Microdata Disclosure Limitation in Statistical Databases: Query Size and Random Sample Query Control

George T. Duncan
School of Urban and Public Affairs
Carnegie Mellon University
Pittsburgh, PA 15213

Sumitra Mukherjee
School of Urban and Public Affairs
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

A probabilistic framework can be employed to assess the risk of disclosure of confidential information in statistical databases that use disclosure control mechanisms. We illustrate how the method may be used to assess the strengths and weaknesses of two existing disclosure control mechanisms - query set size restriction control and random sample query control mechanisms. Our results indicate that neither scheme provides adequate security. The framework is then further exploited to analyze an alternative scheme combining query set size restriction and random sample query control. We show that this combination results in a significant decrease in the risk of disclosure.

1 Introduction

Microdata—data on individual respondents—is increasingly maintained in computer databases. Providing researchers with access to such microdata can further research in the social, policy, and management sciences [13]. While technically now easier, this information sharing may run counter to confidentiality protection for respondents; both are major and conflicting themes guiding data collection and dissemination [17]. The need for devising methods to prevent the disclosure of confidential information while providing access to microdata has been emphasized by several researchers [4,5,6,7,10,16,19,24]. A survey of security control methods for statistical databases is found in [1].

Much of the existing literature relates to the risk of disclosure of confidential information for data released to the user as a complete file. With the rapid advance in communication technology, however, remote access to computer databases in the form of sequential queries is becoming increasingly common.

The need for studies addressing the confidentiality issue in sequentially accessed databases has been highlighted in recent years [2,15,20,22,25].

The disclosure problem concerns the possibility of identifying individuals in a database and revealing sensitive information [3]. If the database includes sufficiently rich information about individuals, even if explicit identifiers such as name and social security number have been deleted, it may contain some identifiable respondent. Fellegi [16] considers a case where the database contains information about the sales figures of firms. Since the broad classification and the identity of at least the largest firms are often common knowledge, particular care must be taken to prevent the identifiability of the sales figures. A common policy is to withhold, not only information on individual respondents, but also aggregate information based on fewer than three respondents on the assumption that any two respondents of a particular kind may easily know each other. Hence, if a statistic based on two respondents was published, any one of the two could subtract his own report from the aggregate figure and hence deduce the figure reported by the other. Thus, the existence of only a few individuals with some unique combinations of characteristics may significantly increase the risk of disclosure.

A commonly used method for limiting this form of disclosure in sequentially accessed databases is the query size restriction (QSR) approach. The query programs of certain statistical databases report raw statistics for query sets. Query sets are groups of records which satisfy the characteristics laid down by the questioner. The raw statistics include query set size and the sums of powers of values in the query set. Many users and database designers have apparently believed that individual records will remain confidential as long as query programs refuse to
report the statistics of query sets which have either too few or too many elements [8]. It has, however, been demonstrated that if there are enough distinguishable classes of individuals in the database, compromise of small or large query sets can in fact always be accomplished with the help of sequences of queries [26]. This notion has been formalized to suggest tools called trackers which lead to disclosure. The general tracker introduced by [8] permits the calculation of statistics for arbitrarily small query sets, without requiring knowledge about individual records. Their results demonstrate that employing a tracker is generally feasible, and this poses the greatest threat to security under QSR control.

An alternative approach to disclosure control is the use of random sample query (RSQ) control. In this technique, a probabilistic sample is constructed from the set of records corresponding to a given query and the answer to the query just for the probabilistic sample is presented to the database user. RSQ control deals directly with the basic principle of compromise by making it impossible for a questioner to control precisely the formation of query sets. Thus trackers cannot be employed. The sampling control strategy is said to permit the release of accurate and timely statistics and is implementable at low cost [9].

In this paper we develop a probabilistic framework to assess the risk of disclosure of confidential information in online statistical databases under various inference control mechanisms. We show that neither QSR control nor RSQ control offer adequate protection against disclosure. Next we suggest a control strategy by combining QSR with RSQ and show how this results in a significant decrease in the risk of disclosure. The database user attempting the disclosure is termed a snooper. The snooper is modeled as a Bayesian decision maker who has subjective probabilities that individual respondents belong to categories into which the database has been partitioned. These probabilities are successively revised as more information is available to the snooper in the form of responses to queries. Duncan and Lambert's disclosure limiting (DL) framework [12,13] suggests that the risk of disclosure depends on the perceived cost of identifying a record, the perceived probability of success, and the value of the information to the snooper. Given the loss function function of the snooper, the risk of disclosure would hence depend on the snooper’s posterior probabilities.

In Section 2 we present the model adopted for our analysis and derive expressions for the snooper's subjective probabilities that a particular respondent is in a certain category, based on the information obtained from queries. Results are established which illustrate how information about one particular category may affect the probabilities associated with the other categories. In Section 3 we consider the use of query size restriction control. Based on our probabilistic framework we show that this control mechanism does not offer adequate protection against disclosure. In Section 4 the effect of RSQ control on these subjective probabilities are obtained. The results are used to show that the risk of disclosure may be unacceptably high even under RSQ control. In Section 5 we investigate the risk of disclosure when QSR and RSQ are both employed. Our analysis indicates that there is a significant reduction of the risk. Section 6 concludes with discussions on the contributions of this research and possible extensions of this work.

2 The Model

The database is considered to be partitioned into K categories; N respondents are distributed among these K categories. Queries are restricted to questions about the number of respondents in each category or a union of categories. Note that typically this is not a restrictive assumption since all statistics such as the mean and regression coefficients can be approximated by a sequence of categorical queries to the extent permitted by the refinement of categories.

Table I. Distribution of respondents to an AIDS survey.

<table>
<thead>
<tr>
<th>HIV Status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>High Risk Group</td>
<td>1 9 10</td>
</tr>
<tr>
<td>Low Risk Group</td>
<td>2 88 90</td>
</tr>
<tr>
<td>Total</td>
<td>3 97 100</td>
</tr>
</tbody>
</table>

As an example of such a database, consider hypothetically, N = 100 respondents distributed among...
K = 4 categories based on the two attributes "Risk Group" and "HIV Status", each with two possible values. Table I shows the number of respondents in each category. For instance, there is only 1 respondent who belongs to the High Risk Group and has a positive HIV Status and a query about this category would return a response 1. Dissemination of this information would impose unacceptable disclosure risks.

The snooper is assumed to have a joint probability distribution for the assignment of respondents to categories. We use the following notation in our analysis:

- $X_t = k$: Event that the $t^{th}$ respondent belongs to the $k^{th}$ category.
- $X = k$: Event that the $N$ respondents belong to certain specified categories $k$. (Note that $X$ and $k$ are $N$ component vectors).
- $P[X = k]$: Prior probability of the snooper that the $N$ respondents are distributed as specified by $k$.

The notion that a snooper has no specific prior knowledge about the individual respondents implies that the snooper does not distinguish among the $N$ respondents. This can be interpreted as exchangeability, and hence $P[X = k]$ is invariant under different permutations of the components of $k$. It follows that the probability that a respondent $t$ belongs to a particular category $k$ is the same for all respondents. That is, there is some probability vector $\pi = [\pi_1, ..., \pi_K]$ such that

$$P[X_t = k] = \pi_k \text{ for } t = 1, ..., N.$$  

In the context of our example, we could say that the snooper is assumed to have exchangeable prior probabilities for the distribution of the respondents among the categories. Such a snooper is called undiscriminating. As mentioned before, an undiscriminating snooper would not have any specific information about any of the respondents and would hence have priors based on beliefs about the population. Table II presents one such set of priors where the prior probability that any respondent is from the high risk group and has his HIV status negative is 0.76.

Further, let $F_j = f_j$ denote the event that the $j^{th}$ category has $f_j$ respondents, and, for a sequence of queries regarding categories $a_1, ..., a_s$ where

$$S = (a_1, ..., a_s) \subseteq (1, ..., K),$$

let

$F^s = f^s$ denote the event that $F_{a1} = f_{a1}, ..., F_{as} = f_{as}$.

With this structure we can show how much information an undiscriminating snooper gains about whether a particular respondent belongs to a specified category. The result is summarized in Proposition 1.

**Proposition 1.** Under exchangeability, after the queries, "How many respondents are in category $j$?" have been answered for all categories in $S$, the probability that respondent $t$ is in category $k$ changes from its prior value of $P[X_t = k] = \pi_k$ for $k = 1, ..., K$, to a posterior value of:

$$P[X_t = k | F^s = f^s] = \frac{f_j/N}{(1 - \sum_{i \in S} f_j/N)\pi_k} P[F^s = f^s | X_t = k]$$

for $k \in S$.

Proof: Refer to [14].

Table III. Posterior probabilities for the distribution of respondents

<table>
<thead>
<tr>
<th></th>
<th>HIV Status Positive</th>
<th>HIV Status Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Risk Group</td>
<td>0.01</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Low Risk Group</td>
<td>0.04</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>Total</td>
<td>0.05</td>
<td>0.95</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Thus, after a query has been made about the number
of respondents in the low risk group whose HIV status is negative, we find the posterior probabilities (computed in accordance with Proposition 1) to be as shown in Table III.

Based on these results, a question about any category affects not only the posterior probability of that category, but also the posterior probabilities associated with other categories about which a query has not been made. This is because a query contains information not only about the number of respondents in a particular category, but also about the number of respondents in all other categories. As we shall observe in the next section, the disclosure control mechanism under the QSR strategy focuses solely on the category being queried. It is in essence this failure to take into account the effect of a query on all other categories that can often lead to a disclosure.

Also, the results from Proposition 1 indicate that once a query about a particular category has been made, subsequent queries do not alter the probability that a respondent belongs to that category. Under the Bayesian framework, this can be interpreted as the fact that a query about a category \( k \) determines the posterior probability that \( X_t = k \) as \( f_k/N \) with probability 1. It is this certainty in the query set size determination that allows the compromise of the database using trackers. As noted earlier, RSQ control deals with the basic principle of compromise by making it impossible for a questioner to control precisely the formation of query sets. This notion will be reflected in our analysis of RSQ in Section 4. We shall further demonstrate that repeated queries about a category may reduce the uncertainty significantly, and hence lead to disclosure. We now investigate the query size restriction approach within our probabilistic framework.

3 Query Set Size Restriction Control

The QSR control method permits a statistic to be released only if the size of the query set lies within some defined range. In the context of our model, this could be stated as the following disclosure limiting rule:

**Respond to the query regarding category \( j \) if and only if:**

\[ z < f_j < N - z, \]

where, \( z \) and \( N - z \) form lower and upper bounds, respectively, for the number of respondents in the category of interest, for a query to be allowed.

The intuition behind such a restriction is that if there are only a few respondents with some unique characteristics, then they may be easily identified and sensitive information about them consequently disclosed [3].

For instance, in the context of our example, consider the case where the database contains an additional attribute about the number of drug related convictions for each respondent. The snooper, let us assume, represents a medical insurer and is interested in this sensitive information about an individual who belongs to the high risk group and has tested HIV positive. Without RSQ control, the snooper’s query about the number of respondents in the category of his interest would yield a response 1. A subsequent query about the total number of drug related convictions for all respondents in that category would then provide the snooper with the exact sensitive information he desired. QSR control could prevent such direct disclosure. Thus, if queries of size less than 3 or greater than 97 were not allowed, then the above compromise would apparently not have been possible.

We shall now show how even under QSR, the probability that the snooper correctly infers the sensitive information is high. Consider the case where the snooper first makes a legitimate query about the number of individuals in the low risk group who have tested HIV negative. Based on the answer to this query, the snooper arrives at the posterior probabilities presented in Table III. Notice that the probability that a respondent belongs to the high risk group and has tested positive is now only 0.01. Hence the probability (from the binomial \( [100, .01] \) distribution) that more than one respondent belongs to this category is less than 0.26. Thus, with a degree of certainty, the snooper may infer that the subject of his interest is the only individual in the category.

The above example illustrates that a disclosure control rule which focuses solely on the category being queried may lead to an unacceptably high risk of disclosure. In order to minimize such risks, we modify the RSQ control rule and express it within our probabilistic framework as:
Respond to the query regarding category \( j \) if and only if:
\[
\alpha < \pi_i < 1 - \alpha \quad \text{for} \quad i = 1, \ldots, K
\]
where \( \alpha \) and \( 1-\alpha \) form the lower and upper bound for the posterior probability \( \pi_i \) that a respondent belongs to any category \( i \), for a query to be allowed.

Because of Proposition 1, for the specific category \( j \) that is queried, this modified rule is equivalent to the usual rule for a choice of \( \alpha = f_j/N \). But, unlike the usual rule, this modified rule does accommodate the changes in the probabilities for the categories \( i \neq j \) that are not queried, hence reducing the risk of indirect disclosure. This notion may be formalized as follows:

**Definition 1.** A disclosure control is said to be query specific if it is a function of only the posterior probability of the category regarding which the query was made. That is, the disclosure limiting rule is query specific if for some \( \alpha < 1/2 \) it states: For a query regarding category \( i \), respond if:
\[
\alpha < \frac{f_i}{N} < 1 - \alpha.
\]

**Definition 2.** A query specific disclosure control is said to be universal if and only if it protects all other categories. That is, for a query regarding category \( i \), if \( \pi_1, \ldots, \pi_K \) are the posterior probabilities for the categories, then,
\[
\alpha < \frac{f_i}{N} < 1 - \alpha \Rightarrow \alpha < \pi_j < 1 - \alpha \quad \text{for} \quad j = 1, \ldots, K.
\]

**Definition 3.** A snooper is said to be ignorant if the prior probabilities for the \( K \) categories are all equal, i.e., \( \pi_k = 1/K \) for \( k = 1, \ldots, K \).

With this structure we can prove the following result:

**Proposition 2.** A universal query specific disclosure control exists if and only if the snooper is ignorant and \( K = 2 \).

**Proof:** Refer to [14].

Proposition 2 indicates that for any practical database a universal query specific disclosure control will not exist. The disclosure limiting rule hence requires that at least a subset of the remaining categories must be checked in order to protect every category. In practice, this task can be prohibitively expensive because of the computation required.

In conclusion to this section, our analysis lends further support to the growing view in the literature [1,10] that query set size restriction control is in itself not very promising. Based on this negative evidence, we next explore a quite different form of disclosure limitation, random sample query control.

### 4. Random Sample Query Control

An alternative to the query size restriction approach is random sample query (RSQ) control [9,21]. Building on the analysis of the previous section, we investigate the effect of using RSQ on the snooper’s probabilities.

RSQ control works this way: Upon a query with true response \( F=f \), a response set \( R \) is constructed from these \( f \) respondents such that the probability of including any respondent in \( R \) is independently \( p \), \( 0 < p < 1 \). The number of respondents in the set \( R \) is released to the questioner. The questioner is also informed of the value of \( p \).

With RSQ control we can determine the snooper’s gain in information about the number \( f \) of respondents in a particular category, given the information that the response set \( R \) has \( f \) respondents. The result (based on [9]) is summarized in Proposition 3.

**Proposition 3.** Under RSQ, given that the response to a query about a particular category is \( F^* = f^* \), the probability that the actual number of respondents in that category is \( F = f \), changes from a prior value of \( P[F = f] = \pi_k \) for \( k = 0, \ldots, N \), to a posterior value of
\[
P[F = f | F^* = f^*] = \begin{cases} 0 & \text{for } f < f^* \\ \frac{f !}{f^* !}\frac{\pi_k^{(f-p)}}{G^{(r)}(q)}(f^*-f^*)! & \text{for } f \geq f^* \end{cases}
\]

where \( q = 1-p \), \( G(z) \) is the probability generating function for \( \{\pi_k\} \), and \( G^{(r)}(z) \) is the \( r \)th derivative of \( G(z) \).

**Proof:** Refer to [14].

As a special case, we consider the distribution of \( F \) to be binomial, with parameters \( N \) and \( \pi \). The resulting generating function is hence given by
\[
G(q) = (1 - \pi q)^N, \quad \text{and, the } r\text{th derivative of } G(q) \text{ is}
\]
\[
G^{(r)}(q) = \frac{N! \pi^r (1-\pi q)^{N-r}}{(N-r)!}.
\]

Substituting these expressions in the results of
Proposition 3, we obtain,

\[ P[F=f \mid F^*=f^*] = \begin{cases} 0 & \text{for } f < f^* \\ \frac{(N-f^*) \pi f^* \gamma f f^* \left[ 1-\pi f^* \right]^\gamma \left[ 1-\pi f \right]^{\gamma f^*} \gamma f & \text{for } f \geq f^* \end{cases} \]

From the above expressions, it can be recognized that the distribution of \( (F - F^* \mid F^* = f^*) \) is also binomial, now with parameters \( N-f^* \) and \( \pi f^*/(1-\pi f) \).

We exploit these results to gain some interesting insights about the strengths and weaknesses of this control strategy. First, through Proposition 4, we show that RSQ overcomes the problem of set sizes being determined with certainty, as encountered under QSR. And then, through Proposition 5, we extend our analysis to show how repeated queries may significantly reduce this uncertainty about the number of respondents in a category.

The posterior probabilities computed in section 2 may be computed under RSQ control. The evaluated probabilities are summarized in Proposition 4.

**Proposition 4.**

\[ P[X_j = k \mid F^*_i = f^*_i] = \begin{cases} \frac{1}{N} & \text{for } j = k \\ \frac{1-\frac{f^*_i}{N}D_f}{1-\pi \gamma} & \text{for } j \neq k \end{cases} \]

where \( \pi = P[X_j = j] \) for \( j = 1, \ldots, K \).

Proof: Refer to [14].

Notice that unlike in the case of query size restriction control, the probability associated with the category about which a query has been made remains a function of the prior probability and may hence change with further information. This element of uncertainty prevents disclosure by using trackers.

Our analysis thus far has been restricted to the case where the snooper makes a single query about each category. However, by making repeated queries about the same category, the snooper could reduce his uncertainty about the true number of respondent in that category. We formalize this notion and present our results in Proposition 5.

**Proposition 5.**

Under RSQ control, if \( k \) repeated queries about a particular category yields the response sequence \( F^*_1 = f^*_1, \ldots, F^*_k = f^*_k \), then the probability that the true number of respondents in that category changes from a prior value of \( P[F=f] = a_f \) to a posterior value of

\[ P[F=f \mid \bigcap_{i=1}^k [F^*_i = f^*_i]] = \begin{cases} 0 & \text{for } F < \max_i f^*_i \\ \frac{\prod_{i=1}^k P[F^*_i = f^*_i]}{\prod_{i=1}^k [P[F^*_i = f^*_i]]} & \text{otherwise}. \end{cases} \]

Proof: Refer to [14].

In order to investigate the implications of the above results, we computed the posterior probabilities assuming a binomial distribution for \( F \) with parameters \( N \) and \( \pi \).

Figure 1. Decreasing uncertainty with increasing number of queries. The enclosed region includes 99.8 percent of the posterior probability.

Figure 1 presents the case where \( N=1000, \pi=0.01, p=0.75 \) and the "true" value of \( F \) is taken to be 10. The maximum probability region (which includes 99.8%) is shown for the cases when 1, 5, 10, 15, and 20 queries are made about the category of interest. For
each query, the number of respondents in the response set under RSQ was obtained by simulating binomial random variables with parameters \( F \) and \( p \). As evident from the figure, as the number of queries, \( k \), increases, the actual number of respondents in the category becomes increasingly apparent to the snooper. This is particularly true for categories with few respondents and hence suggests the need for a limit on the number of queries about a category. The literature on query overlap restriction adopts this approach [11]. In view of the fact that such control mechanisms are expensive, perhaps prohibitively so, in terms of their computational and storage requirements, we shall not consider this control strategy in this paper. Instead, we exploit our framework to suggest an alternative scheme by combining QSR with RSQ and show how this leads to a significant reduction in the risk of disclosure.

5. Risk of disclosure when QSR and RSQ are combined

As we have discussed, the Query Set Size Restriction (QSR) technique aims to control the disclosure of confidential information in on-line statistical databases by restricting queries which involve query sets which are either too large or too small. However, this technique can be easily subverted by the use of tools known as trackers [8]. A suggested remedy is the Random Sample Query (RSQ) control [9] which makes it impossible for a questioner to control precisely the formation of query sets and hence reduces the risk of compromise using trackers. In this section we investigate the probability of the disclosure of confidential information using trackers in statistical databases which employ both QSR and RSQ as disclosure control mechanisms. Since trackers are recognized to pose the most threatening problem for database security [23] we focus our analysis on the risk of disclosure using trackers.

Simplifying the notation presented in Section 2, we recall that the statistical database is modeled as a collection of \( N \) records. A query on the database is an arbitrary logical formula \( C \) specifying a subset of records and \( |C| \) returns the number of records which satisfy \( C \). The QSR inference control is based on restricting responses to only those queries for which:

\[ k \leq |C| \leq n-k \]

for some \( 0 < k < n/2 \)

As noted earlier, though this prevents the direct identification of any small set of records with some unique characteristics, trackers may be used to infer the number of records in such restricted sets. In particular, a general tracker is a logical formula \( T \) such that \( 2k < |T| < n-2k \). Even if a query \( C \) is unanswerable (because \( |C| < k \) or \( |C| > n-k \)), the tracker can be employed to calculate \( |C| \) using the formula:

\[ |C| = |C+T| + |C+\neg T| - |T| - |\neg T| \]

if \( |C| < k \)

\[ = 2|T| + |\neg T| - |C+T| - |\neg C+T| \text{ if } |C| > n-k. \]

Each of the above subqueries are answerable, and hence \( |C| \) can always be computed. The general tracker is guaranteed to work for \( k \leq n/4 \). (Note that in set theoretic terms, \( + \) represents union and \( \neg \) represents complement.)

When RSQ is used as a control mechanism, for any query \( |C| \) a response set is formed by including, with probability \( p \), each record that satisfies the formula \( C \) in the response set. The response given to the questioner is the number of records, \( |C|^* \), in the resulting response set. Thus \( |C|^* \) follows a binomial distribution with parameters \( |C| \) and \( p \). However, for small query sets the true set size \( |C| \) can be determined quite accurately from the RSQ response \( |C|^* \) by repeating the query a number of times.

We now investigate the effect of RSQ control on the ability of the general tracker to infer the true number of records \( |C| \) for query sets in the range restricted by QSR control. Consider the case:

\[ |C| = f < k. \]

Without RSQ control, a general tracker \( T \) could be used to compute

\[ |C| = |C+T| + |C+\neg T| - |T| - |\neg T|. \]

However, under RSQ, the responses to the queries in the right hand side of the above equation are \( |C+T|^* \), \( |C+\neg T|^* \), \( |T|^* \), and \( |\neg T|^* \). These values may be used to estimate:

\[ |C|^* = |C+T|^* + |C+\neg T|^* - |T|^* - |\neg T|^*. \]

Now \( |C+T|^* \) and \( |C+\neg T|^* \) are independent binomial variables with the same parameter \( p \) as their probability of success and with number of trials \( |C+T| \) and \( |C+\neg T| \) respectively. Hence \( X_1 = \frac{|C+T|^* + |C+\neg T|^*}{|T|^* + |\neg T|^*} \) has a binomial distribution with parameters \( |C+T| + |C+\neg T| \) and \( p \). But \( |C+T| + |C+\neg T| = N+f \) and hence \( X_1 \) has a binomial distribution with parameters \( N+f \) and \( p \). In databases where the total number of records, \( N \), is large, the distribution of \( X_1 \) can be approximated by a normal distribution with a mean of \( (N+f)p \) and a variance of \( (N+f)p(1-p) \). By a similar argument \( X_2 = \frac{|T|^* + |\neg T|^*}{|T|^* + |\neg T|^*} \) has a binomial distribution with parameters \( N \) and \( p \). Again, for a large \( N \), this can be
approximated by a normal distribution with mean $Np$ and variance $Np(1-p)$. Now, $|C| = X_1 - X_2$. But $X_1$ and $X_2$ are independent with distributions that can be approximated by normal distributions. Hence $|C|^*$ can be approximated by a normal distribution with a mean of $fp$ and a variance of $(2N+f)p(1-p)$.

Using the approximate distribution for $|C|^*$, we can estimate the probability that a tracker may be used to determine the true number of records $f < k$ in a restricted range, under RSQ control. For example, in the case where $N = 100$, $p = 0.9$ and $f = 2$, we have: $P[\text{tracker disclosure}] = P[1.5 < |C|^* < 2.5] = 0.0932$. Comparing this result with that obtained using the exact binomial distributions for $X_1$ and $X_2$ we find: $P[\text{tracker disclosure}] = P[X_1 - X_2 = f] = 0.0937$. Thus the error introduced by the normal approximation is less than 1 percent for this example.

The following table displays probabilities of disclosure using trackers under RSQ with three different values for $N$ (100, 200, and 300) and 4 different values for $f$ (1, 2, 4, and 10). The numbers in parenthesis indicate the percentage error involved if the approximate distribution for $|C|^*$ is used. For example, with $N = 100$, $f = 1$ and $p = 0.9$, $P[|C|^* = f] = 0.0940$ and if the normal approximation is used to compute this probability, the error is 0.43 percent.

Table IV. Probabilities of disclosure using trackers.
Percentage error of normal approximation appear in parentheses. ($p=0.9$)

<table>
<thead>
<tr>
<th>$f/N$</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0940 (.43)</td>
<td>0.0665 (.30)</td>
<td>0.0543 (.18)</td>
</tr>
<tr>
<td>2</td>
<td>0.0937 (.53)</td>
<td>0.0664 (.30)</td>
<td>0.0542 (.18)</td>
</tr>
<tr>
<td>4</td>
<td>0.0930 (.54)</td>
<td>0.0661 (.27)</td>
<td>0.0541 (.18)</td>
</tr>
<tr>
<td>10</td>
<td>0.0897 (.56)</td>
<td>0.0649 (.31)</td>
<td>0.0534 (.19)</td>
</tr>
</tbody>
</table>

The above results indicate that the normal approximation is fairly accurate for these values of $N$ and $p$. More significantly, if the database contains at least a hundred records, even with $p$ as high as 0.9, the probability that the true value $f$ in the restricted range can be determined is less than 10 percent. In many practical applications the resulting risk of disclosure may be considered acceptable.

In our analysis thus far we have not considered the possibility of repeated queries being made about the same set of records. Let the responses to any such $r$ queries for which the true response is $|C|$ be represented by the sequence $|C|_1^*, ..., |C|_r^*$, under RSQ. Now $|C|_1^*, ..., |C|_r^*$ are independent binomial random variables with parameters $|C|$ and $p$. The problem of determining the true number of records from the sequence of responses obtained hence involves the estimation of the binomial parameter $|C|$. Results from Feldman and Fox [18] show that the maximum likelihood estimator (MLE) for the binomial parameter $n$ with known $p$ is asymptotically normal with mean $n$ and variance $n(1-p)/pr$.

In order to investigate the properties of the MLE for a range of values of $n$, $p$, and $r$, we generated binomial random observations with $n = 100, 200, and 300$, $p = 0.7, 0.8, 0.9, and 0.95$, and $r = 5, 10, and 20$. Based on groups of $r$ observations for each combination of $n$ and $p$ we computed the MLE of $n$ at least a hundred times. Simulation results indicate that the hypothesis that the MLE has a normal distribution cannot be rejected at a level of significance $\alpha = 0.5$. The sample mean and variance for the MLE were within 1 percent of the predicted results. This suggests that the asymptotic results well approximate the distribution of the MLE for the ranges of $n$ and $p$ examined. We shall now show that for databases with the number of records $n \geq 100$, this result can help us determine the probability of disclosure using trackers if queries are repeated $r$ times.

As we have observed before, $X_1 = |C+T|^* + |C+\neg T|^*$, has a binomial distribution with parameters $N+f$ and $p$. Queries are repeated to obtain $r$ observations for $X_1$ based on which the maximum likelihood estimate for $N+f$ is computed. Let $\hat{n}_1$ be the MLE for the parameter $N+f$. Hence $\hat{n}_1$ has an approximate normal distribution with a mean of $N+f$ and a variance of $(N+f)(1-p)/pr$. By a similar argument, the MLE $\hat{n}_2$ of the parameter $n$ for the distribution of $X_2 = |T|^* + |\neg T|^*$ has an approximate normal distribution with mean $n$ and variance $N(1-p)/pr$.

If an estimator $\hat{n} = \hat{n}_1 - \hat{n}_2$ is used to estimate $|C|$, then the distribution of $\hat{n}$ can hence be approximated by a normal distribution with a mean of $f$ and a variance of $(2n+f)(1-p)/pr$. This distribution can be used to estimate the probability that $\hat{n}$ yields the true value $f$ of $|C|$ and hence constitutes a tracker.
Tracker disclosure refers to the correct inference about the number of records in a restricted range using a tracker. For example, with $N = 100$, $p = 0.9$, $f = 2$ and $r = 5$, we get:

$$P[\text{tracker disclosure}] = P[1.5 < \hat{\mu} < 2.5] = 0.1866,$$

and by increasing the number of repeated queries $r$ to 10, we have:

$$P[\text{tracker disclosure}] = P[1.5 < \hat{\mu} < 2.5] = 0.2614.$$

Figure 2. Probability of disclosure using trackers

Figure 2. shows how the probability of tracker disclosure increases as the number of repeated queries, $r$, increases from 1 to 100. With $N = 100$, $p = 0.9$, and $f = 2$, each query must be repeated at least 40 times in order to have a probability of tracker disclosure of 0.5. Even when the queries are repeated 100 times, the probability increases to less than 0.7. In statistical databases where audit trails are maintained, such an unusually high repetition of queries could be easily detected. Besides, the cost involved in repeating each query so many times may act as a deterrent against attempted disclosure.

6. Conclusions

Our framework allows the analysis of disclosure control mechanisms using a probabilistic approach. We have illustrated how the method may be employed to assess the strengths and weaknesses of two existing disclosure control strategies -- query size restriction (QSR) control and random sample query (RSQ) control. The framework can further be exploited to investigate alternative schemes and suggest new mechanisms with desirable characteristics. In order to demonstrate this ability, we suggested one such disclosure control mechanism by combining QSR and RSQ control schemes, which significantly reduces the risk of disclosure of confidential information.

In any such control, the probability of disclosure decreases with an increasing restricted range, $k$, and a decreasing probability of inclusion of records, $p$. However, as $k$ increases and $p$ decreases, less and less information is made available to the database user. The objective of providing maximum possible access to information hence conflicts with the need to limit the risk of disclosure of confidential information. The problem can be viewed as one of selecting optimal design parameters $k$ and $p$ for the control mechanism so as to minimize the loss of information while maintaining the risk of disclosure below some acceptable threshold. Our ongoing research addresses this particular problem.

References.


interfile exchange for useful statistical analysis. Evaluation Quarterly, 1, 269-299.


