Proceedings of the 4th WSEAS Int. Conf. on Information Security, Communications and Computers, Tenerife, Spain, December 16-18, 2005 (pp158-162)

Matching Records on Birthdate, Sex And Locality

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

Abstract:- This paper investigates error rates when matching between datasets on date of birth, sex and locality. Using US health insurance data and voter registration rolls, Sweeney demonstrated how "sanitised" data retaining this information can be matched with records containing uniquely identifying information. Individuals are uniquely identified by year of birth, sex and locality unless they share a birthday with someone with the same attributes. The distribution of multiple occurrences of birthdays is therefore investigated for various size populations, using Australian data to illustrate. Assuming typical age/sex distributions for localities in which no more than 1% of the population is in any cell, a locality population of 4000 will enable an expected 95% of the population to be identified. A locality population of half a million is needed to reduce the percentage of matches to 0.1%.

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

1 Introduction

Benefits to individuals and to society can result when potentially sensitive record-level data is shared between organisations. Such data, sometimes termed microdata, can for example aid research into public health, or help administrators plan future critical infrastructure.

In an effort to satisfy privacy considerations, record level data may be "sanitized" by removing names but, for operational reasons, leaving demographic information, in particular date of birth (DoB), sex, and locality, which are non-identifying in large populations. However, other datasets may provide identifying information along with this same basic demographic data. Such datasets might be generated to store less sensitive information -- for example, as mailing lists – but when linked to databases containing information not intended for disclosure, can support re-identification and lead to breaches of privacy. In a compelling example, Sweeney demonstrated matching across datasets on DoB, sex and locality using United States health insurance and voter registration records [5].

Investigations into sets of attributes which form quasi-identifiers and allow linking of datasets form an active area in privacy research (eg, [1, 2, 4]) Using United States Census data, Sweeney showed that nearly 87 per cent of individuals can be expected to be uniquely identified by their DoB, sex and zipcode; 50 per cent by DoB, sex, and the city, town, or municipality in which the individual lives; and 18 per cent by DoB, sex and county [6]. Linking is interesting not only for its ramifications for privacy, but also from the perspective of those legitimately attempting to match across datasets for which uniquely identifying information is either not available or is corrupted in one or both datasets.

Sweeney's results concern very variable populations; for example, US zip code populations vary from hundreds to many tens of thousands. This paper therefore looks more closely at the role of the number of inhabitants in a locality in determining the expected number of correct matches that can be made between datasets on the basis of DoB, sex and locality.

It is assumed here that the two datasets contain one record per individual. It is also assumed that

(a) attribute data in the datasets D_i accurately represents a population P_i ($i = 1, 2$) and

(b) $P_2 \subseteq P_1$ where the dataset D_1 is the "master" set.

It is not assumed that either set necessarily contains identifying information, only that both contain the demographic data.

Accurate matching is only possible when DoB, sex and locality refer to a unique individual amongst the population recorded in either of the datasets. If multiple individuals share the same identifiers, then matching performance becomes probabilistic. Clearly, matching performance depends on the size of the overall populations P_i. So we need to compute the probability that an individual has the same DoB

and sex as exactly *q-1* other individuals in a locality with population *N*. We will use this to describe how the locality population affects matching across datasets.

The next section looks at co-occurrence of birthdays as a function of group size. Section 3 introduces age-sex distributions in localities to convert these results to findings about co-occurrence of DoB/sex/locality identifiers, illustrating on data drawn from Australian localities.

2 The Birthday Problem

Assuming equal probability of birthdays across the year, and ignoring leap years, it is easy to see that the probability that an individual has unique birthday amongst a set of *N* people is $(364/365)^{N-1}$. The probability p_q that there are exactly *q-1* other individuals with that birthday is

$$p_q(N) = \binom{N-1}{q-1} \left(\frac{1}{365}\right)^{q-1} \cdot \left(\frac{364}{365}\right)^{N-q} \qquad (1)$$

where $\binom{k}{j} = \frac{k!}{j!(k-j)!}$ and $q \geq 1$.

Figure 1 depicts p_q as a function of population size *N*, for $q = 1 \ldots 4$. That is, Fig. 1 shows how increasing population size *N* affects the probability of an individual having a unique birthday, or equivalently, having a unique DoB/age/sex identifier in a locality in which *N* people are of their age and sex. A population of 250 gives a 50% chance of being uniquely identified. A population of around 2500 is needed to reduce the number of unique birthdays to 0.1%.

The well known birthday matching problem [3] explores the more general probability of a given number *m* of unique birthdays occurring amongst *N* individuals. The solution is based on the probability *p(N, k, j)* that when *N* things are distributed randomly into *k* baskets, then exactly *j* of the baskets will be empty. The probability is derived from a combinatorial argument [1].

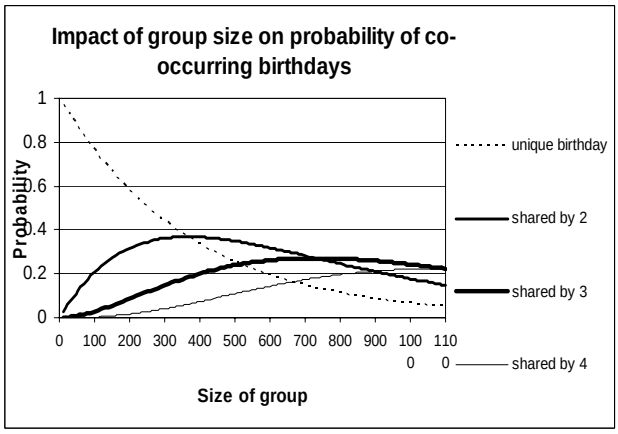

Fig. 1. Predicted probability that an individual shares a birthday for none, one, two or three other individuals, as a function of group size.

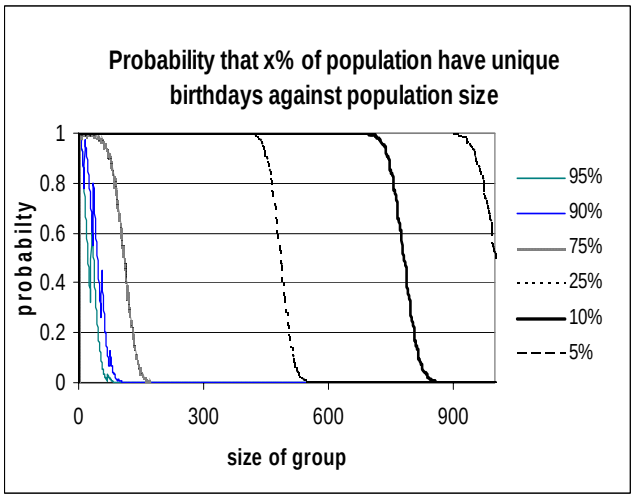

Fig. 2. Probability that a given proportion of individuals have a unique birthday, as a function of group size

Introducing seasonal variation in the distribution of birthdays has no appreciable effect on solutions to the birthday problem using American data [3]. There, the distribution of birthdays varies by no more than 5-7% throughout the year. Although the reversal of the seasons and the coincidence of the summer vacation with Christmas might alter the distribution for Southern Hemisphere countries, adjustment for non-uniform distributions has not been judged to be warranted in this analysis.

Assuming equal probability of birthdays across the year and ignoring leap years, the probability that *N* people have between them exactly *j* birthdays is *p(N, 365, 365-j)*. Here, by necessity, *j* is at least *1* and is no greater than the minimum of *N* and *365*.

[1] The formulation of *p(N, k, j)* is :

$$\binom{k}{j} \sum_{m=0 \text{ to } k-j} \binom{k-j}{m} (-1)^m \left(\frac{k-j-m}{k}\right)^N$$

Take out at random one person for each of the j birthdays that the group shares. Then anyone who has a unique birthday must be in this set of j individuals. The probability that exactly m of the j birthdays is unique is the probability that there are exactly m empty "baskets" after randomly distributing the remaining N-j birthdays into the j birthday "baskets". This is $p(N\text{-}j, j, m)$. The probability $\rho(N, m)$ of m unique birthdays amongst a group of N individuals is therefore

$$\rho(N, m) = \sum_{j=m \text{ to } \min(N, 365)} p(N, 365, 365\text{-}j)\, p(N\text{-}j, j, m) \quad (2)$$

This is the probability that m/N of the population has a unique birthday. The expected proportion of the population whose birthdays are unique in that population, $p_1(N)$, is $\sum_{m=0 \,..\, N,} \rho(N, m)m/N$. Fig. 2 depicts the probability of various percentages of the individuals in a group having a unique birthday, each as a function of group size. The jagged form of some of these plots is due to the fact that the percentages have to be converted to integral numbers of unique birthdays[2].

The maximum group sizes at which a given proportion of the population has unique birthdays, to a given confidence level, is listed in Table 1.

Table 1. Maximum population size for a given proportion of the population to have unique birthdays to a given probability level.

Proportion/ Confidence level	*5%*	*95%*
0.8 Confidence	972	14
0.9 Confidence	952	11
0.95 Confidence	940	6

3 Allowing for age-sex distribution

The age-sex distribution of different localities is needed to translate the birthday problem findings into the predictions about co-occurrence of DoB, sex and locality that are required for matching between the data sets D_i. These distributions vary between countries, and within countries they may vary markedly between rural and urban localities.

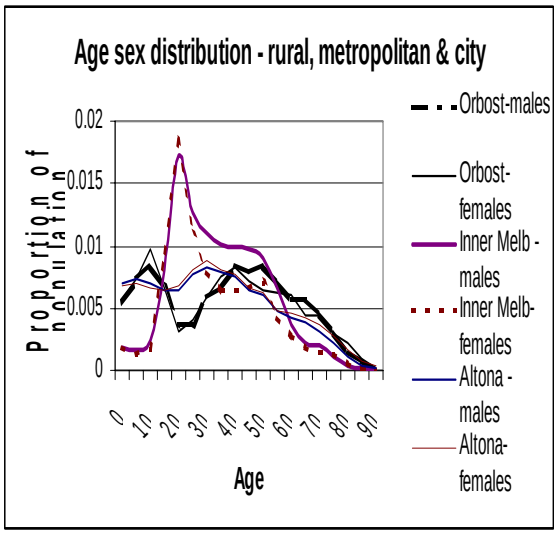

Fig. 3. Sample age-sex distributions for different types of localities. Smoothed ABS Statistical Local Area data 2001 census for rural (Orbost), metropolitan (Hobson's Bay-Altona) and city (Inner Melbourne)

As in many countries, in Australia there are relatively more children and older people in rural areas. Fig. 3 depicts typical Australian age distributions $d_{s,t}$ *(age)* of subpopulations for sex s and locality type t for inner city (inner Melbourne), metropolitan (Altona) and rural (Orbost) localities. Here, $d_{s,t}$ *(age)* = $N(age, s, L)/ N$ for the subpopulation of sex s in the (metropolitan or rural type) locality L with total population N.

From this data, less than 1% of a locality's population has the same age and sex except in inner city areas. In general, therefore, if a locality has 20,000 people, say, then at most 200 will be in the same age-sex cell and so from Fig. 1 the probability that a person is uniquely identified by birthday, age and sex is more than 0.6. If a locality has 8000 people, then that probability rises to 0.8; or for 3000 people, over 0.9.

A more accurate estimate of the expected proportion of unique DoB/sex/locality identifiers as a function of locality population N, for rural or metropolitan locality type t, can be obtained by summing over each age-sex cell, viz,

$$\sum_{A=0 \text{ to } 110} p_1(d_{males,t}(A)*N) + p_1(d_{females,t}(A)*N). \quad (3)$$

[2] For example, the graph for 95% unique has a local minimum at $N=30$, because this is the smallest size at which $.95\ N$ rounds down to $N\text{-}2$ and hence the percentage includes the possibility of one pair of coincident birthdays. (Note that there is always zero probability of exactly $N\text{-}1$ unique birthdays out of a group of N.)

Here, $d_{s,t}$ (A) * N is the expected number of individuals of sex s and age A in a locality of type t and population N so that $p_1(d_{males,t}$ $(A)*N)$ is the expected proportion of males with unique birthdays, and hence with unique birthdate/sex/locality identifier.

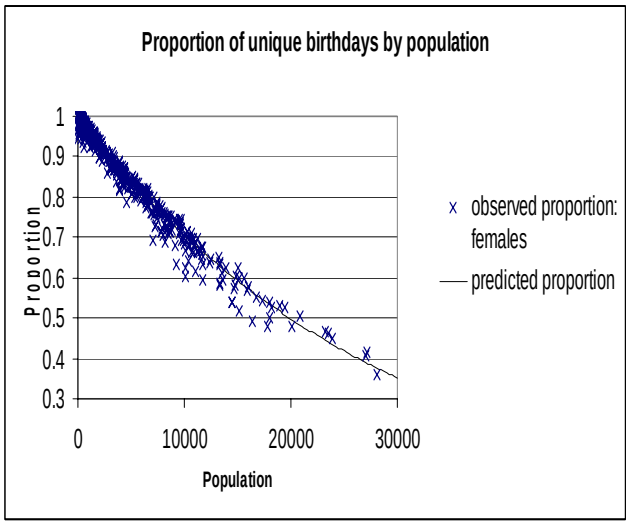

Fig. 4. Proportion of females with unique DoB by population and predicted proportion

The difference between rural and metropolitan age distributions has very little effect on the expected proportion of unique DoB in the overall population of a locality, although it does impact on the age groups in which those birthdates are most likely to occur. So Fig. 4 depicts the result of applying equation (3) to get the predicted proportion against population size for either metropolitan or rural localities. Here the localities are postcode areas, a common way of identifying locality.

Also shown in Fig. 4 are actual percentages of female individuals for whom DoB is a unique identifier within a locality, as a function of locality population. Even in the largest postcode areas have over one third of people uniquely identified by the demographic information. Outliers for which the proportion of unique identifiers is significantly below the prediction may be identified as localities whose population tends to congregate within some age groups relative to the generic distributions.

Fig. 5 shows the actual proportions of females who share their DoB with others in their postcode, again as a function of population of postcode. This figure shows that the actuals accord with the results suggested by Fig. 1. Again, the results for males are similar.

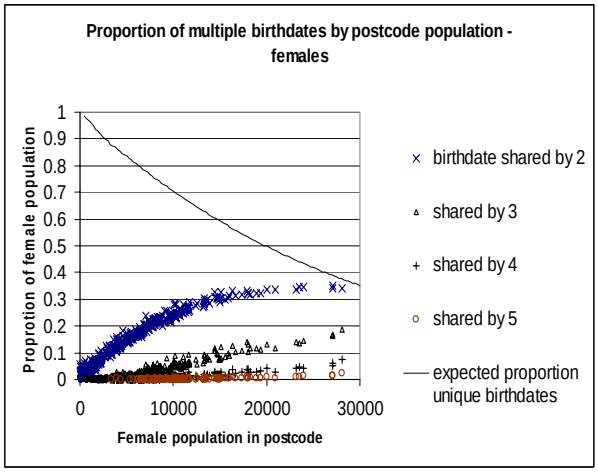

Fig. 5. Proportion of females in postcodes who share DoB identifiers. Also shown is predicted proportion of females with unique DoB, which, from Figure 4, is a good estimate of actuals.

If sex is taken into account, these figures indicate that 95% of the individuals in a locality are expected to be uniquely identified by DoB and sex when the locality has a population of about four thousand. If certain age groups are of interest – say, those over 60 -- then the number differs. In the case of over-60s, a local population of eight to ten thousand may still give an expected 95% unique match, provided the local distribution accords with the generic distribution. As noted, Sweeney's analysis of US census data by zip code concerned localities varying from hundreds to tens of thousands.

4 Conclusion

This article assumed that data are error free, that an individual's DoB/sex/locality identifier is fixed, and that individuals are uniquely identified in each of the two datasets. Our analysis indicated that an age/sex cell population of around 20 is then required for 95% of individuals in a given locality to be uniquely identified by DoB and sex, and a population of around 1100 to reduce this to 5%. This translates roughly into a locality population of 4000 for 95% identification, and a population of 22,000 for 5% identification. In an actual Australian dataset, around 95% of individuals had a unique birthdate at population sizes of 2000, or 4000 if sex is included. More detailed modelling involving likely population distributions showed that if half of the total population was in P_2 then 7-8% of these individuals would not have unique DoB/sex identifiers in a

locality size of 4000, and so would be discarded under a matching strategy that did not tolerate any errors.

Reducing the expected number of uniquely identified individuals to less than 0.1% requires localities with populations of half a million or more.

Analysis of the likelihood of matching across datasets on a set of attribute values requires a detailed knowledge of the underlying multivariate distributions, as this study has indicated. Such distributions may be available for attribute sets such as DoB, sex and locality, but because the form of the distributions can vary greatly, care must be taken in generalising results

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

[3] Peterson, I "Birthday Surprises" *Mathtrek MAA Online*. The Mathematical Association of America. Nov. 22, 1998

[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, 2002,557-570.

[6] Sweeney, L. Comments To the Department of Health and Human Services On "Standards of Privacy of Individually Identifiable Health Information" 26 April 2002. http://privacy.cs.cmu.edu/dataprivacy/HIPAA/HIPAAcomments.html