sum_{v=0}^{F_j-R} \frac{(-1)^v}{R!}$$

which, turns out to be equal to 1. Hence, the expected total proportion of records re-identified in the identification database is

$$\text{equal to } \frac{\tilde{J}}{N}.$$

In another study, Truta et al [34] presented a measure of disclosure risk that considers the distribution of the non-unique records in the sample. The measure represents the record linkage success probability for all records in the sample. The measure is

the same as ours: $\sum_{j=1}^{\tilde{J}} \frac{f_j/F_j}{n}$, and was presented as a

generalization of the sample and population uniqueness measure

$$[34]: \sum_{j:F_j=1} \frac{f_j}{n}.$$

In the case where the disclosed database is a sample of the identification database as illustrated in Figure 1 (i.e., $U \subset D$), the data custodian often does not have access to an identification database to compute the marketer risk before disclosing the data.

In such a case we need to estimate the marketer risk, $\hat{\lambda}$. The values of f_j would be known to the data custodian, therefore we

need to estimate the values $\frac{1}{F_j}$ using only the information in

the disclosed database. In the remainder of this paper we evaluate three different estimators for doing so.

2.4 ESTIMATORS

Three estimators can be used to operationalize the marketer risk metric when only a sample is being disclosed: the Argus estimator [35], the Poisson log-linear model [36], and the negative binomial model [37, 38].

Recall that N denotes the total population number, and n the size of the sample. Denote by p_j the probability that a member of the class G_j is sampled (i.e., belongs to g_j), and by γ_j the probability that a member of the population belongs to the equivalence class G_j .

2.4.1 Argus

Mu-Argus [35] proposes a model where $F_j | f_j$ is a random variable with a negative binomial distribution, where f_j is the number of successes with the probability of a success being p_j :

$$P(F_j = H | f_j) = \binom{H-1}{f_j-1} p_j^{f_j} (1-p_j)^{H-f_j}$$

$$H \geq f_j > 0$$

With the above assumptions, the expected value of $\frac{1}{F_j}$ is

given by:

$$E\left(\frac{1}{F_j} \middle| f_j\right) = \sum_{i=f_j}^{\infty} \frac{1}{i} \Pr(F_j = i | f_j) \dots\dots\dots(2)$$

Equation (2) can be calculated using the moment generation function $M_{F_j|f_j}$ [39] as follows:

$$E\left(\frac{1}{F_j} \middle| f_j\right) = \int_0^{\infty} M_{F_j|f_j}(-t) dt = \int_0^{\infty} \left\{ \frac{p_j e^{-t}}{1 - (1-p_j)e^{-t}} \right\}^{f_j} dt$$

To estimate $E\left(\frac{1}{F_j}\right)$, we need to first estimate p_j . In [35],

each record i in the sample is assumed to have a weighting factor w_i (also known as inflation factor) which represents the number of units in the population similar to unit i . The authors in [35] also proposed: $\hat{p}_j = \frac{f_j}{\hat{F}_j^D}$ where $\hat{F}_j^D = \sum_{i:j(i)=j} w_i$ is the initial estimate for the population, where $j(i) = j$ indicates that record i belongs to g_j .

In our paper, since the weight factors w_i are unknown, we assume that p_j is constant across all equivalence classes and that

$$p_j = \frac{n}{N}.$$

Note that the estimated value for F_j depends only on f_j and is independent of the sample frequency in the other classes (i.e., there is no learning from other cells). Hence the information that one gains from the frequencies in neighboring cells is not used. However Argus has the advantage of being monotonic and simple to calculate.

2.4.2 Poisson log-linear model

In this model, the F_j 's are realizations of independent Poisson random variables with mean $N\gamma_j$: $F_j | \gamma_j \sim \text{Poisson}(N\gamma_j)$. Assuming that the sample is drawn by Bernoulli sampling with probability p_j , we obtain:

$$P(F_j = H | f_j) = \frac{1}{(H - f_j)!} (N\gamma_j(1-p_j))^{H-f_j} e^{-N\gamma_j(1-p_j)}$$

$$H \geq f_j > 0$$

Hence $E_{p_j}\left(\frac{1}{F_j} \middle| f_j\right)$ depends on f_j , γ_j and p_j . Which

can be calculated using the moment generation function $M_{F_j|f_j}$

$$[39] \text{ as: } E_{p_j}\left(\frac{1}{F_j} \middle| f_j\right) = \int_0^{\infty} e^{-tf_j} e^{N\gamma_j(1-p_j)(e^{-t}-1)} dt$$

Usually, a simple random sampling design is assumed where $n = p_j N$. To estimate the parameters γ_j , a log-linear model is used.

Log linear modeling consists of fitting models to the observed frequency (f_j) in the sample. The goodness of fit of the observed frequencies to the expected frequencies (u_j) is then

$$\text{computed. The estimate for } \gamma_j \text{ is then set to } \frac{u_j}{p_j}.$$

The log linear modeling approach uses data from neighborhood cells to determine the risk in a given cell (i.e., the estimated value of F_j does not depend only on f_j), the extent of this dependence is a function of the log-linear model used.

The choice of the model is crucial in providing good risk estimates. Skinner and Shlomo [36] showed through empirical work that for large and sparse data, no standard approach for model assessment works, so they present a novel approach for model assessment. The goodness of fit criterion was designed to detect underfitting (overestimation). Knowing that the independence model always leads to overestimation, and that overestimation decreases as we add more and more dependencies, a forward search algorithm was used [36]:

However, the approach in [36] is based on fitting the equivalence classes in the sample that are of size 1 (i.e., for $f_j = 1$), as the risk they are mainly interested in is the risk due to sample uniques.

The goodness of fit measure they developed [36] shows the impact of underfitting that is due to model misspecification. In other words, it represents the bias arising from the difference between the estimated γ_j , say $\hat{\gamma}_j$, and the actual γ_j as follows:

$$B_1 = \sum_j E(I(f_j = 1)) [h(\hat{\gamma}_j) - h(\gamma_j)] \text{ where } h(\gamma_j)$$

is the disclosure risk due to uniques in the sample:

$$h(\gamma_j) = \sum_{f_j=1} \frac{1/F_j}{N}.$$

In our case, since the risk measure entails the risk due to any equivalence class size, we generalized the Skinner-Shlomo goodness of fit measure to any fixed equivalence class size. We also generalized their method to cover all equivalence class sizes as described below.

For every equivalence class size in the sample, say s , we search for the log-linear model that presents a good fit for these equivalence classes using an iterative method [36]. Once a good fit is found, we compute the portion of the risk that is due to the

equivalence classes of size s , i.e. $\sum_{f_j=s} \frac{s/F_j}{N}$. We repeat the

procedure, fitting different log-linear models for every equivalence class size until we cover all class sizes present in the

sample, at which time the overall risk would have been calculated. The goodness of fit measure that we use for the different equivalence class sizes is a generalization of the uniqueness goodness of fit B_1 introduced in [36]:

If we denote by h^k the disclosure risk due to equivalence class of size k , in other words $h^k(\gamma_j) = \sum_{f_j=k} \left(\frac{k/F_j}{N} \right)$, then to

measure the model misspecification in equivalence classes of size k we use: $B_k = \sum_j E(I(f_j = k)) [h^k(\hat{\gamma}_j) - h^k(\gamma_j)]$.

2.4.3 Negative binomial model

In this model, a prior distribution for γ_j is assumed: $\gamma_j \sim \text{Gamma}(\alpha_j, \beta_j)$. The population cell frequencies F_j are independent Poisson random variables with mean $N\gamma_j$: $F_j | \gamma_j = \text{Poisson}(N\gamma_j)$.

It is often assumed that α is constant with $\alpha\beta = 1/\tilde{J}$, thus ensuring that $E(\sum \gamma_j = 1)$,

Bethlehem et al [37] considered only the case of sampling with equal probabilities, $n = \tilde{p}N$. Under these assumptions we get:

$$P(F_j = H | f_j) = \binom{\alpha + H - 1}{H - f_j} \left(\frac{Np_j + 1/\beta}{N + 1/\beta} \right)^{\alpha + f_j} \left(\frac{N(1 - p_j)}{N + 1/\beta} \right)^{H - f_j}$$

$H \geq f_j > 0$

The expected value of $1/F_j$ can be calculated from the above equation using the moment generation function $M_{F_j|f_j}$ [39] as

$$\text{follows: } E\left(\frac{1}{F_j} | f_j\right) = \int_0^\infty e^{-tf_j} p^{\alpha + f_j} \{1 - (1-p)e^{-t}\}^{-\alpha - f_j} dt$$

Notice that the expected value of $1/F_j$ depends on α .

The authors in [38] obtain an estimate for α , which includes estimating the variance for f_j and the fact that $\alpha\beta = 1/\tilde{J}$.

One of the difficulties of this model is the need to define the number of cells \tilde{J} in the population table. But since in most cases the population is not known, we used the estimator in [40] to estimate the number of classes \tilde{J} in the population.

2.5 Empirical Comparison of Estimators

Our objective is to evaluate the three methods described above for estimating the $1/F_j$ term in equation (1), and compare the

performance of the resulting $\hat{\lambda}$ marketer risk estimate relative to the actual marketer risk value. We therefore performed a simulation study to evaluate $\hat{\lambda}$ using each of the three population estimators relative to the actual λ .

Data Set	Quasi-identifiers	λ
FARS: fatal crash information database from the department of transportation; n=27,529	<ul style="list-style-type: none"> Year (21) Age (99) Race (19) Drinking Level (4) 	0.229
Adult (US Census); n=30,162	<ul style="list-style-type: none"> Age (72) Education (16) Race (5) Gender (2) 	0.104
Emergency department at children's hospital (6 months); n=25,470	<ul style="list-style-type: none"> Postal Code - 2 chars (105) Age (42) Gender (2) 	0.033
Niday (provincial birth registry); n=57,679	<ul style="list-style-type: none"> Postal Code - 3 chars (678) Date of Birth - mth/yr (7) Maternal Age (42) Gender (2) 	0.687
Pharmacy prescriptions for inpatients from a children's hospital; n=3,507	<ul style="list-style-type: none"> Gender (2) Age (22) Postal Code - 3 chars (154) Length of Stay (89) 	0.75

Table 1: Summary of the five data sets that were used for the simulation. The first column includes the number of records (the n value). The second column includes the quasi-identifiers and the number of equivalence classes that they have, and the third column is the actual value of marketer risk. The data sets were generalized to provide variation in the actual marketer risk value from quite low to quite high.

The five data sets which we used in our analysis are summarized in Table 1. Each data set is treated as the population and two thousands five hundreds random samples were drawn from it at five different sampling fractions (0.1 to 0.9 in increments of 0.2).

For each sample we estimated marketer risk and computed the relative error:

$$RE = \frac{\hat{\lambda} - \lambda}{\lambda} \dots\dots\dots(3)$$

The mean relative error was computed across all of the samples.

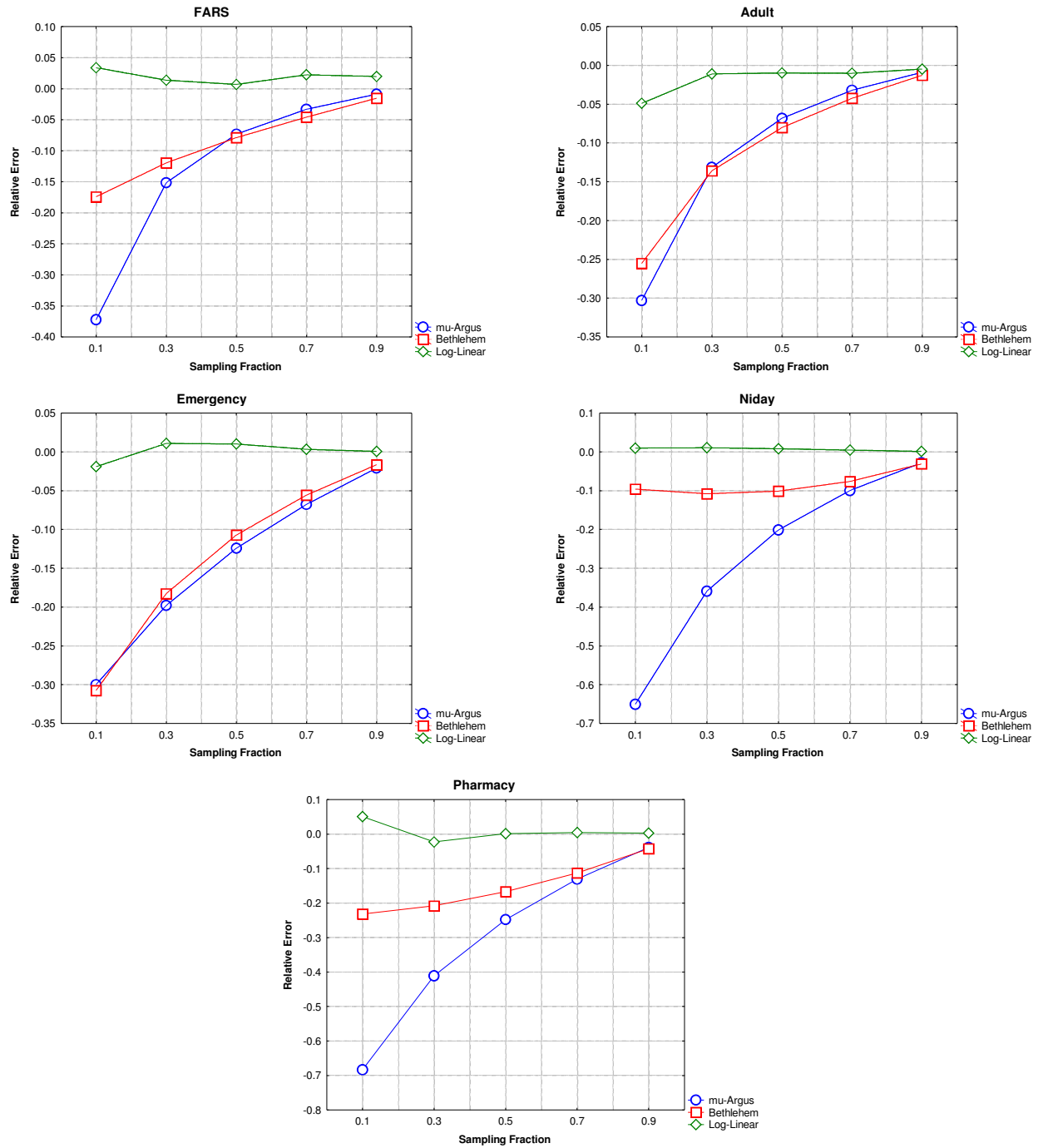


Figure 2: Graphs showing the relative error for each of the five data sets as we varied the sampling fraction from 0.1 to 0.9.

3. RESULTS

The results for the Adult data set in terms of the relative error (equation 3) are shown in Figure 2 for the three estimators. As can be seen, the log-linear modeling approach has a significantly lower relative error than mu-Argus and the Bethlehem estimators. This is the case across all sampling fractions and data sets.

4. DISCUSSION

Our results provide compelling evidence that, in the context of estimating marketer risk for a sample, the log-linear modeling approach is the most accurate in relative and absolute terms across multiple data sets of varying size and nature.

4.1 Application of the Marketer Risk Measure

An important question is how does a data custodian decide when is the expected proportion of records that would be correctly re-identified too high? We can provide some guidance based on precedents. Previous disclosures of cancer registry data have deemed thresholds of 5% and 20% of high risk records as acceptable for public release and research use respectively [41-43]. These can be used as a basis for setting acceptability thresholds for marketer risk values.

Another practical consideration is that the modified log-linear modeling approach we have used is computationally intensive and for very large J such a computation may not be feasible. In that case we recommend using the negative binomial estimator as it, in general, still outperforms the Argus estimator across the spectrum of sampling fractions.

4.2 Relationship to Other Risk Measures

Two other risk measures for identity disclosure have been defined [3]. The first is *prosecutor* risk, which is applicable when

$U = D$, and is computed as: $R_p = \frac{1}{\min_j(f_j)}$. The

second is *journalist* risk, which is applicable when $U \subset D$, and

is computed as: $R_j = \frac{1}{\min_j(F_j)}$. In both of these cases the

risk measure captures the worse case probability when re-identifying a *single* record, whereas for marketer risk we are evaluating the expected number (proportion) of records that would be correctly re-identified. Another important difference is that marketer risk does not help identify which records in U are likely to be re-identified. However, with journalist and prosecutor risk measures it is possible to identify the highest risk records and focus disclosure control action only on those.

4.3 Controlling Marketer Risk

Currently there are no algorithms specifically designed to control marketer risk. However, we can use existing k-anonymity algorithms to control marketer risk.

Let's assume that an intruder wishes to ensure that marketer risk is below some threshold, say τ . Then

$$\frac{1}{n} \sum_j \frac{f_j}{F_j} \leq \left(\frac{1}{\min_j(F_j)} \cdot \frac{\sum_j f_j}{n} \right) = \frac{1}{\min_j(F_j)} \quad \dots\dots(4)$$

Therefore, by ensuring that $R_j \leq \tau$ we can also ensure that the marketer risk is below that threshold. Any k-anonymity algorithm can be used to guarantee that inequality [3].

A disadvantage of using k-anonymity algorithms is that they may cause more de-identification than necessary. The marketer risk value can be quite a bit smaller than R_j in practice. For example, consider a population data set with 3 equivalence classes $F_j \in \{5, 20, 23\}$ and the sample consisting of uniques. In this case the marketer risk value would be half the R_j value. Therefore, while using existing k-anonymity algorithms is a suitable approach for now, it behooves the research community to develop algorithms that directly manage marketer risk.

4.4 When to Use Marketer Risk ?

If an intruder has an identification database, he can use it for re-identifying a single individual or for re-identifying as many individuals as possible. In the former case either the prosecutor or journalist risk metrics should be used, and in the latter case the marketer risk metric should be used. Therefore, the selection of a risk measure will depend on the motive of the intruder. While discerning motive is difficult, there will be scenarios where it is clear that marketer risk is applicable and represents the primary risk to be assessed and managed.

One scenario involves an intruder who is motivated to market a product to all of the individuals in the disclosed database. In that case the intruder may use an identification database, say a voter list, to re-identify the individuals. The intruder does not need to know which records were re-identified incorrectly because the incremental cost of including an individual in the marketing campaign is low. As long as the expected number of correct re-identifications is sufficiently high, that would provide an adequate return to the intruder. A data custodian, knowing that a marketing potential exists, would estimate marketer risk and may adjust it down to create a disincentive for such linking.

A second scenario is when a data custodian, such as a registry, is disclosing data to multiple parties. For example, the registry may disclose a data set A with ethnicity and socioeconomic indicators to a researcher and a data set B with mental health information to another researcher. Both data sets share the same core demographics on the patients. The registry would not release both ethnicity and socioeconomic, as well as mental health data to the same researcher because of the sensitivity of the data and the potential for group harm, but would do so to different researchers. However, the two researchers may collude and link A and B against the wishes of the registry. Before disclosing the data, the registry managers can evaluate the marketer risk to assess the expected number of records that can be correctly matched on the common demographics if the researchers colluded in linking data, and adjust the granularity of core demographics to make such linking unfruitful.

Consider a third scenario where a hospital has a list of all patients who have presented to emergency, D' . This data is then de-identified and sent to a municipal public health unit as D to provide general situational awareness for syndromic surveillance. The data set does not contain any unique identifiers. But a breach occurs at the public health unit and say 10% of the records, U , are exposed to an intruder. The public health unit is compelled by law to notify these patients that their data has been breached. Because D is de-identified, the public health unit would have to re-identify the patients first before notifying them, with the help of the hospital or at its own expense. The more patients that are notified the greater the cost for the public health unit and possibly also increases compensation costs. The simplest thing to do, and the most expensive one, is to work with the hospital to notify all of the patients in D' . However, the public health unit can use U to estimate $\hat{\lambda}$ and determine whether matching the breached subset with the original data D' from the hospital would yield a sufficiently high success rate. If $\hat{\lambda}$ is high then the public health unit would request linking U to D' and only notify the re-identified patients, which would be the most cost effective option that would be compliant with the legal notification requirement. If $\hat{\lambda}$ is low then all patients in D' , whether included in the breached subset or not, would be notified even though 90% of them were not affected by the breach.

As a final scenario, detailed identity information can be useful for committing financial fraud and medical identity theft. However, individual records are not worth much to an intruder. In the underground economy, the rate for the basic demographics of a Canadian has been estimated to be \$50 [44]. Another study determined that full-identities are worth \$1-\$15 [45]. Symantec has published an on-line calculator to determine the worth of an individual record, and it is generally quite low [46]. Furthermore, there is evidence that a market for individual identifiable medical records exists [47, 48]. This kind of identifiable health information can also be monetized through extortion, as demonstrated recently with hackers requesting large ransoms [49, 50]. In one case, where the ransom amount is known, the value per patient's health information is \$1.20 [50]. Given the low value of individual records, a disclosed database would only be worthwhile to such an intruder if a large number of records can be re-identified. If the marketer risk value is small, then there would be less incentive for a financially motivated intruder to attempt re-identification.

4.5 Limitations

The measure of marketer risk assumes exact matching. Exact matching is appropriate if there are no or few errors in the data. However, where there are many data errors an intruder may use probabilistic or distance-based matching techniques instead to obtain higher success rates. In such a case, the marketer risk measure would likely underestimate the proportion of records that would be correctly re-identified.

5. REFERENCES

- [1] C. Safran, M. Bloomrosen, E. Hammond, S. Labkoff, K.-F. S, P. Tang, and D. Detmer, "Toward a national framework for the secondary use of health data: An American Medical

- Informatics Association white paper," *Journal of the American Medical Informatics Association*, vol. 14, pp. 1-9, 2007.
- [2] "Transforming healthcare through secondary use of health data," *PriceWaterhouseCoopers* 2009.
- [3] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," *Journal of the American Medical Informatics Association*, vol. 15, pp. 627-637, September/October 2008.
- [4] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records," *Journal of Official Statistics*, vol. 2, pp. 329-336, 1986.
- [5] K. El Emam, A. Brown, and P. Abdelmalik, "Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk," *Journal of the American Medical Informatics Association*, vol. (accepted), 2008.
- [6] K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, and M. Power, "Evaluating common de-identification heuristics for personal health information," *Journal of Medical Internet Research*, vol. 8, p. e28, 2006.
- [7] K. El Emam, E. Jonker, S. Sams, E. Neri, A. Neisa, T. Gao, and S. Chowdhury, "Pan-Canadian De-Identification Guidelines for Personal Health Information (report prepared for the Office of the Privacy Commissioner of Canada)," 2007.
- [8] Canadian Institutes of Health Research, "CIHR best practices for protecting privacy in health research," Canadian Institutes of Health Research 2005.
- [9] ISO/TS 25237, "Health Informatics: Pseudonymization," 2008.
- [10] K. Benitz and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," *Journal of the American Medical Informatics Association*, 2010.
- [11] P. Kosseim and K. El Emam, "Privacy Interests in Prescription Records, Part 1: Prescriber Privacy," *IEEE Security and Privacy*, vol. 7, pp. 72-76, 2009.
- [12] K. El Emam and P. Kosseim, "Privacy Interests in Prescription Records, Part 2: Patient Privacy," *IEEE Security and Privacy*, vol. 7, pp. 75-78, 2009.
- [13] K. El Emam, F. K. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk, "Evaluating patient re-identification risk from hospital prescription records," *Canadian Journal of Hospital Pharmacy*, vol. 62, pp. 307-319, 2009.
- [14] National Institutes of Health, "Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)". vol. NOT-OD-07-088, 2007.
- [15] G. Loukides, J. Denny, and B. Malin, "Do clinical profiles constitute privacy risks for research participants?," in *Proceedings of the AMIA Symposium*, 2009.
- [16] S. Stolfo, S. Bellovin, S. Herkshop, A. Keromytis, S. Sinclair, and S. Smith, *Insider attack and cyber security: Springer-verlag*, 2008.
- [17] E. Kowalski, D. Cappelli, and A. Moore, "Insider Threat Study: Illicit Cyber Activity in the Information Technology

- and Telecommunications Sector," in Carnegie Mellon University and the United States Secret Service, 2008.
- [18] M. Randazzo, M. Keeney, E. Kowalski, D. Cappelli, and A. Moore, "Insider Threat Study: Illicit Cyber Activity in the Banking and Finance Sector," in Carnegie Mellon University and the United States Secret Service, 2004.
- [19] M. Keeney, E. Kowalski, D. Cappelli, A. Moore, T. Shimeall, and S. Rogers, "Insider Threat Study: Computer System Sabotage in Critical Infrastructure Sectors," in Carnegie Mellon University and the United States Secret Service, 2005.
- [20] E. Kowalski, T. Conway, S. Keverline, M. Williams, D. Cappelli, B. Willke, and A. Moore, "Insider Threat Study: Illicit Cyber Activity in the Government Sector," in Carnegie Mellon University and the United States Secret Service, 2008.
- [21] Kroll Fraud Solutions, "HIMSS Analytics Report: Security of Patient ", 2008.
- [22] CSO Magazine, "2006 e-crime watch survey," US Secret Service, Carnegie Mellon University, and Microsoft Corporation 2006.
- [23] J. Predd, J. Hunker, and C. Bulford, "Insiders behaving badly," *IEEE Security and Privacy*, vol. 6, pp. 66-70, 2008.
- [24] CSI, "14th Annual CSI Computer Crime and Security Survey," 2009.
- [25] Ponemon Institute, "Data loss risks during downsizing," 2009.
- [26] Verizon Business Risk Team, "2009 data breach investigations report," 2009.
- [27] PriceWaterhouseCoopers, "Safeguarding the new currency of business: Findings from the 2008 global state of information security study," 2008.
- [28] K. El Emam and M. King, "The data breach analyzer," 2009, [Available at: <http://www.ehealthinformation.ca/dataloss/>].
- [29] E. Shaw, K. Ruby, and M. Jerrold, "The insider Threat to Information Systems," Department of Defense Security Institute 1998.
- [30] S. Bell, "Alleged LTTE front had voter lists," in *National Post*, 2006.
- [31] S. Bell, "Privacy chief probes how group got voter lists," in *National Post*, 2006.
- [32] C. Freeze and C. Clark, "Voters lists 'most disturbing' items seized in Tamil raids, documents say," in *Globe and Mail*, May 7, 2008.
- [33] J. Domingo-Ferrer and V. Torra, "Disclosure risk assessment in statistical microdata protection via advanced record linkage," *Statistics and Computing*, vol. 13, 2003.
- [34] T. M. Truta, F. Fotouhi, and D. Barth-Jones, "Assessing global disclosure risk in masked microdata," in *Proceedings of the Workshop on Privacy and Electronic Society* (WPES2004), in conjunction with 11th ACM CCS, 2004, pp. 85 – 93.
- [35] R. Benedetti, A. Capobianchi, and L. Franconi, "Individual risk of disclosure using sampling design information," *Istituto nazionale di statistica (Italy)* 2003.
- [36] C. Skinner and N. Shlomo, "Assessing identification risk in survey microdata using log-linear models," *Journal of the American Statistical Association*, vol. 103, pp. 989-1001, 2008.
- [37] J. Bethlehem, W. Keller, and J. Pannekoek, "Disclosure control of microdata," *Journal of the American Statistical Association*, vol. 85, pp. 38-45, 1990.
- [38] Y. Rinott, "On models for statistical disclosure risk estimation," *Joint ECE/Eurostat Working Session on Statistical Data Confidentiality* 2003.
- [39] N. Cressie, A. S. Davis, J. Folks, and G. E. Policello, "The Moment Generating Function and Negative Integer Moments," *American Statistician*, vol. 35, pp. 148-150, 1981.
- [40] P. Haas and L. Stokes, "Estimating the number of classes in a finite population," *Journal of the American Statistical Association*, vol. 93, pp. 1475-1487, 1998.
- [41] H. Howe, A. Lake, and T. Shen, "Method to assess identifiability in electronic data files," *American Journal of Epidemiology*, vol. 165, pp. 597-601, 2007.
- [42] H. Howe, A. Lake, M. Lehnerr, and D. Roney, "Unique record identification on public use files as tested on the 1994-1998 CINA analytic file," *North American Association of Central Cancer Registries* 2002.
- [43] K. El Emam, "Heuristics for de-identifying health data," *IEEE Security and Privacy*, pp. 72-75, July/August 2008.
- [44] S. Polsky, "Witness statement to the Standing Senate Committee on Legal and Constitutional Affairs," *Parliament of Canada* 2007.
- [45] Symantec, "Symantec Global Internet Threat Report - Trends for July-December 07," *Symantec Enterprise Security* 2008.
- [46] Symantec, "What is the underground economy?," 2009, [Available at: <http://www.everyclickmatters.com/victim/assessment-tool.html>].
- [47] N. Luck and J. Burns, "Your secrets for sale," in *Daily Express*, 1994.
- [48] L. Rogers and D. Leppard, "For sale: Your secret medical record for 150," in *Sunday Times*, 1995, p. 2.
- [49] D. Kravets, "Extortion Plot Threatens to Divulge Millions of Patients' Prescriptions," in *Wired*, 2008.
- [50] B. Krebs, "Hackers Break Into Virginia Health Professions Database, Demand Ransom," in *Washington Post*. vol. May 4, 2009.