Online Supplement to
Privacy Protection of Microdata with Individual Identifiers

Practical Illustrations of the Proposed Technique

In the paper "Privacy Protection of Microdata with Individual Identifiers" we develop a technique for the following scenario:

> An organization has the right and need to collect and view exact individually identifiable information. It wishes to share data with another organization that has a legitimate use for the data, but does not have the right or need to access exact information.

In the online supplement we provide some illustrative examples of settings where our technique can be implemented effectively. Two of the examples serve to illustrate the public health benefits that can be realized. A third example illustrates how our technique can be used to support inter-organizational sharing of IIM (individually identifiable microdata) along with association rules mined from customer-specific transaction data.

Clinical Trials

One example of how the technique developed in this work could be successfully applied in the domain of public health and research is the identification of clinical trial study subjects. In this context a substantial benefit accrues to the data user because identifying appropriate trial subjects can be time consuming and costly. Presently, subjects are obtained either by physician referral, or by "advertising" on public websites such as Center Watch (www.centerwatch.com). One example, for a study seeking women at high risk for breast cancer, is available at

http://www.centerwatch.com/patient/studies/stu56452.html. On the other hand, some studies deliberately seek low-risk individuals, which can also be challenging. One example is the study seeking men over the age of 55 that have never been diagnosed with cancer available at http://www.centerwatch.com/patient/studies/stu38876.html.

Absent a more efficient method of identifying subjects, both of these research efforts will suffer from delays and high screening costs. Delays are associated with the amount of time needed to identify the appropriate number of subjects. While studies can begin as soon as the first suitable subject is enrolled, they cannot end until the required number of subjects has been evaluated. Screening costs are incurred because it is often not enough that subjects verbally attest to the fact that they qualify for a medical study. Medical records must be screened in order to verify that each subject is an accurate historian. This is especially true when subjects receive financial compensation.

Payment for participation is so common that there is a website, http://www.biotrax.com/home.php, dedicated to listing for-pay clinical research projects. Using our technique, clinical trial directors would be able to obtain statistical information that would enable them to identify groups that exhibit certain risk characteristics (e.g. at high risk for breast cancer or low risk for cancer). They could then query the recoded database and obtain a list of subjects such that all subjects meeting the search criteria specified in the query are included in the list, along with some spurious subjects. They could then contact these subjects about the possibility of enrolling in the research study. Note that the study director, who does not have the right or the need to view exact medical information, has no way of knowing which subjects truly meet the search

criteria. Furthermore, there is no need for the initial invitation to include specific details about the relative risk for, say breast cancer.

The benefits are two-fold. First, the study directors obtain contact information for all subjects meeting their search criteria. This eliminates the delay associated with the current mechanism. Second, the cost of screening the subjects is greatly reduced (it is not eliminated since not all of these subjects would truly meet the search criteria. So prior to enrollment some basic screening would be required). Data subjects that truly meet the search criteria have an opportunity to participate in a research study that could be beneficial to their long-term health. Those that do not will be screened out as potential study candidates, or perhaps included as a control group depending on study design.

Health Newsletters

The ability to generate customized health newsletters is another important public health application of our technique. For instance, consider the problem faced by a health insurer or a medical center that wishes to disseminate useful health-related information to individuals. Two examples of this are the health and wellness information provided by Aetna health insurance (see http://www.aetna.com/members/health_wellness/health_wellness.html) and the myGeorgetownMD newsletter provided by the Georgetown University Hospital (see http://www.georgetownuniversityhospital.org/body.cfm?id=1296).

Confidential medical information is readily available to both Aetna and Georgetown University Hospital. Aetna obtains this information in the context of processing medical claims. Georgetown University Hospital obtains this information

during the process of providing health care services. However, that confidential information can only be accessed by individuals with the right and need to know. As a result, the health newsletters currently disseminated by both of these organizations are not customized in any way, *even though subjects opt-in to receive them*.

All subjects receive the same generic material and the issue of what, if any, of that information is relevant to a given subject is not considered. The result is that subjects receive a seemingly random sample of information. For example, the most recent issue of myGeorgetownMD newsletter (included in our response as an attachment) covers topics ranging from drug-resistant epilepsy to inflammatory bowel disease.

Given the vast amount of health information that could be included in such a newsletter, the contents represent a very small proportion of what could be included. The obvious problem with including everything is information overload. While theoretically everyone could benefit from receiving all available information, it is more practical to provide individuals with information that is relevant to them.

Our technique provides the opportunity to use the available medical information to provide more specific targeting of individuals. At the same time, the information that is shared (even though it is not leaving the organization) has been recoded such that data recipients (those charged with distributing customized newsletters) are not able to see, or infer, confidential medical information. The benefit to the data subject is access to current, relevant medical information while minimizing information overload. As with the existing newsletter, the subjects are not being told that they have, or are at high-risk for, a particular ailment. The purpose is to ensure that the information they are presented with is relevant. The benefit to the data users (the hospital or the insurer) is the more

efficient utilization of resources. In both cases these organizations expend financial resources in order to educate people. Our recoding scheme allows those financial resources to be deployed more efficiently.

Data mining

Increasingly businesses are looking to share knowledge gleaned from customer transaction data. In many cases the data mining takes the form of association rule mining. The general form of an association rule is {if A, then B} where A and B are disjoint sets of items. The set A is typically termed the antecedent and the set B is termed the consequent. The association rule is accompanied by two numbers that measure the degree of uncertainty of the rule. The first is the *support* which is the probability that a transaction drawn at random will contain all items in A. The second is the *confidence* which is the conditional probability that a selected transaction will include B, given that it includes A. For example in a database with 1,000 transactions, 30 of which include A and 6 of which include A and B then support = 30/1000 = .03 and confidence = 6/30 = .2. Typically the mining of association rules is driven by user-specified levels of support and confidence where the number of rules extracted is limited to those with high levels of support and confidence.

Association rules are useful because they enable a number of marketing initiatives ranging from promotional efforts to store merchandizing. In this general context our technique can be applied to a data set that includes individual customer information as well as the association rules that have been mined. Consider the following database:

Name	ZIP	Gender	Occupation	Preference Category 1	Preference Category 2	Income	Association Rules
M.A.	06040	M	Banker	Low	High	102	1
G.P.	06269	M	Manager	Medium	Medium	78	2
M.L.	14260	F	Nurse	High	Low	49	3
W.F.	14260	M	Banker	Low	High	121	1
R.H.	06040	F	Banker	Low	Medium	97	1
F.J.	06269	M	Manager	High	Medium	80	2
M.G.	98195	F	Nurse	High	Low	29	2
J.M.	98195	F	Nurse	Medium	Medium	61	3
A.B.	98195	F	Banker	Medium	High	96	1
J.R.	14260	M	Manager	Low	Low	48	3
R.S.	98195	M	Nurse	Medium	Medium	59	1

Table 1: Customer data and association rules

The Table 1 depicted here is analogous to the Table 1 used to begin our public health illustration in the paper. Here the database includes individual identifiers for each customer (name, zip code, and gender). The database also includes information on four intermediate fields (occupation, salary, and the customer's preferences for products in category 1 and 2). This intermediate data is typical of the type of data collected when, for example, a customer registers as a new user on a website. The database also includes a confidential field that shows the specific association rules that each customer has exhibited based on past transactions. Here we do not expect that the number of association rules will be excessively large since most data mining initiatives are designed to extract a relatively small set of the most powerful associations (in terms of support and confidence).

The owner of this data would like to share (or perhaps sell) the information in this database but would also like to prevent the data recipient from inferring the value of the confidential field. By applying our technique (which could be applied seamlessly) the owner of the data could provide the data recipient with the association rules mined from the transaction data as well as a perturbed microdata set with individual identifiers intact.

The data recipient could then use the perturbed microdata set to generate marketing campaigns that are more precisely targeted to customer behavior. At the same time the data recipient would be unable to infer the exact association rules exhibited by a given customer.