Proceedings of the 4th WSEAS Int. Conf. on Information Security, Communications and Computers, Tenerife, Spain, December 16-18, 2005 (pp143-146)

Strategies For Matching Between Datasets Using Quasi-identifiers

JANET AISBETT and GREG GIBBON
School of Design, Communication and Information Technology
The University of Newcastle
University Drive, Callaghan NSW 2308
AUSTRALIA

Abstract: - This paper investigates the effectiveness of three strategies for matching records in two datasets using quasi-identifiers, that is, sets of attributes which potentially allow identification of individuals. For simplicity we assume that one of the datasets is a subset of the other. The strategies are: discard all records in the larger set which have non-unique quasi-identifier; discard those records in the smaller set which are not uniquely identified then randomly match if there are any co-occurrences in the larger set; or retain all records and randomly match when there are non-unique identifiers. The optimal strategy depends on the cost of mis-matching records versus the cost of not attempting any match.

Key-Words: - data privacy, data anonymity, quasi-identifiers, re-identification, database linking

1 Introduction

There are many reasons for sharing data about individuals, for their overall individual good (for example, linking medical or academic records) or societal good (for example, using de-identified records to research economic development or population health). A balance is required between preserving privacy and making data available to support a specified usage [2]. The privacy goal of *minimizing* the number of correct record-level matches expected between datasets has the sometimes legitimate converse goal of *maximizing* the number of such matches. Both goals are of practical interest.

The fact that multiple attributes values may allow unique identification of a significant number of members of a population is well recognized. The term *quasi-identifier* has been used to describe a minimal set of attributes that can be joined with external information to re-identify individual records. Computational mechanisms to protect privacy modify such attributes by generalisation or removal of the attribute from publicly-released datasets [1, 3]. One general approach to measuring the likelihood of identification uses the notion of *k*-anonymity [3, 4, 5], where released information is designed to refer to at least *k* distinct individuals, $k > 1$. The factor *k* is estimated by starting at a high (general) level, then progressively detailing information until a base privacy rule is no longer satisfied. Determining the potential for identifying record-level data is a difficult problem in practice, and Sweeney [5] discusses how information released about individuals using attributes additional to the quasi-identifier may allow unforeseen matching.

In this paper, we compare three simple strategies for matching between two datasets sharing a quasi-identifier, when one dataset is a subset of the other. Quasi-identifiers can be composed of many possible sets of attributes, and their expected ability to uniquely identify individuals depends on a probabilistic analysis of the (dependent) distributions of the attributes concerned. We illustrate using the quasi-identifier of date of birth (DoB), sex and locality, in an analysis that neglects the slight effect that leap years have on DoB distributions.

2 Strategies for dealing with multiple matches across datasets

Consider two populations $P_2 \subset P_1$ and two datasets D_1 and D_2 where dataset D_i has exactly one record for each individual in the population P_i. Assume the data in both sets are error-free. Suppose that the datasets have at least a set Q of attributes $A_1, A_2, \ldots A_n$ in common. Note that if A_i can take m_i possible values, then there are $m_1.m_2.m_3.\ldots.m_n$ possible quasi-identifier labels, using which each of the populations P_1 and P_2 can be partitioned into $m_1.m_2.m_3.\ldots.m_n$ cells. Call this the *Q partition*. Many cells in a Q partition may be empty.

Suppose we are trying to identify each individual in the population P_2 represented in D_2 as

an individual in P_1 using Q as quasi-identifier. One of several strategies could be adopted, including:

1. Discard individuals with non-unique quasi-identifiers in P_1. The discarded population will not have same characteristics as the general population, because the distributions of attribute values will not be uniform in practice, and so the sub-populations in the cells in the Q partition of P_1 will vary. Individuals in cells with larger sub-populations are more likely to be discarded. Later results will have to be adjusted to take this into account.
2. Discard individuals with non-unique quasi-identifiers in P_2, and randomly match the remaining individuals to individuals in P_1 with the same quasi-identifier. Again, the discarded population will not have the same characteristics as the general population, so later adjustment will be needed.
3. Randomly match individuals in P_2 to individuals in P_1 with the same quasi- identifier.

3 Comparing the strategies

To analyse the likely success of the three strategies as a function of locality population N, we introduce the probability f that an individual in P_1 with a given quasi-identifier label is in P_2. That is, if $N(a_1, a_2,\ldots a_n)$ is the number of individuals with records taking value a_i on attribute A_i in P_1, then we expect $f.\ N(a_1, a_2,\ldots a_n)$ individuals in P_2 to have those attributes. Here, f is a function of the attributes in Q, because subpopulations may vary with each of these factors.

Let $p_{k,m}(N(a_1, a_2,\ldots a_n))$ be the expected proportion of individuals in P_1 with records taking value a_i on attribute A_i and who are in a cell with $k-1$ other individuals in the Q partition of P_1 and with $m - 1$ others in the Q partition of P_2. Then

$$p_{k,m}(N) = \binom{k}{m} f^m (1-f)^{k-m} p_k(N). \qquad (1)$$

where $p_k(N)$ is the probability that there are exactly k individuals in a subpopulation N with non-unique quasi-identifier, and where $\binom{k}{j} = \frac{k!}{j!(k-j)!}$ and $q \geq 1$.

Under strategy 1, there are no incorrect matches. For each cell in the Q partition of P_1 of size N an expected $\sum_{k=2..N} \sum_{1 \leq m \leq k} p_{k,m}(N).f.\ N$ individuals will be discarded from P_2. Over all of P_2, an expected

$$\sum_{a1,\ a2,\ldots an} \sum_{k=2..N} \sum_{1 \leq m \leq k} p_{k,m}(N(a_1, a_2,\ldots a_n)).f.\ N(a_1, a_2,\ldots a_n)) \qquad (2)$$

individuals will be removed from the matching process.

Under strategy 2, an individual in P_2 will be correctly matched in P_1 unless he or she shares a quasi-identifier label with one or more others with the same quasi-identifier label, and none of the others is also in P_2. (For if more than one appears in P_2, they will have been discarded.) If the individual has not been discarded and a match is made at random in P_1 from a choice of k with the same quasi-identifier label then the probability of a correct match is $1/k$. Hence the expected proportion of mismatches of individuals in that cell in the Q partition of P_2 is

$$\sum_{k=2..N} (1-1/k)\ p_{k,1}(N).f.N. \qquad (3)$$

Over the entire population, the expected number of individuals in P_2 for which incorrect matches are made is:

$$\sum_{a1,\ a2,\ldots an} \sum_{k=2\ \ldots} (1-1/k)\ p_{k,1}(N(a_1, a_2,\ldots a_n))\ f.N(a_1, a_2,\ldots a_n). \qquad (4)$$

Under strategy 2, on average $\sum_{k=2..N} \sum_{2 \leq m \leq k} p_{k,m}(N(a_1, a_2,\ldots a_n))f\ N(a_1, a_2,\ldots a_n)$ individuals in P_2 will be discarded in each cell. The reward for risking mismatches is that fewer individuals have been discarded.

Under strategy 3, no individuals are discarded, but the probability of a mismatch is increased. The problem is well known as "the matching problem", usually studied when the numbers in the match set P_2 and target set P_1 are the same. Suppose m individuals have pairs in a set of k individuals, where $k \geq m$. There are $k!/(k-m)!$ possible ways of matching the individuals, but averaged over all these ways, the expected number of correct matches is always less than 1 unless $m = k$, when it is exactly 1.

Figure 1 shows the expected number of correct matches as a function of m (shown on the x-axis) for various k. The expected number of mismatches, is always between $m-1$ and m, whatever the k. Let $e(k,m)$ be the expected proportion of mismatches, which is therefore between $1-1/m$ and 1. Then the expected number of mismatches in a cell is

$$\sum_{1 < k \leq N} \sum_{1 \leq m \leq k} e(k, m)\ p_{k,m}\ (N(a_1, a_2,\ldots a_n)f\ N(a_1, a_2,\ldots a_n). \qquad (5)$$

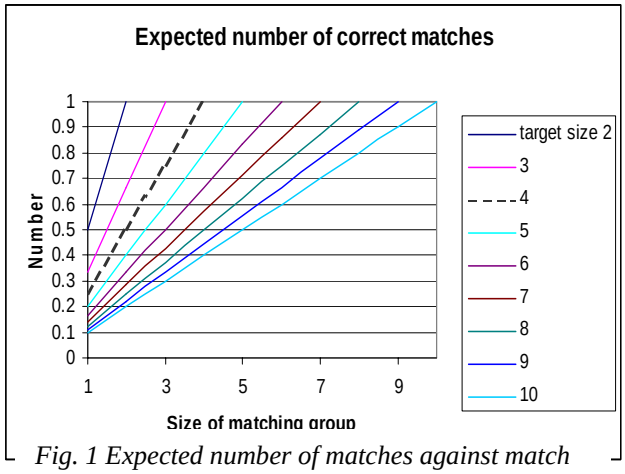
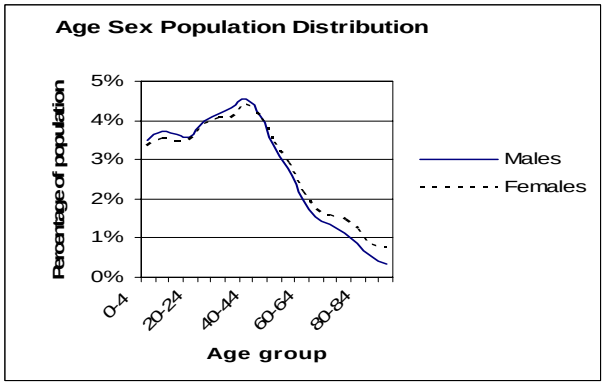


Fig. 1 Expected number of matches against match group size, for various target group sizes



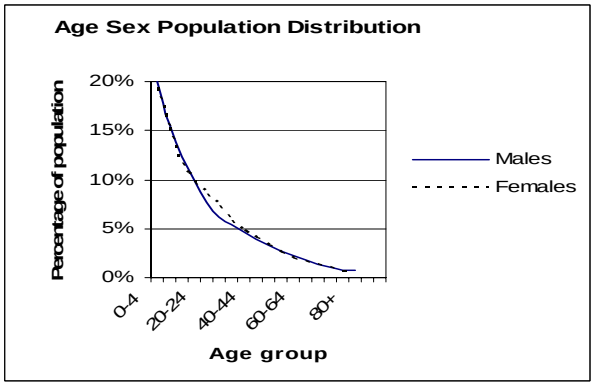


Fig. 2 Age sex distribution: Top (a) Orange County US Census 2000; Bottom (b) Tanzania 1980

4 Matching on quasi-identifier set DoB, age and locality

Distributions of age and even gender vary with country and locality within that country. Fig. 2a illustrates a typical shape for OECD countries, with a bulge caused by mid-20th century births, longevity and relatively low birth rates. In contrast, distributions in developing countries tend to be J-curves with high birth rates and high death rates in the early years (Fig. 2b) although urban and rural distributions differ markedly.

Figure 3 shows the expected proportion of mismatches and discards for the three strategies for the case when the ratio f between cells in the Q partitions of P_1 and P_2 is a constant and an OECD-like distribution is used. Fig. 3a refers to the situation when the population P_2 is expected to be a quarter that of P_1, Fig. 3b to the situation when P_2 is expected to be a half that of P_1 and Fig. 3c to the situation when P_2 is expected to be three quarters that of P_1 .

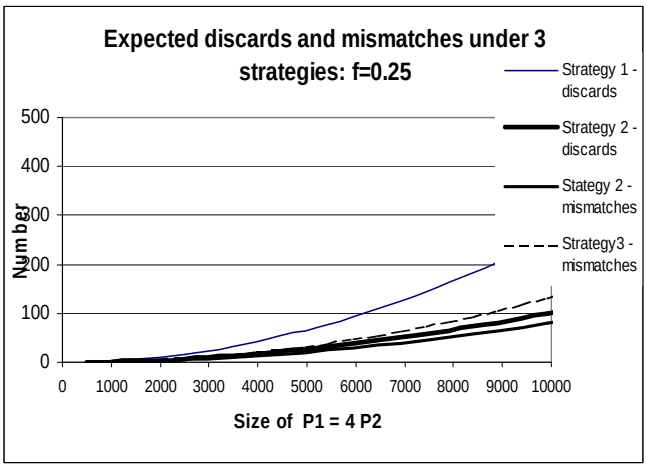
When the total population is 10,000 and f=.5 (so P_2 is 5000) the safe strategy 1 of discarding records with multiplicities results in an expected 16% of discards of the 5000 in P_2. With strategy 3 an expected 8% will be mismatched, only half the discard rate. When the total population is 5000 and P_2 is 2500 there are an expected 8% discards with strategy 1 and 4% mismatches with strategy 3. When the total population is 4000, these figures drop to 7% and about 3.5%.

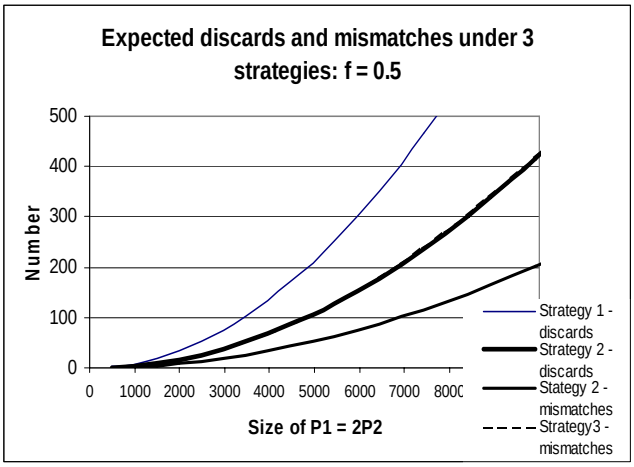
The distribution of a sub population is unlikely to be the same as that of the broader population, so the assumption of a constant ratio f is unrealistic. However, it is obvious from its derivation that Strategy 1 always results in the largest number of discards+mismatches over the 3 strategies, and Strategy 3 in the lowest. The optimal strategy will depend on the cost of mismatching, as well as on the relationship between the sizes of the populations.

5 Conclusion

Depending on the task involved, incorrect matches may be tolerated if the overall number of correct matches is increased. If this is so, then random matching of individuals sharing the same attribute values may be permitted. We presented two strategies based on random matching and compared them with the error-free strategy of discarding all individuals who did not have unique attribute value sets.

Computing expected correct match rates in datasets when matching on a quasi-identifier depends on understanding the joint distribution of attribute values in the quasi-identifier. A quasi-identifier of practical interest, and for which some relevant distribution data are available, is age/sex/locality. Our analysis revealed quite different performance from the three strategies, as well as the dependence on the relative sizes of the two datasets.





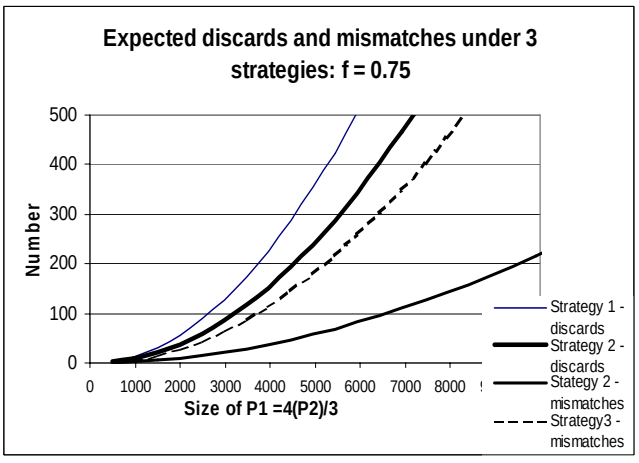


Fig.3. Expected number of discarded individuals and mismatched individuals when matching between 2 populations on birthdate, sex and locality. The total population is that in the larger set. A generic metropolitan age-sex distribution is assumed. See text for description of strategies.

The best strategy for increasing the matching rate appears to be to remove any non-uniquely identified records prior to matching and to compensate in later analyses for the discarded records on the basis of demographics.

Only random matching was considered. Of course, the underlying problem in privacy is that information in attributes other than those considered in the quasi-identifier may help identify an individual. It is possible that attributes in the datasets, even when not shared between them, could be used to do better than random matching. For example, the small dataset might be known to be collected from an Internet hip hop music site, and the larger dataset has an attribute describing musical preference; or the smaller dataset is derived from patients in an emergency department and the larger has attributes describing medical conditions.

The analysis presented here was based on assumptions that are unrealistic, namely that data are error free, that one dataset has records pertaining to each of the individuals represented in the other dataset, that an individual's DoB/sex/locality identifier is fixed between datasets, and that individuals only have one record in each of the two datasets. Further work should allow for more realistic data conditions, as well as exploring the effects of using collateral information.

References:

[1] Fung, B. C. M., Wang, K., and Yu, P.S.. Top-Down Specialization for Information and Privacy Preservation, *21st International Conference on Data Engineering* (ICDE'05), 2005, pp. 205-216.

[2] Iyengar, V. S. Intrusion and privacy: Transforming data to satisfy privacy constraints, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, July 2002, pp.279-288.

[3] LeFevre, K., DeWitt, D. J and Ramakrishnan, R. Incognito: Efficient Full-Domain K-Anonymity. *SIGMOD* Baltimore 2005 pp 49-60.

[4] Samarati, P. Protecting Respondents' Identities in Microdata Release, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No. 6, 2001, pp. 1010-1027.

[5] Sweeney, L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol 10 No. 5, 2002, pp. 557-570.