Protecting patient privacy while preserving medical information for research

Gang Xiang[1)], Jason O'Rawe[1,2)], Vladik Krienovich[3)], Janos Hajagos[2)], and Scott Ferson[1)]
[1)]Applied Biomathematics, Setauket, New York 11733 USA, scott@ramas.com
[2)]Stony Brook University, Stony Brook, New York 11794 USA, jazony33@gmail.com
[3)]University of Texas at El Paso, El Paso, Texas 79968 USA, vladik@utep.edu

Abstract: Patient health records possess a great deal of information that would be useful in medical research, but access to these data is impossible or severely limited because of the private nature of most personal health records. Anonymization strategies, to be effective, must usually go much further than simply omitting explicit identifiers because even statistics computed from groups of records can often be leveraged by hackers to re-identify individuals. Methods of balancing the informativeness of data for research with the information loss required to minimize disclosure risk are needed before these private data can be widely released to researchers who can use them to improve medical knowledge and public health. We are developing an integrated software system that provides solutions for anonymizing data based on interval generalization, controlling data utility, and performing statistical analyses and making inferences using interval statistics.

Keywords: intervals; patient privacy; data informativeness; anonymization; *k*-anonymity.

1. Introduction

Although data collected during health care delivery and public health surveys possess a great deal of information that could be used in biomedical and epidemiological research, access to these data is usually restricted because of the private nature of most personal health records. Simple strategies to anonymize data do not protect privacy. For example, dropping explicit identifiers such as name, social security number and address makes the data appear anonymous, but the remaining attributes can often be used to re-identify individuals. For instance, Golle (2006) showed that 63% of the US population can be uniquely identified by gender, ZIP code, and date of birth, and those three attributes often exist in health data put in open archives (Sweeney 2002). This information is widely distributed, so it can easily be obtained and used for re-identification.

The difficulty in balancing privacy with scientific utility of released health care data has been a substantial barrier to progress in medical research. Enormous bodies of scientifically valuable health care data, often collected at considerable public and private expense, are held by a wide variety of organizations such as NIH, CDC, FDA, EPA, hospitals, universities, and insurance companies. Even organizations such as prison systems, cruise lines, etc., are also custodians of significant information resources that would be useful in furthering medical and scientific research that could be used to improve public health. But these data sets generally contain private and sensitive information about patients. When releasing data, an organization must guarantee that individuals cannot be identified from the data, otherwise their sensitive information could be disclosed, which would subject the data custodian to legal liability under health

© 2014 Xiang et al. *Printed in USA.*

privacy laws such as HIPAA. This very often means that there is little after-use of the extensive data originally collected in medical claims, clinical studies, Medicaid administration, public surveys, and even medical research at universities and hospitals.

Techniques for statistical disclosure control and methods for data anonymization based on attribute generalization have been developed to ensure privacy but information truthfulness is often not well preserved by the former so that unreliable results may be released, and information loss in the latter do not provide sufficient control for maintaining data utility. To be useful in medical research, any released data must be accurate and as precise as possible, which entails preserving the statistical information present in those data, including importantly the dependencies among variables. Management tools are needed for electronic health records that can help balance privacy concerns with scientific informativeness needed to improve medical knowledge by providing the technical capability to find relationships in the large bodies of data that have been collected but are currently inaccessible. Such tools require strategies that somehow protect both the privacy of individuals as well as the integrity of the statistical relationships in the data. The problem is that there is an inherent tradeoff between these. Protecting privacy always loses information. However, for a given anonymization strategy, there are often several ways of masking the data that all meet the given disclosure risk criteria. This multiplicity can be exploited to choose the solution that best preserves statistical information while still meeting the disclosure risk criteria. We describe efforts to develop an integrated software system that provides solutions for managers of data sets so they can minimize disclosure risks while maximizing data informativeness.

2. Anonymization Strategies

Maintaining privacy in released data is universally acknowledged as important and it is not a new problem, but current anonymization methods have severe limitations. The most basic approach is to release data after removing explicit identifiers like name, social security number and full address, so that the data appear anonymous. However, this approach is inadequate for privacy protection because the residual data are not actually anonymous; the remaining attributes in the released data may be used to re-identify individuals. Golle (2006) showed that more than half of people in the United States can be uniquely indentified by gender five-digit ZIP code and birthday. Sweeney (2002) notes that three attributes often exist in health data put in open archives (Sweeney 2002). This information is seemingly non-sensitive and is widely distributed, so it can easily be obtained and used for re-identification. Common health care data sets can be much smaller than the census data, so re-identification will often be even easier.

The failure of simple remove-explicit-identifiers approaches has spawned research on statistical disclosure control techniques (Willenborg and De Waal 1996). Statistical databases are sometimes equipped with a limited querying interface that only allows the release of aggregate statistical information. However, such limited querying interfaces alone are insufficient to protect privacy. Even though statistical results themselves might appear insensitive because they are at the population level, an adversary may still be able to craft a series of queries and combine multiple statistical results to re-identify an individual and disclose sensitive information. For instance, if statistical results show that the average systolic blood pressure of 10 patients aged from 40 to 50 in a study group is 110 mmHg and the average systolic pressure of 11 patients aged from 40 to 51 in the group is 115 mmHg, the adversary can get the individual systolic blood pressure of the patient aged 51 as 115×11−110×10 = 165 mmHg, and this patient is easy to identify since s/he can be recognized as the only 51-year old in the group.

Many sophisticated methods exist for such re-identification attacks on statistical data. For this reason, various statistical disclosure control techniques such as cell suppression, data perturbation, data swapping, and generation of synthetic data are used to maintain privacy in statistical databases. These techniques have been widely used in releasing statistical information by federal data holders like CDC and US Census. Unfortunately, traditional statistical disclosure control techniques are inadequate in today's data rich settings. These methods can protect privacy with varying degrees of success, but they often have substantial costs to the utility of the data because information truthfulness is not well preserved and these methods may release unreliable results (Sweeney 2002a) when they alter the statistical information in the data.

In response to the limitations of traditional statistical disclosure control techniques, *generalization-based anonymization approaches* have been developed, in which explicit identifiers are removed and non-sensitive attributes are then used to re-identify individuals using quasi-identifiers (Sweeney 2002a). The data set is then modified so that this re-identification becomes harder. Re-identification can be avoided by attribute generalization on quasi-identifiers (Samarati 2001; Sweeney 2002b). Numerical attributes are replaced by intervals that contain the original exact value (e.g., [10, 20] in place of 12), and categorical attributes are usually replaced by wider classifications (e.g., "student" in place of "high school student"). The more records that share the same values of their quasi-identifiers, the more difficult re-identifying individuals becomes.

There are several different methods of applying generalization-based anonymization approaches. Sweeney (2002a) introduced a privacy protection model called *k*-anonymity, in which every record in the released data is indistinguishable from at least $k-1$ other records with respect to every set of quasi-identifier attributes. A set of such indistinguishable records is called an equivalence class (Li et al. 2007). The use of *k*-anonymity is popular due to its conceptual simplicity and the fact that there are many efficient algorithms for creating a *k*-anonymity data set (Bayardo and Agrawal 2005; LeFevre et al. 2005; Meyerson and Williams 2004; Sweeney 2002a; Zhong et al. 2005).

During recent years, various enhancements on *k*-anonymity have also been proposed to overcome its shortcomings. One of the limitations of *k*-anonymity is that sensitive information can be disclosed due to lack of diversity in the attribute among records in an equivalence class (Machanavajjhala et al. 2006; Truta and Vinay 2006; Xiao and Tao 2006). For example, if all patients in one equivalence class have the same disease, the sensitive information of disease specific to any individual in this equivalence class will be disclosed even though they are indistinguishable. In response to this limitation, Machanavajjhala et al. (2006) described an enhanced model, *l*-diversity, in which an equivalence class contains at least *l* "well-represented" values for the sensitive attribute, and an algorithm for *k*-anonymity was extended for creating *l*-diversity data sets (LeFevre et al. 2005). Compared to *k*-anonymity, *l*-diversity usually replaces values of quasi-identifiers with more general values, so the equivalence class is larger to ensure diversity in the sensitive attribute. Further extensions of the *k*-anonymity model include *t*-closeness (Li et al. 2007), *p*-sensitive *k*-anonymity (Truta and Vinay 2006), (α, *k*)-anonymity (Wong et al. 2006), (*k*, *e*)-anonymity (Zhang et al. 2007), (*c*, *k*)-safety (Martin et al. 2007), *m*-confidentiality (Wong et al. 2007), personalized privacy (Xiao and Tao 2006), etc**.**

It also worth mentioning that even though the common definition of quasi-identifiers only refers to non-sensitive attributes whose values are easily accessible (Machanavajjhala et al. 2006), sensitive attributes can also be used for re-identification (Gal et al. 2008; Li and Ye 2007). Sensitive attributes are not widely distributed but are frequently attainable. If multiple sensitive attributes are linked to one record, knowledge of one sensitive attribute specific to an individual can be used to reveal other sensitive attributes. For example, knowing a patient's salary might help disclose a medical condition. For this possibility, Li and Ye (2007) introduced an improved generalization-based privacy protection procedure which also generalizes sensitive attributes.

Compared to traditional statistical disclosure control techniques, generalization-based approaches preserve information truthfulness and the resulting anonymized data are guaranteed to be reliable in the sense that they do not contain misrepresentations or fabrications of the information in the original data, just less precision about it. However, because attribute generalization adds uncertainty there is unavoidable information loss, thus the utility of the released data is decreased. For example, from an interval [10, 20] in the released data we cannot determine the original exact value, and we only know it is between 10 and 20. This situation is called *interval uncertainty*. It is important to have proper control on utility during anonymization, so that the released data continue to have inferential power.

All generalization-based anonymization algorithms aim to gain privacy protection with control on utility. Otherwise, maximum privacy protection could always be obtained by replacing quasi-identifiers with the most general values possible (e.g. [0%, 100%]). Generalization-based algorithms optimize anonymization solutions using one of two approaches: *privacy-constrained* (Ghinita et al. 2009) anonymization maximizes utility while meeting the requirements for privacy (such as a given value of k for k-anonymity or l for l-diversity), while *utility-constrained* or *accuracy-constrained* (Ghinita et al. 2009) anonymization maximizes degree of privacy (e.g., the value of k or l) while meeting the requirements for utility. In both approaches, data utility must be quantified, but directly quantifying utility is generally viewed as difficult (Machanavajjhala et al. 2006), so it is of interest to develop cost metrics measuring information loss. Common information loss metrics include Generalization Height (LeFevre et al. 2005, Samarati 2001), average size of the equivalence classes, and Discernibility Metric (Bayardo and Agrawal 2005). Generalization Height, intuitively, is the number of generalization steps that were performed. The second metric is the average number of records in the equivalence classes generated by the anonymization algorithm. The Discernibility Metric measures the number of tuples that are indistinguishable from each other. It is equivalent to the sum of the squares of the sizes of the equivalence classes. Other information loss metrics developed recently include Classification Metric (Iyengar 2002), Ambiguity Metric (Nergiz and Clifton 2006), and Loss Metric (Iyengar 2002, Nergiz and Clifton 2006).

Efficient anonymization algorithms are available with control of information loss based on various cost metrics (Bayardo and Agrawal 2005; Ghinita et al. 2007; Ghinita et al. 2009; Gionis and Tassa 2009; LeFevre et al. 2005; Meyerson and Williams 2004; Miller et al. 2008; Sweeney 2002a). However, current cost metrics are insufficient to control utility of anonymized data, because they only measure information loss, or degrees of anonymization, but not utility of the data for statistical inference. Although more information loss/higher degree of anonymization implies less utility, the maps from information loss to utility are unclear. Given an anonymized data set with some degree of information loss/anonymization in term of such a cost metric, it is still unclear how useful the data would be.

3. Optimizing Anonymization to Improve Utility

We can use cost metrics to control anonymization procedures. These cost metrics should directly reflect data utility and should be understandable to data consumers, so that data holders could release anonymized data to meet both privacy protection requirements and data consumers' expectations of utility. To find appropriate cost metrics, we need to look at how data are actually used in practical situations. In fact, using data in health care research usually involves statistical analysis, i.e., computing statistics on numerical attributes. For example, researchers might want to know the average and variance of the ages of patients who have a certain disease. In a generalization-based anonymization procedure, original values of many numerical attributes are replaced by intervals for the purpose of privacy protection. Storing intervals $[\underline{x}_1, \overline{x}_1], \dots, [\underline{x}_n, \overline{x}_n]$ instead of exact numerical values leads to difficulty in computing statistics. In principle we can only get the range of the statistic $\{C(x_1, \dots, x_n): x_1 \in [\underline{x}_1, \overline{x}_1], \dots, x_n \in [\underline{x}_n, \overline{x}_n]\}$, which is the set of all values of some statistic C that could arise from any combination of input values from their respective intervals (Kreinovich et al. 2006, Kreinovich et al. 2007). For example, based on some intervalized age data [20, 25], [25, 30], …, the average and variance of ages might be computed as ranges [32, 37] and [2, 5], respectively. Problems of computing statistics on interval data have been studied by computer scientists and mathematicians during recent years, and efficient algorithms have been developed (Kreinovich and Xiang 2008; Ferson et al. 2007).

A computed range of a statistic on interval data is reliable, in the sense that it is guaranteed to contain the exact (but unknown) value of the statistic. However, its inferential power varies depending on the width of the range. An average of ages like [30, 35] is useful in research, but one like [0, 99] barely provides any beneficial information. Basically, more useful anonymized data have narrower computed ranges of statistics, which implies that values of numerical attributes should not be generalized into intervals too wide. Therefore, one natural idea is to use widths of computed ranges of statistics to measure utility of anonymized data, and to use these cost metrics to control generalizations of numerical attributes. For example, for research about the relation between age and a certain disease, it might be required that after anonymization the width of ranges be kept less than 5 for the average of ages and be kept less than 10 for the variance of ages. In this case, utility cost metrics are width of computed range of age average and width of computed range of age variance. The generalizations for categorical attributes is somewhat more complex but has the same motivations.

The above statistics-related utility cost metrics have simple mathematical formulas, and they are quite understandable to data consumers. Data consumers could easily exploit these utility cost metrics to clearly define their specific expectations of utility of released data and pass the expectations to data holders. On the other side, data holders could use data-consumer-specified utility cost metrics to control the anonymization procedure, so that the released anonymized data are guaranteed to meet data consumers' expectations of utility (and to meet privacy protection requirements as well). However, it is possible that such a balanced point of disclosure risk with inferential power cannot be reached by running an anonymization procedure. This can often happen in situations when the number of records in the original data set is not big enough. In this case, data holders could reach the conclusion that the original data set is not suitable for releasing yet. Further adjustments are suggested for the original data set, e.g., adding more records, before it can be released.

It is also worth mentioning some issues that arise when computing statistics on interval data. As described above, special methods to compute interval statistics are necessary to provide data consumers with reliable statistical results on intervalized data generated from anonymization procedures. However, computing statistics on interval data is, in general, computationally difficult. For example, it is known that the problem of computing the exact range $[\underline{V}, \overline{V}]$ for the variance V over interval data $[\underline{x}_1, \overline{x}_1], \dots, [\underline{x}_n, \overline{x}_n]$ is, in general, NP-hard (Ferson et al. 2002; Kreinovich et al. 2006; Kreinovich et al. 2007), [which means the problem becomes prohibitively expensive when n gets large]. However, efficient statistical algorithms exist for several practically important situations of interval uncertainty. For example, when no interval is a proper subset of another, such as [0,10], [5,15], [10,20], …, a linear-time algorithm exists for computing variance and an algorithm exists for computing skewness on the order of n^2 (Kreinovich et al. 2007, Xiang et al. 2007).

More computationally efficient results for several situations and the corresponding efficient statistical algorithms are described in recent reviews (Kreinovich and Xiang 2008; Ferson et al. 2007). However, even though these algorithms have been found to be computationally efficient, software exploiting them for the practical use has not been implemented yet. Our proposed software will also implement efficient algorithms for computing statistics on interval data as tools for data consumers to conduct statistical analysis on anonymized data. Most of these algorithms are results of our previous research, while new algorithms will be developed for conducting more complicated analyses in health care research.

4. Approach

Anonymization can be based on a privacy-constrained generalization-based approach (i.e., the approach which maximizes utility while meeting the requirements for privacy). This procedure is illustrated in Figure 1.

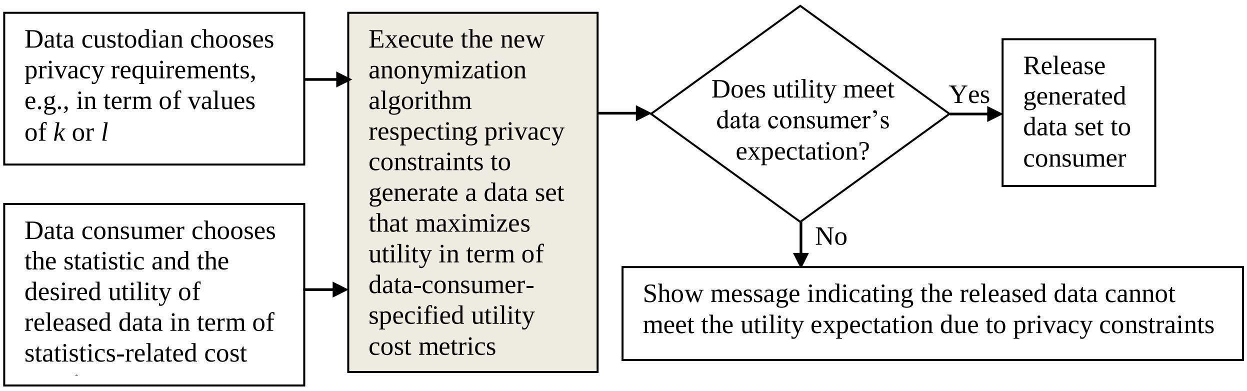
Data custodian chooses privacy requirements, e.g., in term of values of k or l
Data consumer chooses the statistic and the desired utility of released data in term of statistics-related cost
Execute the new anonymization algorithm respecting privacy constraints to generate a data set that maximizes utility in term of data-consumer-specified utility cost metrics
Does utility meet data consumer’s expectation?
Yes
Release generated data set to consumer
No
Show message indicating the released data cannot meet the utility expectation due to privacy constraints

Figure 1. Privacy-constrained generalization-based approach.

The critical component of this procedure is a set of new privacy-constrained anonymization algorithms (highlighted in the flowchart) which maximize utility with respect to the proposed statistics-related cost metrics. The algorithms are described in the next section.

Statistical analysis with intervals is generally computationally difficult (Ferson et al. 2002). To minimize the computational challenges associated with subsequent statistical analyses, it is also possible to select the *shapes* of the intervals used in the anonymization to exploit reduced-complexity algorithms available for those analyses. While computing the range of the average E under interval uncertainty is straightforward in any situation, computing the range for variance V, skewness $S = ((x_1 - E)^3 + \cdots + (x_n - E)^3)/n$, or the lower and upper endpoints of a confidence interval $L = E - k_0 \cdot \sqrt{V}$ and $U = E + k_0 \cdot \sqrt{V}$ (where k_0 is specified by the confidence level) over interval data has been proven to be NP-hard in general, or its difficulty is still unknown (Kreinovich et al. 2006; Kreinovich et al. 2007; Kreinovich and Xiang 2008, Xiang 2006). But we have developed efficient algorithms in several practical situations. In particular, in the case of “no-nesting”, for which no interval is a proper subset of another, we have developed an efficient polynomial-time algorithm for computing ranges for *V, S, L* and *U* (Kreinovich et al. 2006; Kreinovich et al. 2007; Kreinovich and Xiang 2008, Xiang 2006). We can generate non-nested intervals when we use attribute-generalization approaches for protecting privacy. This allows us to exploit these efficient algorithms within the new utility cost control functions.

5. Algorithm

Xiang (2012) showed that the optimal subdivision for maintaining *k*-anonymity, where there are at least *k* points within a box, while maximizing the utility for computing a statistical characteristic *C,* is achieved by selecting a box around each point $x = (x_1, \ldots, x_n)$ with half-widths

$$\Delta_i = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod_{j=1}^{n} a_j}}{a_i},$$

where $a_i = |\partial C/\partial x_i|$, and $\rho(x)$ is the data density around point *x*. However, Xiang (2012) did not provide algorithmic details to flesh out how this formula can be implemented. Here, we describe an easy-to-implement algorithm, which is based on this result with algorithmic enhancements.

The data density $\rho(x)$ can be estimated by using kernel methods. However, kernel methods usually lead to approximate values of $\rho(x)$, in which case we cannot guarantee that the derived box with half-widths Δ_i contains at least *k* points. Of course, we might attempt to get around this by trying gradually larger values of Δ_i, but this would increase complexity and make the algorithm more difficult to implement.

In principle, we do not necessarily need to calculate $\rho(x)$ to specify the box. For instance, when $n = 1$, we don’t need to consider any of the details in the equation for Δ_i. We can simply select the $k - 1$ points closest to the point *x* and use the minimum and maximum of these *k* one-dimensional points (including the point *x* itself) as the lower and upper bounds of the one-dimensional anonymization box. In the multi-dimensional case when $n > 1$, if we use *L* to denote the term

$$\frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \sqrt[n]{\prod_{j=1}^{n} a_j},$$

then the half widths of the box are $\Delta_i = L/a_i$, for each *i* from 1 to *n*. This means that the box keeps fixed ratios between any two different half widths when we define the box by changing the value of *L*. Because we need to select the smallest possible box to contain the point *x* and $k - 1$ points around *x*, this implies finding the smallest possible value of *L*.

We need to define the distance from a point $x^{(p)}$ to $x^{(q)}$, where both points are from the *N* given points $x^{(1)}, \ldots, x^{(N)}$. Naturally, the distance from $x^{(p)}$ to $x^{(q)}$ should be shorter than the distance from $x^{(p)}$ to $x^{(r)}$ if

1. the point $x^{(q)}$ lies in the box around the point $x^{(p)}$ with half widths L_1/a_i, and
2. the point $x^{(r)}$ lies in the box around the point $x^{(p)}$ with half widths L_2/a_i, but not in the box around $x^{(p)}$ with half widths L_1/a_i as $x^{(q)}$, where $L_1 < L_2$.

From this natural interpretation, we can define distance from a point $x^{(p)} = (x_1^{(p)},..., x_n^{(p)})$ to another point $x^{(q)} = (x_1^{(q)}, \dots, x_n^{(q)})$ as

$$d(x^{(p)}, x^{(q)}) = \max_{1\le i\le n} \frac{\left|x_i^{(p)} - x_i^{(q)}\right|}{\frac{1}{a_i}} = \max_{1\le i\le n}\left(a_i \cdot \left|x_i^{(p)} - x_i^{(q)}\right|\right)$$

where $a_i = |\partial C/\partial x_i|$. Note that under this definition, $d(x^{(p)},x^{(q)}) \neq d(x^{(q)},x^{(p)})$, so d is not a true mathematical metric. The point as defined in the first parameter is the point around which the anonymization box is constructed.

Having defined this distance $d(x^{(p)}, x^{(q)})$, we now can construct an anonymization box in the multi-dimensional case. We select $k - 1$ points around the point $x^{(p)}$ closest to it under the distance d. Then the anonymization box is the one with smallest possible half widths $\Delta_i = L/a_i$ that contains these k points (including the point $x^{(p)}$ itself).

There is another possible enhancement to the algorithm that lets the anonymization box be even smaller. Let us denote these k points in the box including $x^{(p)}$ as $x^{(p,1)},\dots, x^{(p,k)}$. Then in each dimension i, we can compute

$$\left[\min_{1\le q\le k} x_i^{(p,q)},\ \max_{1\le q\le k} x_i^{(p,q)}\right]$$

as the lower and upper bounds of the anonymization box in this dimension. Let us denote the anonymization box around the point $x^{(p)}$ as $B(x^{(p)})$. After selecting $k - 1$ closest points around point $x^{(p)}$ and determining $B(x^{(p)})$, we can anonymize these k points including $x^{(p)}$ by assigning $B(x^{(p)})$ as their common value. However, we do not construct an anonymization box for every point, i.e., we do not calculate anonymization boxes $B(x^{(1)}),\dots, B(x^{(N)})$ for all points $x^{(1)},\dots, x^{(N)}$. Instead, we anonymize a point $x^{(p)}$ only if neither it nor any of the $k - 1$ closest points around it have been anonymized yet, that is, they are not contained in a previously calculated anonymization box $B(x^{(q)})$ around some other point $x^{(q)}$.

After completing this anonymization process, there are likely to be some points that have not yet been anonymized in any calculated box. Such points can be merged with existing anonymization boxes. For example, an unanonymized point $x^{(q)}$ can be merged with the calculated anonymization box $B(x^{(p)})$ with the smallest distance $d(x^{(p)}, x^{(q)})$. When merging $x^{(q)}$ with a $B(x^{(p)})$, in each dimension i, we can select

$$\left[\min\left(x_i^{(q)}, \underline{B(x^{(p)})}_i\right), \max\left(x_i^{(q)}, \overline{B(x^{(p)})}_i\right)\right]$$

as the lower and upper bounds of the merged anonymization box, which can be denoted as $B_{\text{new}}(x^{(p)})$, where $\underline{B(x^{(p)})}_i$ and $\overline{B(x^{(p)})}_i$ as the lower and upper bounds of the original anonymization box $B(x^{(p)})$ in this dimension. We then assign $B_{\text{new}}(x^{(p)})$ as the common value of $x^{(q)}$ and all points which have previously assigned the value $B(x^{(p)})$. At this point, $B(x^{(p)})$ merges with $x^{(q)}$, and the box is updated to $B_{\text{new}}(x^{(p)})$. This process is continued until all unanonymized points have been merged with existing anonymization boxes.

The time complexity of this algorithm is $O(N^2)$ where N is the number of point, which suggests that it should be practically applicable to moderately large data sets.

6. Software Implementation

All privacy-constrained generalization-based algorithms try to find optimal anonymization solutions which maximize utility with respect to some cost metric while meeting the requirements for privacy, but they vary widely in complexity. We developed new anonymization algorithms with improved control on data utility based on the privacy-constrained approach by using our new statistics-related utility cost control. The new privacy-constrained anonymization algorithms were developed according to statistics-related utility cost metrics as widths of computed ranges of the statistics mean, variance, covariance, and correlation.. These algorithms were theoretically proven to generate datasets meeting both the privacy requirement (as constrained by a *k*-anonymity model) and user-specified expectation of statistical usefulness (Xiang and Kreinovich 2013; Xiang et al. 2013). Work is underway to explore the scalability of the algorithms for large and very large data sets which are common in health care.

The anonymization software is associated with ancillary software tools for computing basic statistics for interval(ized) data sets (see reviews in Manski 2003; Ferson et al. 2007; Billard and Diday 2007, Kreinovich and Xiang 2008; Nguyen et al. 2012). The software implements algorithms for computing several location statistics such as the median and arithmetic, geometric and harmonic means and endpoints of confidence intervals on the mean. It also supports various dispersion statistics such as variance, standard deviation, coefficient of variation, and interquartile range, as well as other descriptive statistics such as skewness, quantiles, exceedance risks, and the empirical distribution function. These statistics have the form of intervals (or of a p-box in the case of the empirical distribution function). When the interval data exhibit no nesting, as is the case when they are produced by a one-dimensional anonymization, these algorithms are especially efficient, but they can also be applied to other interval data configurations (see Ferson et al. 2007), including the general case in which the interval data have no particular structure. These algorithms allow researchers who receive the anonymized data sets to extract the statistical information still present after individual privacy has been protected.

Figure 2 shows the results of applying the software to an example synthetic data set. The circles indicate the actual magnitudes of 60 data values of the continuous variable systolic blood pressure along the horizontal axis. The line segments are the corresponding anonymization intervals produced by the software using different strengths of *k*-anonymity. For graphical clarity, the circle-interval pairs are displaced vertically. Note that, for small *k*, not all of the data values have been changed. This is because they already satisfy the constraint of having the same value as at least *k*–1 other records. The output intervals have the no-nesting property so none is totally inside another. This makes the calculation of various statistical characteristics of the intervalized data much simpler computationally.

The true mean of the systolic blood pressure is 139.3. When *k* = 3, the anonymized data reveals it is in the interval [138.1 140.3]. When *k* = 4, this interval for the mean widens to [135.9, 142.2]. When *k* = 9, we would only be able to say the mean is somewhere in the range [130.5, 150.1]. A similar loss of information accompanies the calculation of the variance. The keeper of the original data knows that the true variance is 1085 to four digits. When *k* is 3, 4, and 9, the anonymization yields intervals for the variance that expand to [1011, 1160], [890, 1318] and [544, 1595] respectively.

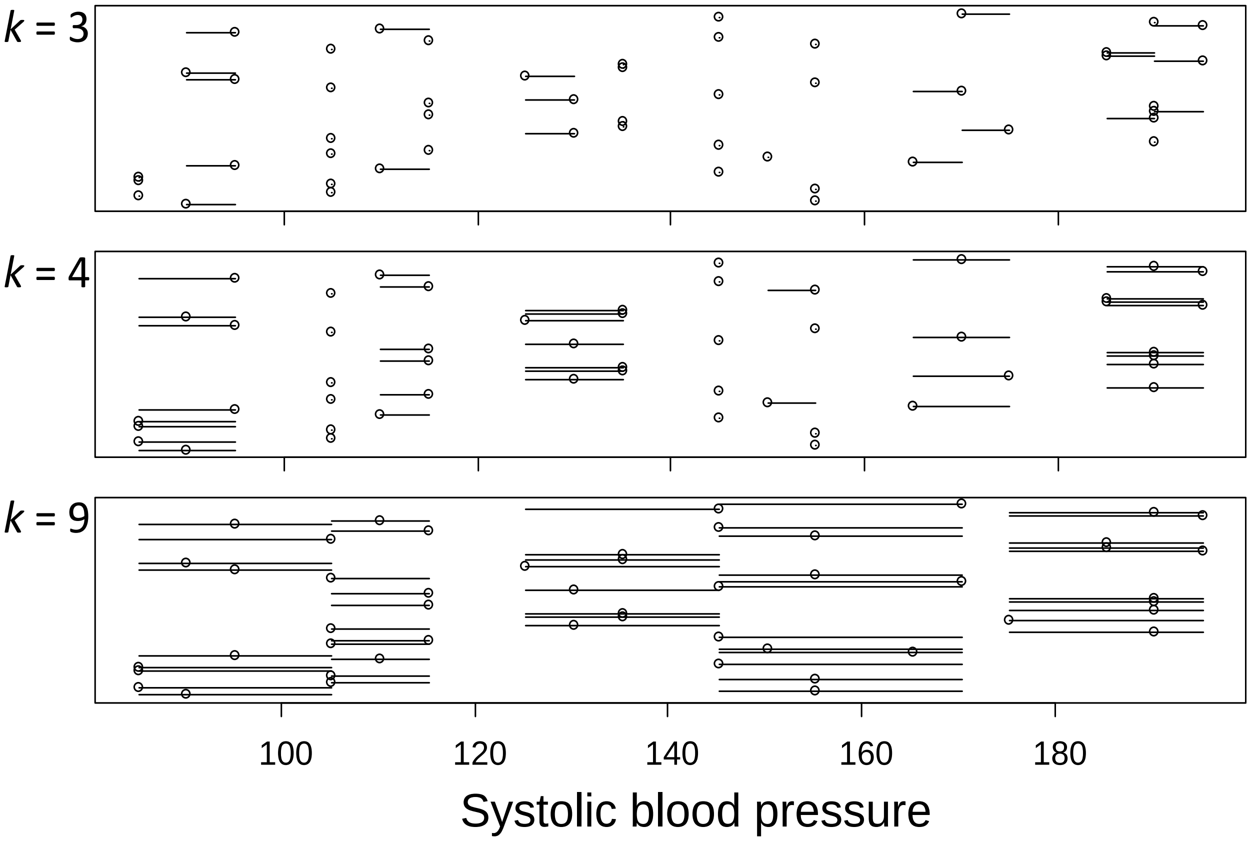

Figure 2. Example anonymization intervals (line segments) compared to original data (circles) for different *k*-anonymities.

7. Conclusions and Future Directions

Controlling disclosure risk and maintaining inferential power are two critical prerequisites to the free distribution of electronic health records among researchers, but more privacy protection usually means less information is released. Previously available data-release techniques do not sufficiently meet the need for balancing disclosure risk with inferential power, but focus almost exclusively on disclosure risk. In statistical disclosure control techniques, information truthfulness is not well preserved so that these methods may release unreliable results. In generalization-based anonymization approaches, the released data are guaranteed to be reliable, but there is still information loss due to attribute generalization and data utility is negatively affected. There have been few attempts to develop enhanced generalization-based anonymization techniques with sufficient and practical control on inferential power during the anonymization procedures.

We are developing a software tool that meets this need for balancing disclosure risk with inferential power in the release of health care data by implementing new generalization-based anonymization procedures with improved control of data utility. Statistics-related utility cost metrics have been adapted in the control processes. These utility cost metrics directly reflect data utility and are quite understandable to data consumers, therefore, researchers will be able to easily specify their expectations of the inferential power of the released data. Data holders will be able to use the software to release anonymized data to meet both disclosure protection requirements and data consumers' expectations of inferential power. The software includes efficient algorithms for computing interval statistics, which health care researchers can use to conduct statistical analyses on released intervalized data generated by the anonymization procedures.

For some generalization-based disclosure protection models, finding optimal solutions with respect to some other cost metrics is in general NP-hard, and the same situation would be expected for statistics-related utility cost metrics. Advanced algorithm design techniques are used to develop practically useful algorithms with reasonable computational complexities—even if theoretically they might not be practical for all cases in general.

The algorithms developed can be extended to more complex generalization-based disclosure protection models and a wider range of statistics-related utility cost metrics. In future efforts, we plan to implement methods for other statistical calculations not yet available, including various types of bivariate regressions and other correlation formulations, and two-sample comparisons. Most of these analyses have been considered by Billard and Diday (2007), but their equidistribution hypothesis, which assumes each interval can be modeled by a uniform distribution, is incompatible with our approach so further algorithm development for the set-theoretic version of interval statistics (Manski 2003; Ferson et al. 2007) will be required. We also plan to develop anonymization procedures based on utility constraints. The goal is to create a software library of novel and pre-existing anonymization methods to empower keepers of medical data with multiple methods by which they can release informatively optimal anonymized data in a way that protects patient privacy.

The methods described here are focused on anonymizing continuous data, but categorical patient information also contains information that can be used for re-identification. An extreme example of patient re-identification was recently demonstrated by Gymrek et. al. (2013) using genetic information to uniquely identify research participants in genome-scale human sequencing studies. Methods for mitigation and re-identification have been reviewed by Erlich and Narayanan (2014), though their treatment of *k*-anonymity in this context is admittedly incomplete. The anonymization methods should be generalized for categorical data.

Acknowledgments

We thank Lev Ginzburg of Stony Brook University, and Jody Sachs and Olga Brazhnik of the National Institutes of Health for discussions and encouragement. Funding was provided by the National Institutes of Health, through a Small Business Innovation Research grant (award numbers 1R43TR000173-01 and 5R43TR000173-02) to Applied Biomathematics. The opinions expressed herein are solely those of the authors.

References

Bayardo, R.J. and R. Agrawal 2005. Data privacy through optimal *k*-anonymization. *Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*. April 5-8, Tokyo.

Bertrand, P. and F. Groupil. 2000. Descriptive statistics for symbolic data. Pages 107-124 in *Analysis of Symbolic Data*, H.-H. Bock and E. Diday (eds.), Springer, Berlin.

Billard and Diday 2007. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley.

Erlich, Y., and A. Narayanan 2014. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, http://www.nature.com/nrg/journal/vaop/ncurrent/pdf/nrg3723.pdf.

Ferson, S., L. Ginzburg, V. Kreinovich, L. Longpré and M. Aviles. 2002. Computing variance for interval data is NP-hard. *ACM SIGACT News* 33(2): 108-118.

Ferson, S., V. Kreinovich, J. Hajagos, W.L. Oberkampf and L. Ginzburg 2007. *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. SAND2007-0939, Sandia National Laboratories, Albuquerque, New Mexico. http://www.ramas.com/intstats.pdf .

Gal, T.S., Z. Chen, and A. Gangopadhyay 2008. A privacy protection model for patient data with multiple sensitive attributes. *International Journal of Information Security and Privacy* 2(3): 28-44.

Ghinita, G., P. Karras, P. Kalnis, and N. Mamoulis 2007. Fast data anonymization with low information loss. *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)*, September 23-27, Vienna, Austria.

Ghinita, G., P. Karras, P. Kalnis, and N. Mamoulis 2009. A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Transactions on Database Systems (TODS)* 34(2).

Gionis A. and T. Tassa 2009. *k*-anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering* 21(2): 206-219.

Golle, P. 2006. Revisiting the uniqueness of simple demographics in the US population. *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, October 30, Alexandria, VA.

Gymrek, M., A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich 2013. Identifying personal genomes by surname inference. *Science* 339: 321-324.

Iyengar, V 2002. Transforming data to satisfy privacy constraints. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02),* July 2002, Edmonton, Alberta, Canada. Pages 279–288.

Kreinovich, V. and G. Xiang 2008. Fast algorithms for computing statistics under interval uncertainty: an overview. *Interval/Probabilistic Uncertainty and Non-Classical Logics*, Springer-Verlag, Berlin, pp. 19-31.

Kreinovich,V., G. Xiang, S.A. Starks, L. Longpre, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos 2006. Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity. *Reliable Computing* 12(6): 471-501.

Kreinovich, V., L. Longpre, S.A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos 2007. Interval versions of statistical techniques, with applications to environmental analysis, bioinformatics, and privacy in statistical databases. *Journal of Computational and Applied Mathematics* 199(2): 418-423.

LeFevre, K., D. DeWitt, and R. Ramakrishnan 2005. Incognito: efficient full-domain *k*-anonymity. *Proceedings of ACM Conference on Management of Data (SIGMOD'05)*, June 13-16, Baltimore.

Li, N., T. Li, and S. Venkatasubramanian 2007. *t*-Closeness: privacy beyond *k*-anonymity and *l*-diversity. *Proceedings of the 23rd International Conference on Data Engineering (ICDE'07).* April 17-20, Istanbul.

Li, Z. and X. Ye 2007. Privacy protection on multiple sensitive attributes. *Information and Communications Security,* Lecture Notes in Computer Science 4861, S. Qing et al. (eds.), Springer, Berlin. Pages 141-152.

Machanavajjhala, A., J. Gehrke, D. Kifer, and M. Venkitasubramaniam 2006. *l*-diversity: privacy beyond *k*-anonymity. *Proc. 22nd International Conference on Data Engineering (ICDE'06),* April 3-7, Atlanta, Georgia.

Martin, D. J., D. Kifer, A. Machanavajjhala, and J. Gehrke 2007. Worst-case background knowledge for privacy-preserving data publishing. *Proceedings of the 23rd International Conference on Data Engineering, (ICDE'07)*, April 15-20, Istanbul, Turkey. Pages 126–135.

Meyerson, A. and R. Williams 2004. On the complexity of optimal *k*-anonymity. *Proceedings of ACM Conference on Principles of Database Systems (PODS'04)*. June 14-18, Paris, France.

Miller, J., A. Campan, T. M. Truta 2008. Constrained *k*-Anonymity: privacy with generalization boundaries. *Proceedings of 8th SIAM International Conference on Data Mining (SDM'08).*

Nergiz, M. E. and C. Clifton 2006. Thoughts on *k*-anonymization. *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW '06),* Washington, DC, USA. Page 96.

Nguyen, T.H., V. Kreinovich, B. Wu, and G. Xiang. 2012. *Computing Statistics under Interval Uncertainty: Applications to Computer Science and Engineering*, Springer-Verlag.

Samarati, P. 2001. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 13(6): 1010–1027.

Sweeney, L. 2002a. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5):557–570.

Sweeney, L. 2002b. Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5): 571–588.

Truta, T.M. and B. Vinay 2006. Privacy protection: *p*-sensitive *k*-anonymity property. *Proceedings of the 22nd International Conference on Data Engineering Workshops, the Second Intenational Workshop on Privacy Data Management (PDM'06)*, April 8, Atlanta, Georgia. Page 94.

Willenborg, L. and T. De Waal 1996. *Statistical Disclosure Control in Practice*, Springer-Verlag.

Wong, R. C. W., J. Li, A. W. C. Fu, and J. Pei 2007. Minimality attack in privacy-preserving data publishing. *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)*, September 23-27, Vienna, Austria. Pages 543–554.

Wong, R. C. W., J. Li, A. W. C. Fu, and K. Wang 2006. (α, *k*)-Anonymity: an enhanced *k*-anonymity model for privacy-preserving data publishing. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, August 20-23, Philadelphia, Pennsylvania. Pages 754–759.

Xiang, G. and V. Kreinovich 2012. Revisiting the uniqueness of simple demographics in the US population. *Proceedings of the IEEE Symposium on Computational Intelligence for Engineering Solutions CIES'2013*, Singapore. Pages 163-170.

Xiang, G., and V. Kreinovich 2013. Data anonymization that leads to the most accurate estimates of statistical characteristics. *Proceedings of the IEEE Series of Symposia on Computational Intelligence SSCI'2013*, Singapore.

Xiang, G., M. Ceberio, and V. Kreinovich 2007. Computing population variance and entropy under interval uncertainty: linear-time algorithms. *Reliable Computing* 13(6): 467-488.

Xiang, G., S. Ferson, L. Ginzburg, L. Longpré, E. Mayorga, and O. Kosheleva 2013. Data anonymization that leads to the most accurate estimates of statistical characteristics: Fuzzy-motivated approach. In *IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint* IEEE. Pages 611–616.

Xiao, X. and Y. Tao 2006. Personalized privacy preservation. *Proceedings of ACM Conference on Management of Data (SIGMOD'06)*, June 26-29, Chicago, Illinois. Pages 229–240.

Zhang, Q., N. Koudas, D. Srivastava, and T. Yu 2007. Aggregate Query Answering on Anonymized Tables. *Proceedings of the 23rd International Conference on Data Engineering, (ICDE'07)*, April 15-20, Istanbul, Turkey. Pages 116–125.

Zhong, S., Z. Yang, R.N. Wright 2005. Privacy-enhancing *k*-anonymization of customer data, *Proc. 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'05)*, June 13-15, Baltimore, MD.