R-U policy frontiers for health data de-identification

RECEIVED 14 July 2014
REVISED 27 December 2014
ACCEPTED 9 January 2015
PUBLISHED ONLINE FIRST 24 April 2015

Weiyi Xia[1], Raymond Heatherly[2], Xiaofeng Ding[3], Jiuyong Li[4], Bradley A Malin[1,2]

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD UNIVERSITY PRESS

ABSTRACT

Objective The Health Insurance Portability and Accountability Act Privacy Rule enables healthcare organizations to share de-identified data via two routes. They can either 1) show re-identification risk is small (e.g., via a formal model, such as *k*-anonymity) with respect to an anticipated recipient or 2) apply a rule-based policy (i.e., Safe Harbor) that enumerates attributes to be altered (e.g., dates to years). The latter is often invoked because it is interpretable, but it fails to tailor protections to the capabilities of the recipient. The paper shows rule-based policies can be mapped to a utility (U) and re-identification risk (R) space, which can be searched for a collection, or frontier, of policies that systematically trade off between these goals.

Methods We extend an algorithm to efficiently compose an R-U frontier using a lattice of policy options. Risk is proportional to the number of patients to which a record corresponds, while utility is proportional to similarity of the original and de-identified distribution. We allow our method to search 20 000 rule-based policies (out of 2^{700}) and compare the resulting frontier with *k*-anonymous solutions and Safe Harbor using the demographics of 10 U.S. states.

Results The results demonstrate the rule-based frontier 1) consists, on average, of 5000 policies, 2% of which enable better utility with less risk than Safe Harbor and 2) the policies cover a broader spectrum of utility and risk than *k*-anonymity frontiers.

Conclusions R-U frontiers of de-identification policies can be discovered efficiently, allowing healthcare organizations to tailor protections to anticipated needs and trustworthiness of recipients.

Keywords: privacy, de-identification, secondary use, policy, optimization

RESEARCH AND APPLICATIONS

INTRODUCTION

In the age of big data, healthcare organizations will accumulate a substantial quantity of detailed personal data.[1,2] These large-scale resources can support the development of novel healthcare applications and innovative services.[3,4] For instance, the data in electronic medical record systems can enable predictive modeling,[5,6] novel association studies,[7–9] as well as the discovery of personalized treatment regimens.[10,11] At the same time, there are many requests to share patient-level data to support information reuse (e.g.,[12]), learning health systems (e.g.,[13]), transparency (e.g.,[14,15]), and to adhere to federal grant policies (e.g.,[16]). In other words, there is a belief that sharing data derived from the clinical domain provides utility (U) to society. While publication can enable broad access, there is a risk (R) of violating the privacy rights of the patients to whom the data corresponds.[17–20]

The myriad definitions of privacy[21] are evolving in the context of big data,[22,23] and it should be clear that the goal of this paper is not to suggest which is best, but rather to provide technical mechanisms to support a certain definition that continues to receive a substantial amount of attention. Specifically, various laws state that data are sufficiently protected when it is "difficult" to ascertain an individual's identity.[24] For example, the European Union's Data Protection Directive refers to such data as "anonymized"[25] and the U.S. Health Insurance Portability and Accountability Act (HIPAA) calls that data "de-identified"[26] (the convention we use henceforth). In so doing, these laws aim to prevent *identity disclosure*, which transpires when a recipient of the data links it with some resource containing explicit identifiers (e.g., a voter registration list[27,28]).

To achieve de-identification, laws often provide publishers with several options. First, they may invoke a set of rules, or a policy, to transform data into a de-identified state. The Safe Harbor model defined by HIPAA is a clear example of such a policy, which specifies 18 rules, including suppression of explicit identifiers (e.g., personal names) and generalization of "quasi-identifiers" which could enable linkage (e.g., dates of events, such as birth, are replaced with time periods no more specific than one year and ages over 89 years-old are recoded as 90+). Yet, the rigidity of such rule-based policies is not ideal for sharing every data set, such as studies with the elderly (e.g., dementia patients).[29]

Thus, the law enables publishers to use an alternative, which permits data to be shared in any format, provided the risk of re-identification is appropriately measured and mitigated. Various formal anonymization models (i.e., *k*-anonymity[30]) have been proposed to ensure that a certain mathematical property of the dataset holds true. However, these models tend to be overly rigid for a number of practical purposes. Notably, HIPAA states that publishers can ensure the "risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."[26] This implies that risk is proportional to the trustworthiness of the recipient and suggests there are many different policies that could be invoked to de-identify the data. For example, if health data are published on the Internet, via a Centers for Medicare and Medicaid Services dataset, the threat is high because the recipients are unknown and the system is completely open. As such, a data manager could select a policy that heavily favors risk mitigation over utility. By contrast, if health data are published to a more

Correspondence to Weiyi Xia 2525 West End Avenue, Suite 1030 Department of Biomedical Informatics Nashville, TN 37205, USA; weiyi.xia@vanderbilt.edu Tel: 615 887 4798

 For numbered affiliations see end of article.

trusted party (e.g., a health services researcher with strong information security practices who agrees to sign a data use agreement), then the threat is much lower and one could apply a policy that favors utility over risk.

This paper shows that an efficient and effective mechanism can be applied to discover rule-based de-identification policy alternatives for patient-level datasets. To do so, we extend an algorithm[31] designed to search a collection of de-identification policies that compose a frontier that optimally balances risk (R) and utility (U). We show this approach allows for guidance, interpretation, and justification of rule-based policies, as opposed to relying on a predefined standard in terms of the re-identification risk and data utility or formal models. As a concrete example, Figure 1 depicts how a record from a dataset investigated in our experimental analysis is transformed by one of the frontier policies in comparison with its Safe Harbor and 10-anonymous (i.e., the record is part of a group of no less than 10 records with the same values) versions. In this example, R is defined as inversely proportional to the size of this demographic group defined by the record in a population set, while U is in terms of an information loss metric which represents the discrepancy between the probability density of the record in the original dataset and the transformed dataset. Safe Harbor transforms this record into a group with a large set of ZIP code areas, while 10-anonymization and a policy on the R-U frontier transform it into two different age groups and small ZIP code groups. Based on the population size in these different groups, the record transformed via the frontier policy has slightly higher risk than its 10-anonymous counterpart, while Safe Harbor has the highest risk. On the other hand, the record in the dataset transformed via a frontier policy has lower information loss than its counterparts of both Safe Harbor and 10-anonymous.

To demonstrate the effectiveness of this approach, we apply it to demographic data from the U.S. Census Bureau (i.e., the "Adult" dataset)[32] in combination with geographic information from nine states with medical facilities involved in the Electronic Medical Records and Genomics (eMERGE)[33] network, as well as the state of Hawaii, which has a unique demographic distribution. The results show the de-identification frontier can recommend policies with less risk and more utility than Safe Harbor and cover a broader spectrum of utility and risk than formal protection models in the form of *k*-anonymity.[30]

BACKGROUND AND RELATED RESEARCH

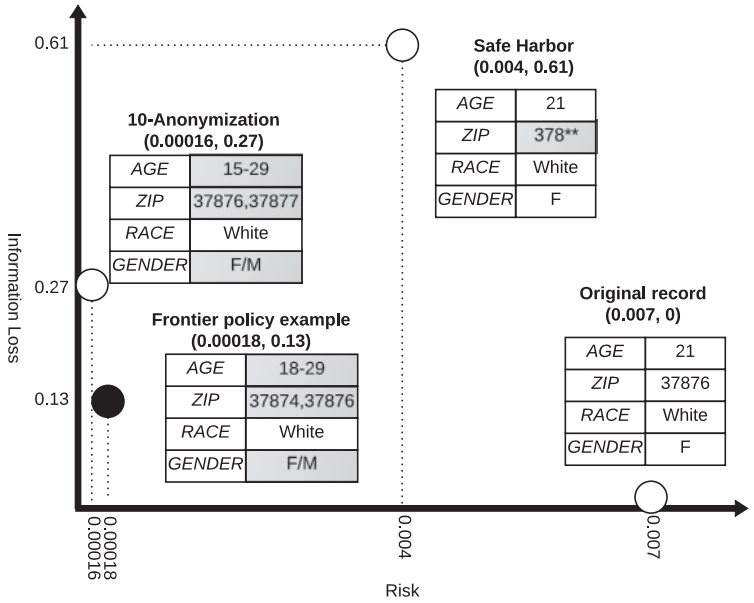
A re-identification occurs when a data recipient matches a published record with the identifiers (e.g., personal name) of the corresponding individual. This is often accomplished by a quasi-identifier (e.g., gender, date of birth, and residential 5-digit ZIP code).[34] While various methods have been developed for mitigating re-identification risk while preserving data utility in the area of privacy-preserving data publishing,[35] in this section, we focus on two topics in this area which are highly related to our work: 1) disclosure control, with particular attention to generalization and suppression strategies and 2) R-U frontier analysis.

Disclosure Control

De-identification policies are rules that guide the processing of the quasi-identifiers to reduce re-identification risk. Many operations can be applied to process quasi-identifiers in medical records, such as generalization, suppression, and randomization.[36] Safe Harbor focuses on a certain set of rules for generalization and, thus, to directly relate our method, we represent the policy space as the set of possible generalizations. Focusing on the protection of only quasi-identifiers has certain vulnerabilities (e.g., homogeneity attacks where most, or all, patients have the same diagnosis,[37,38]), but it is often used in practice.[39,40,41]

Generalization has been invoked by many formal privacy protection models, such as *k*-anonymity,[30] which itself is a specific case of de-identification (where each record must to be equivalent to least $k-1$ other records over the set of quasi-identifiers). As a result, a

Figure 1: De-identification of a record in the Adult-TN dataset and the corresponding risk and utility (in terms of information loss). De-identification is performed according to Safe Harbor, 10-anonymization, and a de-identification policy on the R-U frontier.



significant number of variations on the generalization space have been proposed.[42] In this work, we focus on two in particular because they have the strongest relationship with solutions deemed to be useful by the medical research community.[40] The first is *full-domain generalization*[43] which relies upon a domain generalization hierarchy (DGH) for each quasi-identifier (e.g., age intervals of 1 year, then 5 years, then 10 years, etc.) where the policy space is the cross-product of the level set of each DGH. The second is *full-subtree generalization*,[44] which extends the policy space to allow for arbitrary partitions of a quasi-identifier's domain. It should be recognized that such a space is large and searching for the optimal solution (even for the restricted case of *k*-anonymity) is NP-hard.[45] As such, heuristic-driven strategies,[46] including genetic algorithms[44] and cost-bounding on subsections of the lattice,[47] have been invoked.

R-U Frontier of Data Publishing

The notion of an R-U analysis was proposed to support disclosure control decision making. First, the R-U confidentiality map[48] was introduced to characterize the tradeoff between the level of perturbation (e.g., randomization) applied to a dataset and its predictive accuracy. This concept was then utilized to design algorithms to search for a frontier of *k*-anonymization solutions with minimum utility loss over a range of *k*.[49,50] The frontier method has further been extended for set-valued data anonymization, where a record contains more than one value per attribute; for example, a set of diagnosis codes.[51]

The policy space was structured as a lattice of binary strings.[44,52] However, the algorithm proposed to search the space considered utility only from a syntactic (e.g., level in the DGH), as opposed to a semantic (e.g., the difference between the distribution of the original and de-identified datasets), perspective. Moreover, the algorithm only searched for policies with risk no worse than HIPAA Safe Harbor and, thus, does not guarantee composition of an R-U frontier. This problem is extended into R-U frontier discovery by introducing a semantic utility loss measure which captures the amount of distortion introduced to the dataset by the de-identification process, and proposed a heuristic algorithm based on a random walk to search the policy lattice.[31] Yet, as the experiments in the Supplementary Appendix show, the random process is limited in its ability to compose a high-performing frontier in an efficient manner.

METHODS

De-identification Policy Frontier Search Framework

The Sublattice Heuristic Search (SHS) algorithm[31] was introduced to search a lattice of policies for a frontier in the R-U space. Formally, a frontier is a set of policies that are not strictly dominated by other policies. Intuitively, a policy p_A strictly dominates a policy p_B when both risk and utility loss values of p_A are no greater than the corresponding values of p_B and at least one value is strictly less that of p_B.

SHS is composed mainly of two functions: i) initialize and ii) improve. Figure 2 provides an illustration of how SHS searches for a frontier. Basic versions for these functions in the SHS algorithm relied heavily on randomized processes. For this investigation, we devised functions that used more intelligent heuristics based on patient demographics for frontier composition. Here, we briefly review the basic functions and the new heuristics with the full details in the Supplementary Appendix online. We report on an empirical comparison to prove the improved performance of the heuristics in the results.

SHS Initialization: In the basic algorithm, the frontier is initialized by selecting a random path from the most general to the most specific policy in the lattice. An example of random path initialization (or *RandPath*) is illustrated in Figure 2 (a), where p_1, p_2, p_3, p_4, and p_5 are selected. As an extension, we developed the full domain initialization (*FullDom*) strategy. *FullDom* samples a full-domain generalization space to compose the initial frontier and refines it through interpolation. A balanced DGH for each attribute is built in a top-down fashion. Starting from the most generalized value (i.e., generalize every value to the entire domain; e.g., age as [0–120]), in each level, the construction process splits each generalized interval into two equal size intervals. For example, if a level in the age DGH generalizes to 10-year bins, then its immediate children generalize into 5-year bins. This policy space thus corresponds to a cross-product of the level set of the DGH of each attribute. It covers a more diverse set of policies than a random path because i) the full-domain space contains policies from different paths while the random path does not and ii) the full-domain space tends to sample a policy that is multiple levels away from the previous one while a random path samples a child of a previous one.

SHS Improvement: Following initialization, SHS iteratively selects a *sublattice* to improve the frontier. A sublattice is a subgraph in the lattice containing all policies between a *top* and a *bottom* policy (where the latter is a descendent of the former). Figure 2 (b) provides an illustration of such a structure. Assuming the re-identification risk and utility functions are monotonic over the order defined by the policy lattice, the R-U mappings of policies in a sublattice are contained within a bounding region in the R-U space as the rectangle in Figure 2 (b). The basic algorithm relied on a random sublattice generation (*RandSub*) strategy to iteratively select a sublattice. For each sublattice, *RandSub* computes the proportion of the bounding region not dominated by the frontier (Figure 2 (b), yellow area), to estimate the probability that a randomly selected policy could improve the frontier. When the computed value is below a predefined threshold, the algorithm skips this sublattice. Otherwise, it searches a random path in an effort to improve the frontier. As illustrated in Figure 2 (c), the newly searched policies (i.e., p_a, p_i, and p_b) will be on the new frontier, while policy p_4 in the old frontier is dominated and is removed from further consideration.

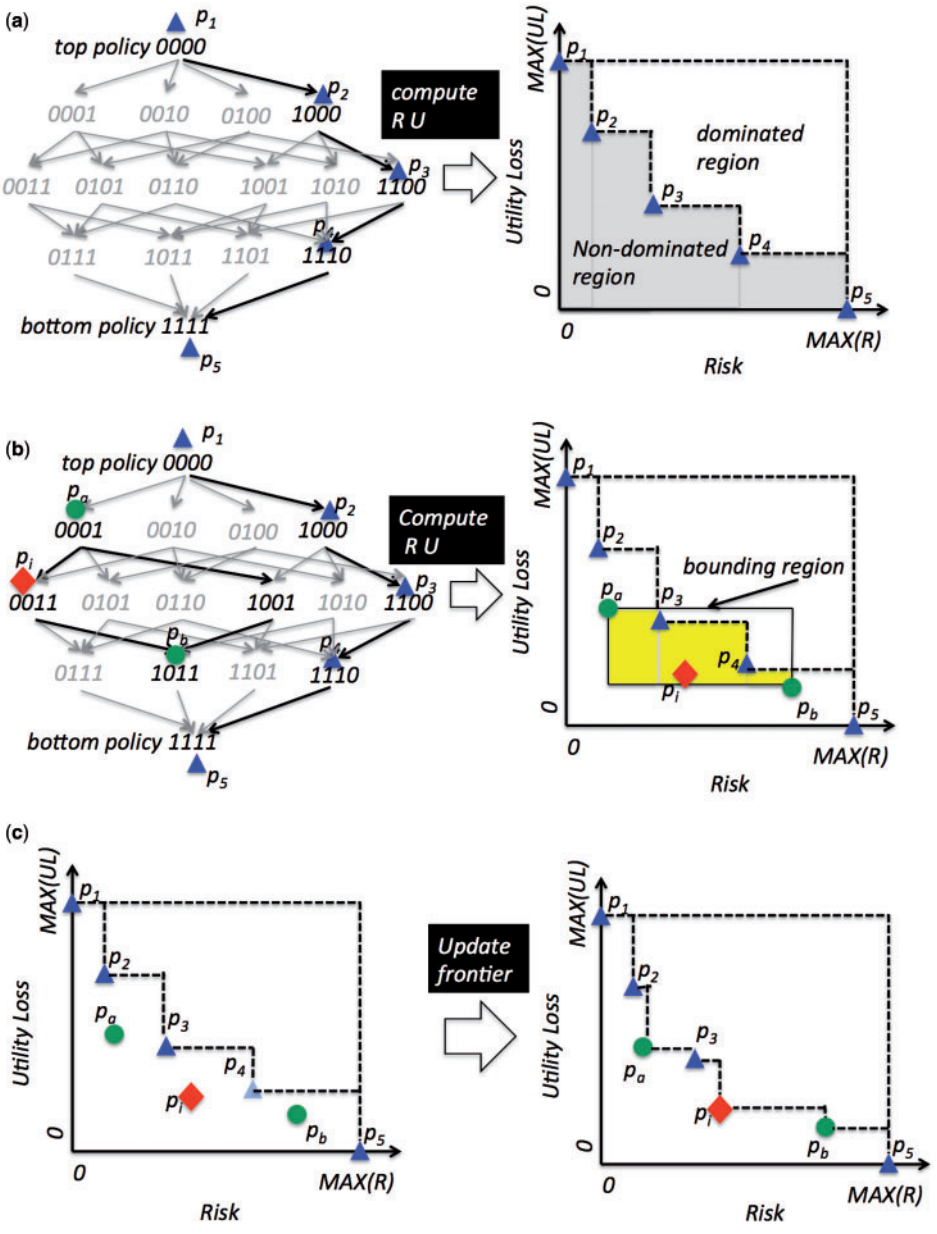
To mitigate randomization effects, we introduce a sublattice chain generation (*SubChain*) strategy for the SHS improvement function. *SubChain* generates a sequence of sublattices between the most generalized and the most specific policies of the lattice by using the bottom policy of the sublattice in the current iteration as the top policy of the sublattice in the following iteration. To limit convergence to local optima, *SubChain* randomly restarts from the most generalized policy when the chain reaches the most specific policy in the lattice. This strategy is based upon the expectation that when a sublattice updates the current frontier, the top and bottom policies are likely to be useful starting points for the subsequent sublattice.

RESULTS

Materials

We evaluated the R-U frontier discovery process using two publicly available datasets. The first is the Adult dataset,[32] which consists of 32 561 records without missing values. For comparison with Safe Harbor, we restrict the quasi-identifiers to the demographics of {Age, Gender, Race}. To enable a comparison with respect to geography, we combine the available demographics data from Adult with state-level demographic information obtained from the US Census Bureau's 2010 Census Tables PCT12A-G[53] to provide each tuple with a 5-digit ZIP code. To mitigate the bias that can be introduced through analysis over a single population, we simulated the Adult dataset for 10 US states: Illinois (IL), Hawaii (HI), Massachusetts (MA), Minnesota (MN),

Figure 2: An example of a policy lattice for five quasi-identifying values and an illustration of the main functions of the SHS algorithm. (a) depicts a random path initialization of the frontier composed of five policies (p_1, p_2, p_3, p_4, and p_5). (b) depicts a sublattice structure *sublattice* (p_a, p_b) in the policy space and its corresponding bounding region in the R-U space, the yellow area of which is not dominated by the frontier. The proportion of the yellow area is used as the heuristic of the probability that a randomly selected policy in the sublattice can update current frontier. (c) depicts the updates to the frontier: p_a, p_b, and p_i from the *sublattice* (p_a, p_b) will be on the frontier while policy p_4 is removed.



New York (NY), Ohio (OH), Pennsylvania (PA), Tennessee (TN), Washington (WA), and Wisconsin (WI). Following the strategy set forth in a previous study,[52] the Census data of the corresponding states are used as the population statistics to compute the re-identification risk of these synthesized datasets.

All of these states, with the exception of HI, correspond to regions that contain academic medical centers participating in the eMERGE network.[33] These centers are collecting and sharing de-identified data on patients to the public and are actively using the Safe Harbor de-identification policy, but are open to alternatives.[29] HI is selected as an additional state because of its unique demographic distribution (e.g., it has the highest percentage of Asians and the lowest percentage of whites in the United States).

To provide analysis on nonsynthetic data, we also conducted experiments on the North Carolina voter registration (NCVR) database,[54] which contains 6 150 562 records without missing values, each record

consists of 18 fields. For this study, we restricted the dataset to a set of four quasi-identifying attributes {Age, Race, Gender, 5-Digit ZIP Code}. We use the entire dataset as the population and randomly sample datasets to publish.

The policy lattice, based on the selected quasi-identifier, contains on the order of 2^{700} policies, which would take a significant length of time to search exhaustively.

Experimental Design

Our experiments were split into two primary sections. First, we evaluated the performance of the frontier initialization and improvement strategies by comparing the resulting frontiers after searching the same quantity of policies. We used the area under the frontier in the R-U space, denoted as AU, as the criteria of the frontier given the orientation of risk and utility loss. We reported the results after every 1000 policies while searching the first 5000 policies. Second, we compared the R-U tradeoffs made by the policies on the frontier discovered by the best SHS configuration to those of two related methods i) a rule-based single policy in the form of HIPAA Safe Harbor and ii) a popular *k*-anonymization algorithm. In this setting, we allowed SHS to search 20 000 policies.

The details of the generalization rules defined by Safe Harbor (which is a policy in the lattice) are provided in the Supplementary Appendix online. For *k*-anonymization, we utilized the Incognito[55] algorithm to find solutions on the frontier for $k = 5$ and 10, which are commonly adopted protection levels.[56] We utilized Incognito because it generates all *k*-anonymous solutions from which the set of nondominated solutions can be selected to create a frontier, which is not the case in many other methods and would lead to an incomplete frontier.

To compute risk, we adopted the *marketer* risk-based disclosure measure[52] using the Census data as population statistics, which is based on the distinguishability metric.[57] Informally, this means that the recipient of the data attacks every record in the dataset and risk is proportional to the average likelihood of successfully re-identifying a record to the population from which it was derived. We used the Kullback–Leibler divergence to measure the utility loss incurred by a generalized dataset with respect to its original form. Informally, the Kullback–Leibler divergence is the difference between the probability distributions of the quasi-identifying values in the original and de-identified datasets. The mathematical definitions for the risk and utility computation can be found in the Supplementary Appendix.

The algorithms were implemented in Python and all experiments were run on an Ubuntu server with 24 Intel(R) Xeon(R) CPUs at 2.4 GHz and 64 GB of RAM. Further details of the parameterization of the algorithms are documented in the Supplementary Appendix.

Performance of SHS strategies

We assessed the performance of SHS using the Adult dataset with TN ZIP codes and a randomly sampled set of 100 000 records from NCVR. We herein report on the average performance of the algorithms over 20 complete runs for each dataset.

Figure 3 provides snapshots of the R-U frontiers after initialization and the end of the improvement search. The snapshots correspond to a single run of each algorithm (though other 19 runs were similar). Figure 3 (a) and (b) show the *FullDom-RandSub* and *FullDom-SubChain* frontiers are closer to the origin than the *RandPath-RandSub* and *RandPath-SubChain* frontiers after the initialization process, which implies they have less risk and more utility.

Figure 4 (a) and (b) provides the average and standard deviation the AU of the initial frontiers using different strategies after visiting 1000 polices (i.e., initialization). For orientation, after initialization in the Adult dataset, the average AU of *FullDom-RandSub* and *FullDom-SubChain* is 0.127 and 0.124, respectively. By contrast, *RandPath-RandSub* and *RandPath-SubChain* exhibited an average AU of 0.172 and 0.195, respectively. The result strongly suggests that *FullDom* is the dominant initialization strategy.

To compare the frontier improvement strategies, Figure 3 (c) and (d) show samples of the frontiers after 5000 policies visited and the average AU of them is reported in Figure 4 (a) and (b). These results show that when the initialization approach is *FullDom*, *SubChain* converges to a better frontier than *RandSub*. For instance, for the NCVR final frontier, the average AU of *FullDom-SubChain* is 0.127, while *FullDom-RandSub* is 0.138. By contrast, when *RandPath* is the initialization process, *RandSub* outperforms *SubChain*. For instance, for the NCVR final frontier, the average is 0.171 for *RandPath-SubChain* and 0.152 for *RandPath-RandSub*. The result indicates that the *SubChain* strategy is more efficient in improving a frontier that is closer to the optimal frontier, since *FullDom* results in a significantly better initial frontier than the *RandPath* algorithm.

The running time of the algorithms is shown in Figure 4 (c) and (d). It should be noted that the running time is significantly longer for the NCVR dataset than the Adult dataset because the former is approximately 3 times larger in sample size. The result shows that, given the same search budget, initialization via *RandPath* is faster than *FullDom*. One possible reason for this is an artifact of our implementation. Our code was optimized to reduce repetition in the risk and utility computations.[31] And *RandPath* tends to search policies on the same path, whereas *FullDom* tends to have more distinct routes. However, even though *RandPath* initialization is faster than *FullDom*, Figure 4 (e) and (f) illustrates that *FullDom* still converges to a better frontier in less time.

These results suggest that *FullDom* initialization in combination with *SubChain* sublattice generalization is the best strategy for the SHS algorithm. As such, the remainder of our experiments uses the *FullDom-SubChain* procedure.

Frontier Case Studies

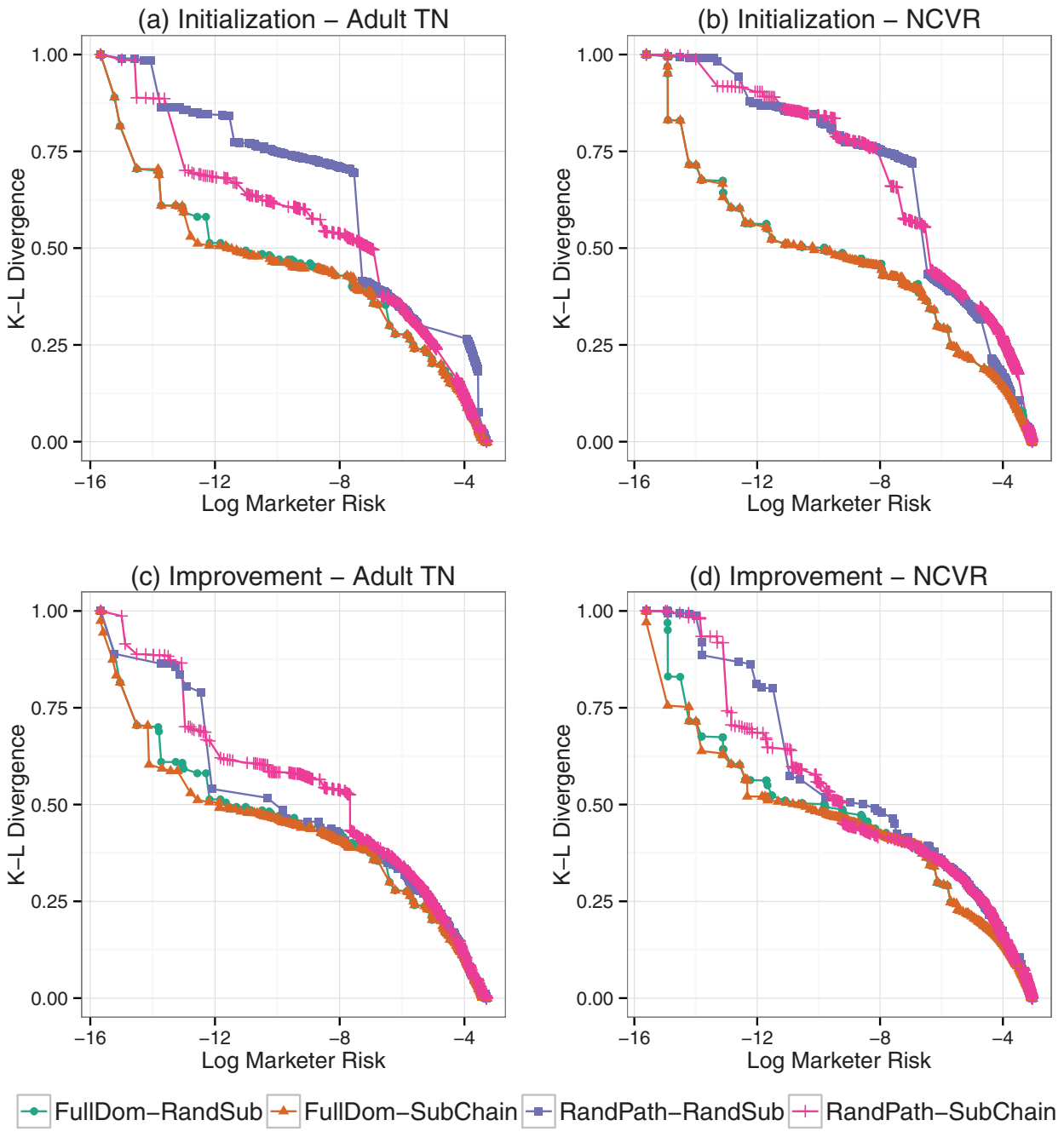
The frontier for SHS and 10-anonymization (the result for 5-anonymization (in the Supplementary Appendix)) is similar for Adult-TN is depicted in Figure 5 (a), while results for the other states are in Figure 6. Notably, the results indicate that a region of the frontier discovered by SHS dominates the Safe Harbor policy in all states. Moreover, the frontier region that dominates Safe Harbor results in both greater utility and risk than the results of 10-anonymization.

For illustration, two policies that dominate Safe Harbor (i.e., less risk and better utility) are highlighted in Figure 5. The discovered policies exhibit notable differences from Safe Harbor. For instance, both policies generalize race and ages below 90 to larger groups than Safe Harbor (as illustrated in Figure 5 (c)), but retain more specific geographic information (as illustrated in Figure 5 (d)). Additionally, the second policy generalizes gender to (Male or Female).

Policies on the Frontier

Table 1 reports the number of policies on each frontier. The SHS frontier contains an average of 4700 policies while the *k*-anonymity frontier contains an average of 33 and 26 policies when $k = 5$ and 10, respectively. This is because SHS can search a significantly larger space than the Incognito *k*-anonymization algorithm, due to the construction of their

Figure 3: Snapshots of the R-U frontiers discovered via a single run of the Sublattice Heuristic Search (SHS) algorithm. The initialized frontier for the (a) Adult-TN and (c) North Carolina Voter Registration (NCVR) datasets are based on 1000 policies. The improved frontier for (b) Adult-TN and (d) NCVR datasets are based on an additional 4000 policies.



respective lattices. Even though the number of policies in the SHS frontier is large, these policies are ordered by their R-U values, so a data publisher can quickly locate the policies they are interested in.

Policies Dominating Safe Harbor

The ratio of policies on the SHS and *k*-anonymization frontiers that dominate Safe Harbor is summarized in Table 2. Notice that the SHS frontier contains policies that dominate Safe Harbor in all states. By contrast, 5-anonymization leads to solutions that dominate Safe Harbor for only HI, TN, MN, WA, and WI, while 10-anonymization can only find dominant solution for HI. This is because *k*-anonymity datasets tend to have more utility loss than does a dataset de-identified through Safe Harbor.

Frontier Ranges

Table 3 summarizes the result of the comparison of ranges of the *k*-anonymity frontier and the SHS frontier. The results indicate *k*-

Figure 4: Effectiveness versus efficiency for the Sublattice Heuristic Search (SHS) algorithm. The average (±standard deviation) of the area under the frontier (AU) for the frontiers as a function of the number of policies visited is shown for (**a**) Adult-TN and (**b**) North Carolina Voter Registration (NCVR). The average (± standard deviation) of the runtime as a function of the number of policies visited is shown for (**c**) Adult-TN and (**d**) NCVR. The average (± standard deviation) of the AU for the frontiers as a function of the runtime is shown for (**e**) Adult-TN and (**f**) NCVR.

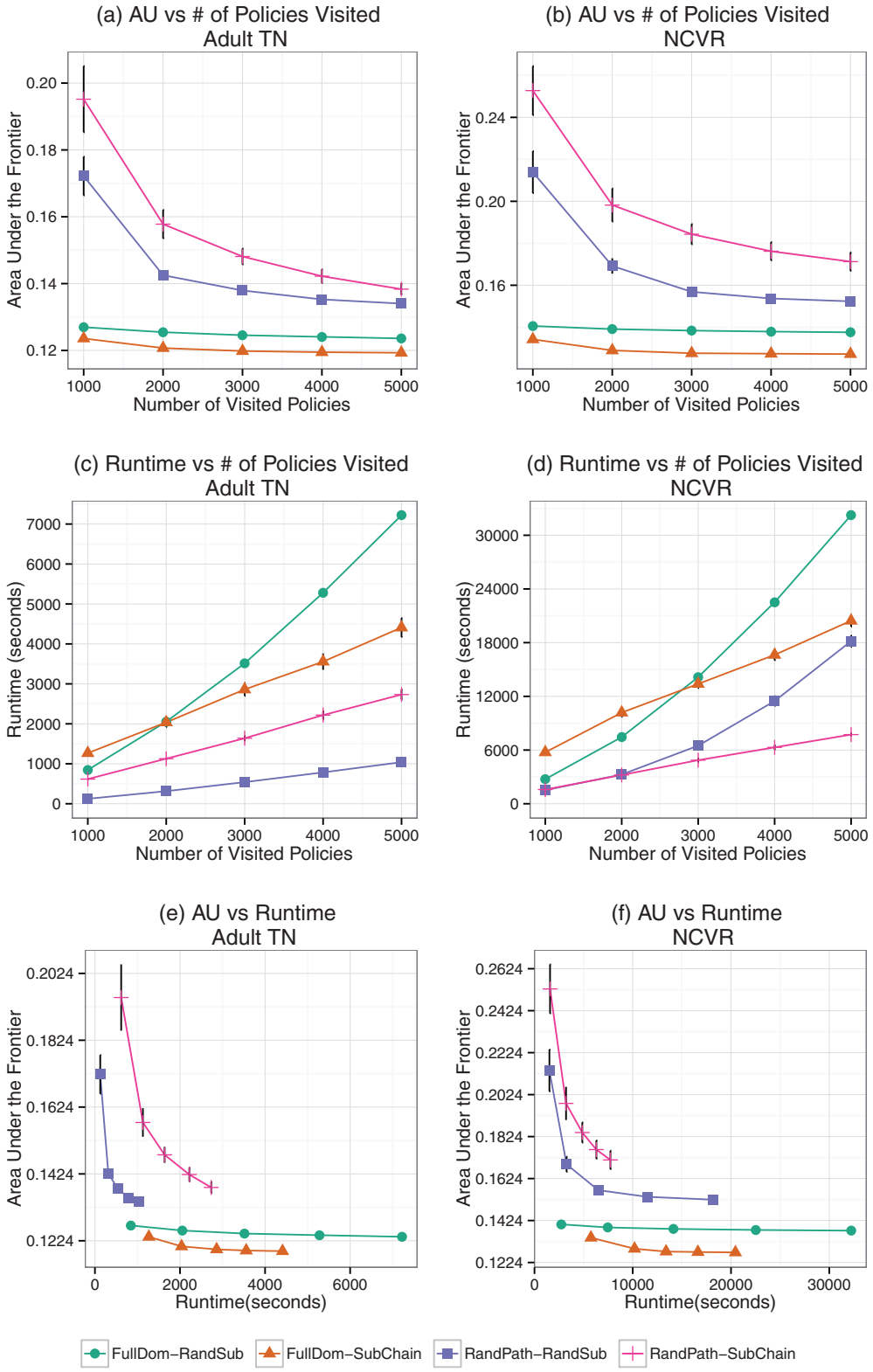
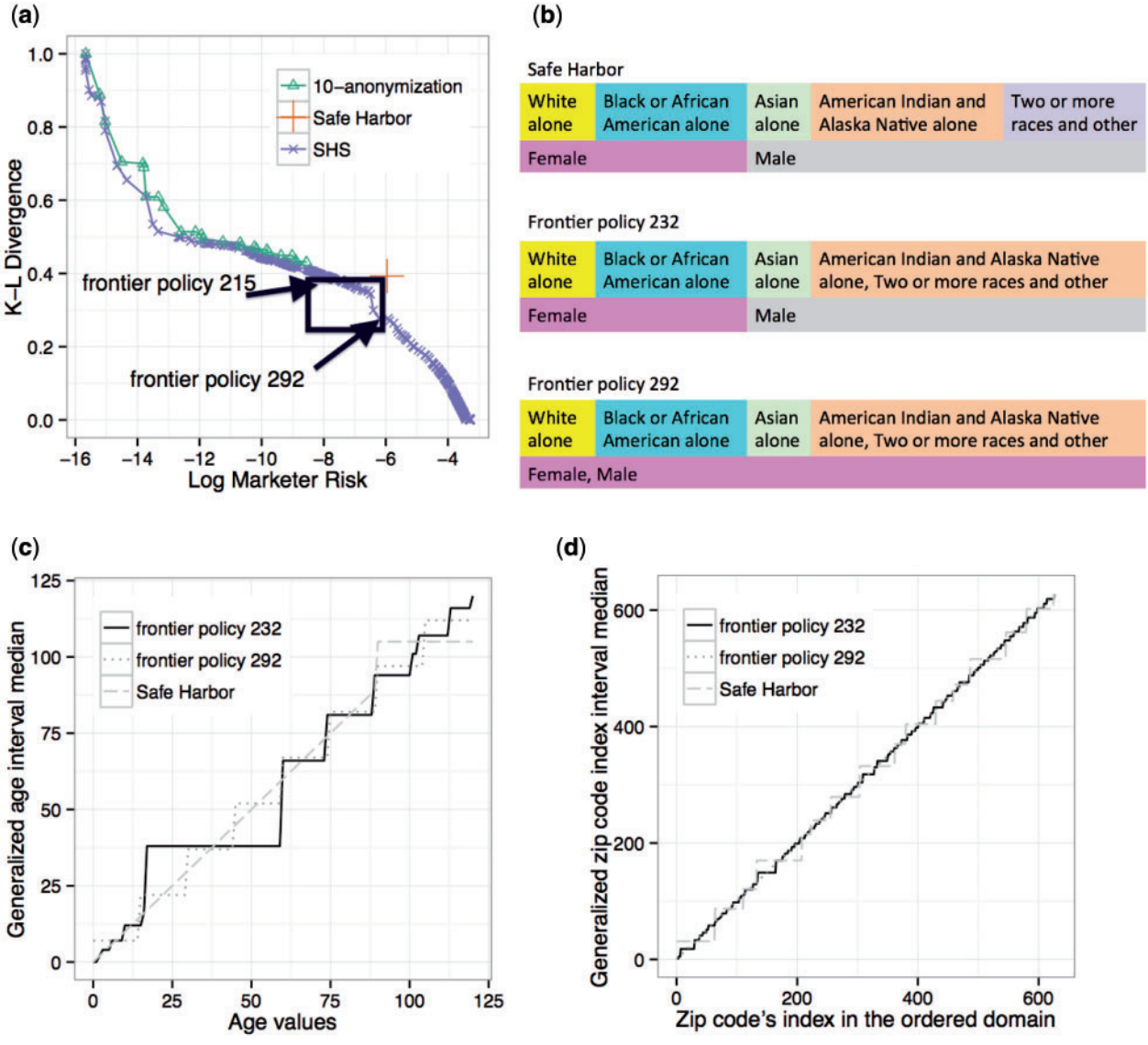


Figure 5: Results from the case study for the Adult-TN dataset. (**a**) A comparison of the 10-anonymization frontier, Safe Harbor policy, and the Sublattice Heuristic Search (SHS) frontier in the R-U space. The policies between the 215th and the 292nd on the SHS frontier (in the rectangle) dominate Safe Harbor. (**b**)–(**d**) provide a comparison of Safe Harbor and two dominating policies −232 and 292. (**b**) A comparison of the generalization rules for *race* and *gender* attributes. (**c**) A comparison of the age generalization rule. The *x*-axis corresponds to the original age, while the *y*-axis corresponds to the median of the generalized age interval. (**d**) A comparison of the ZIP generalization rule. The *x*-axis corresponds to the original ZIP, while the *y*-axis corresponds to the median of the ZIP interval. The ZIP codes are represented as an ordinal index, the translation for which can be found in the Supplementary Appendix online.

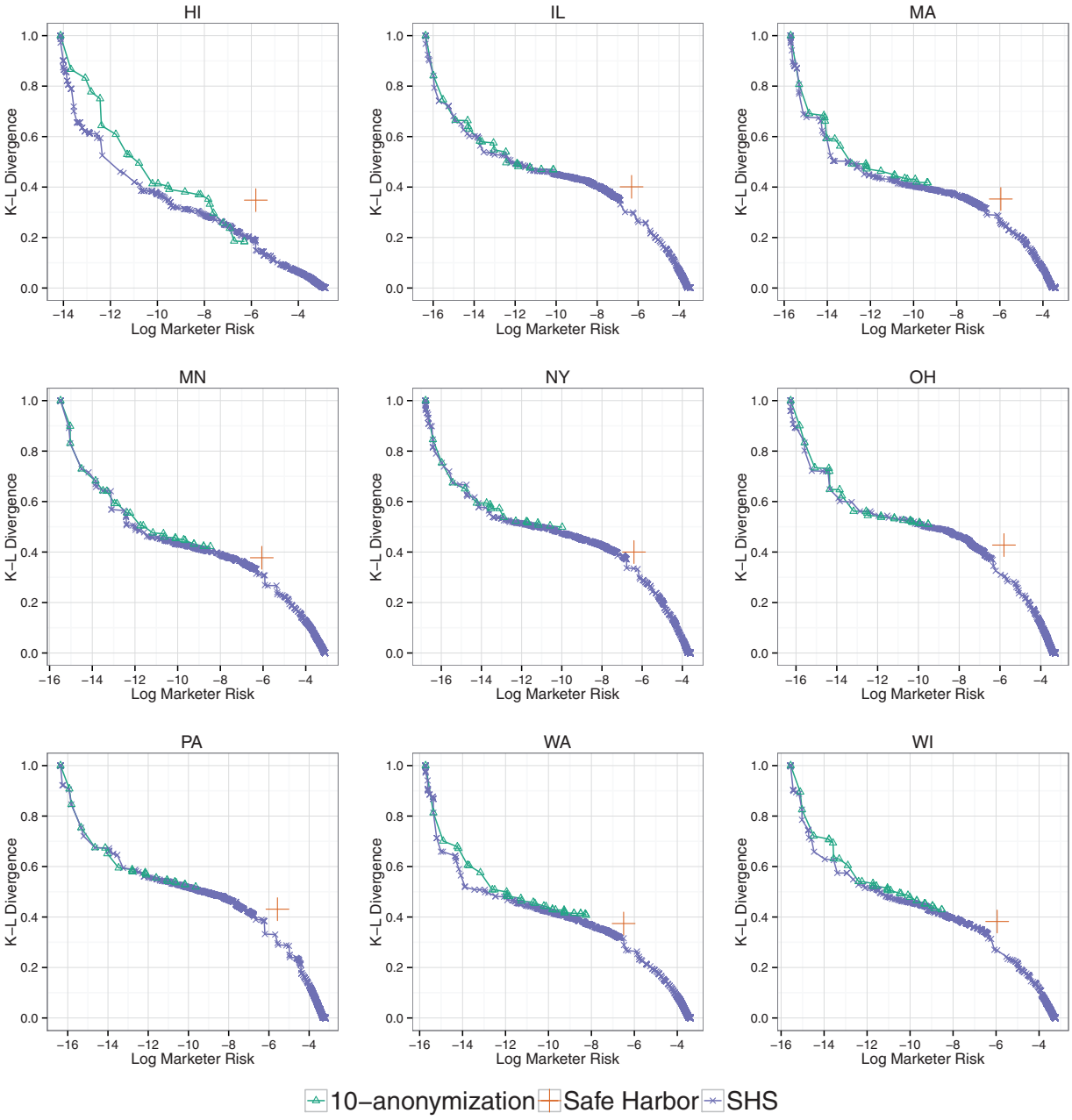


anonymization solutions are constrained in a very small sub-interval of the SHS frontier. This interval tends to have very small risk and large utility loss. Thus, SHS may be particularly useful when the data publisher is interested in solutions with better utility at the cost of an acceptable increase in risk. For instance, in the state of NY, the maximum risk of the 5-anonymization frontier is only 0.003 that of the SHS frontier. On the other hand, the minimum utility loss of the 5-anonymization frontier is between 0.15 and 0.52, while SHS is always at 0. This phenomenon is visualized is in Figure 5 (a), where the 10-anonymization has a much smaller range than that of the SHS frontier. This finding indicates that the SHS frontier can provide solutions in a broader range than the *k*-anonymity frontier.

Improvement of the Frontier R-U Tradeoff

The frontier R-U tradeoff improvement made by SHS over *k*-anonymization is outlined in Table 4. We use the relative difference of AU of the *k*-anonymization frontier F_k and the SHS frontier F_s to represents the R-U tradeoff improvement rate of the SHS frontier over the *k*-anonymization frontier: $\mathrm{IR} = (\mathrm{AU}(F_k) - \mathrm{AU}(F_s))/(\mathrm{AU}(F_s))$. A positive value indicates F_s improves upon F_k. We truncate the SHS frontier to be in the

Figure 6: A comparison of the 10-anonymization frontier, the Safe Harbor policy, and the the Sublattice Heuristic Search (SHS) frontier in the R-U space for the Adult dataset simulated over nine U.S. states.



same range of the corresponding *k*-anonymization frontier for a fair comparison.

In 9 out of 10 states (OH being the exception), the SHS frontier dominates the *k*-anonymization frontier. Recall that a positive value in this table indicates the corresponding *k*-anonymization frontier is dominated by the SHS frontier.

DISCUSSION

The present study illustrates that the SHS framework is, under many conditions, superior to *k*-anonymization strategies, as well as existing one-size-fits-all policies often invoked in practice (e.g., HIPAA Safe Harbor). Specifically, we observed that *k*-anonymity limits a frontier to a very small range with extremely low risk and potentially significant loss in data utility. We emphasize that our empirical analysis was performed over a range of diverse population distributions from 10 U.S. states to mitigate biases in the results. We believe that the SHS strategy has the potential to be a method that overcomes the limitations of a single fixed rule-based policy while being interpretable to health data managers. A healthcare organization, for instance, could present the policy frontier as a "documented method"[26] to an Institutional Review Board or legal counsel to justify its selection of a certain degree of protection when sharing data in a de-identified manner.

At the same time, there are several limitations to this work. First, the policy lattice was constructed under the assumption that the set of quasi-identifying attributes is known to the data publisher as *a priori*. We believe, however, there are several possible ways by which our method could be extended to address this problem. One potential strategy is to construct a policy lattice of the superset of all the possible quasi-identifier sets of attributes and measure re-identification risk as a weighted sum of the risk associated with each potential quasi-identifier. The weight of each quasi-identifier could be dependent on the availability of the corresponding external data resources. An alternative strategy is to construct a policy lattice for every subset with a size no greater than a threshold of the set of all the possible quasi-identifying attributes and search for a policy frontier in each space. Applying the latter would require a strategy to reconcile the de-identification policy associated with attributes that are in the overlapping part

Table 1: Number of policies on the frontier for the Adult dataset with ZIP codes simulated based on U.S. census data

State	Number of policies on frontier		
	SHS	$k=5$	$k=10$
HI	4545	28	25
IL	3999	28	21
MA	3510	29	23
MN	5655	34	25
NY	4374	27	20
OH	3257	39	27
PA	4161	29	23
TN	7766	39	27
WA	5296	33	35
WI	5147	42	34
Average	4771	33	26
St. Dev.	1234	5.5	5.03

Table 2: Proportion of policies that dominate Safe Harbor for the Adult dataset with ZIP codes simulated based on 2010 U.S. census data

State	Proportion of frontier dominating Safe Harbor		
	SHS	$k=5$	$k=10$
HI	0.01	0.36	0.28
IL	0.04	0	0
MA	0.01	0	0
MN	0.03	0.09	0
NY	0.001	0	0
OH	0.06	0	0
PA	0.003	0	0
TN	0.01	0.13	0
WA	0.01	0.12	0
WI	0.01	0.10	0
Average	0.018	0.08	0.028
St. Dev.	0.018	0.11	0.084

Table 3: Maximum risk values (MAX Risk) and minimum utility loss (MIN Utility Loss) of the frontiers for the Adult dataset with ZIP codes simulated from U.S. census data

State	Max. risk			Min. utility loss		
	SHS	$k=5$	$k=10$	SHS	$k=5$	$k=10$
HI	0.057	2.6×10^{-3}	1.8×10^{-3}	0	0.15	0.18
IL	0.031	2.4×10^{-4}	4.0×10^{-5}	0	0.43	0.47
MA	0.032	3.8×10^{-4}	9.0×10^{-5}	0	0.36	0.42
MN	0.045	5.2×10^{-4}	2.1×10^{-4}	0	0.36	0.42
NY	0.027	9.0×10^{-5}	4.0×10^{-5}	0	0.48	0.50
OH	0.037	2.9×10^{-4}	7.0×10^{-5}	0	0.45	0.51
PA	0.041	2.9×10^{-4}	7.0×10^{-5}	0	0.49	0.52
TN	0.037	4.3×10^{-4}	1.9×10^{-4}	0	0.36	0.43
WA	0.033	4.1×10^{-4}	2.6×10^{-4}	0	0.35	0.41
WI	0.039	4.6×10^{-4}	2.0×10^{-4}	0	0.36	0.43
Average	0.038	5.7×10^{-4}	3.0×10^{-4}	0	0.38	0.43
St. Dev.	0.009	7.1×10^{-4}	5.5×10^{-4}	0	0.10	0.10

Table 4: Frontier R-U tradeoff improvement rate of the SHS over *k*-anonymization (IR) for the Adult dataset with ZIP codes simulated based on U.S. census data

State	IR	
	$k=5$	$k=10$
HI	0.054	0.050
IL	0.035	0.018
MA	0.058	0.062
MN	0.020	0.049
NY	0.034	0.034
OH	−0.003	0.005
PA	0.018	0.019
TN	0.018	0.050
WA	0.040	0.060
WI	0.026	0.050
Average	0.030	0.040
St. Dev.	0.017	0.019

of multiple policy spaces (e.g., age, if [Age, Zip, Gender] and [Age, Gender, Race] are both possible quasi-identifiers).

Second, our search strategy does not cover the entire policy space. As such, the frontier is not guaranteed to be optimal. SHS is based on several heuristics and it is possible that more effective approaches could be developed. It may also be possible to develop methods to more systematically and efficiently navigate the space of policies using advanced pruning strategies, such as cost bounding. Moreover, the lattice search process should be amenable to parallel computing techniques as has recently been achieved for *k*-anonymization[58] provided an appropriate master program that minimizes reassessment of sections of the lattice can be designed.

Third, our investigation is based on specific measures of risk and utility. In particular, we rely on the marketer risk model, which amortizes the risk over all records in a published dataset. Yet, this is only one way to define risk. The amortization model itself, for instance, can be refined to allow for a discounting function that applies greater weight to individuals in smaller groups. Beyond the risk model, one could also consider worst-case re-identification scenarios, such as *prosecutor* or *journalist* attacks (which state that the risk of a dataset is equal to that of the most risky record).[39] From the perspective of utility, it is important to recognize that we adopted a generic information loss measure, which was based on the assumption that the specific usage of the dataset is unknown *a priori*. The data utility function is not necessarily consistent with the usage of the dataset in certain clinical data mining or statistical analysis applications. Nonetheless, if it is known that the dataset will be used in a certain study, then the frontier policy search framework can be customized with an alternative utility function defined by domain experts, provided that the function satisfies the monotonicity requirements of our framework.

Finally, while SHS builds a better frontier than other methods, it can yield a very large number of policies. A data manager would still need to determine which policy is best and it is clear that they could not review every policy on the frontier. As such, a strategy to present the most interesting policy options should be devised.

CONCLUSIONS

Current regulations permit institutions to publish de-identified data according to two conceptual routes: 1) fixed rules-based policies and 2) statistically-informed strategies that appropriately mitigate the risk of re-identification. While formal privacy models such as *k*-anonymity provide guarantees of protection, they can, at times, be too rigorous, leading to unacceptable levels of data utility in comparison to a rule-based policy. This paper showed, with evaluation over multiple datasets, that a de-identification policy frontier can provide a broader range of options than a well-known *k*-anonymization algorithm. In most cases, the de-identification frontier dominates 5-anonymization and is always superior to the popular HIPAA Safe Harbor rule-based policy. There are, however, several opportunities to improve the efficiency and selectivity of the SHS method to ensure it is directly usable by health data managers.

FUNDING

This work was supported, in part, by the Australian Research Council grant number [DP110103142], National Science Foundation grant number [CCF0424422], National Institutes of Health grant numbers [U01HG006385, U01HG006378, UL1TR000135, R01HG006844], and National Natural Science Foundation of China grant number [61472148].

COMPETING INTERESTS

None.

CONTRIBUTORS

WX designed the algorithm, performed the study, analyzed the experiment results, and wrote the paper. RH assisted in the experiment design, data analysis, and revised the paper. XD and JL provided insights in the outline of the study, analysis of the results and the refinement of the paper. BM designed and supervised the study, analyzed the data, and revised the paper.

ACKNOWLEDGEMENTS

We thank the Vanderbilt Genome Electronic Records project members (particularly Ellen Clayton, Dana Crawford, Josh Denny, Dan Roden, Jonathan Schildcrout, and Sarah Stallings) for helpful discussions and members of the eMERGE Coordination Center (particularly Melissa Basford and Jonathan Haines) for useful feedback during this work. We also thank Steve Nyemba for configuration of the servers used in this study.

REFERENCES

1. Lohr S. The age of big data. *New York Times*. February 11, 2012.
2. Schneeweiss S. Learning fro big health care data. *N Engl J Med*. 2014;370:2151–2153.
3. Murdoch TB, Detsky AS. THe inevitable application of big data to health care. *JAMA*. 2013;309:1351–1352.
4. Sun J, Reddy CK. Big data analytics for healthcare. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2013:1525; ACM.
5. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *J Biomed Inform*. 2014;48:160–170.
6. Post AR, Kurc T, Cholleti S, *et al*. The Analytic Information Warehouse (AIW): a platform for analytics using electronic health record data. *J Biomed Inform*. 2013;46:410–424.

7. Denny JC Bastarache L, Ritchie M, *et al*. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol.* 2013;31: 1102–1110.
8. Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc.* 2013;20: e147–e154.
9. Pathak J, Kho A, Denny J. Electronic health records-driven phenotypoing challenges: recent advances, and perspectives. *J Am Med Inform Assoc.* 2013;20:e206–e211.
10. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J *Gen Intern Med.* 2013; 28 (Suppl 3): S660–S665.
11. Schildcrout JS, Denny JC, Bowton E, *et al*. Optimizing drug outcomes through pharmacogenetics: a case for preemptive genotyping. *Clin Pharmacol Ther.* 2012;92:235–242.
12. Rea S, Pathak J, Savova G, *et al*. Building a robust, scalable, and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. *J Biomed Inform.* 2012;45:763–771.
13. McGlynn EA, Lieu TA, Durham ML, *et al*. Developing a data infrastructure for a learning health system: the PORTAL network. *J Am Med Inform Assoc.* 2014;21:596–601.
14. Arzberger P, Schroeder P, Beaulieu A, *et al*. Science and government. *An international framework to promote access to data. Science.* 2004;303: 1777–1778.
15. Chalmers I, Altman DG, McHaffie H, *et al*. Data sharing among data monitoring committees and responsibilities to patients and science. *Trials.* 2013; 14:102.
16. National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies. NOT-OD-07-088. Bethesda, MD 2007.
17. Hallinan D, Friedewald M, McCarthy P. Citizens' perceptions of data protection and privacy in Europe. *Computer L Sec Rev* 2012;28:263–272.
18. King T, Brankovic, Gillard P. Perspectives of Australian adults about protecting the privacy of their health information in statistical databases. *Int J Med Inform Assoc.* 2012;81:279–289.
19. Olson JS, Grudin J, Horvitz E. A study of preferences for sharing and privacy. In: Proceedings of the Extended Abstracts on Human Factors in Computing; 2005:1985–1988.
20. Perera G, Holbrook A, Thabane L, Foster G, Willison D. Views on health information sharing and privacy from primary care practices using electronic medical records. *Int J Med Inform.* 2011;80:94–101.
21. Solove DJ. A taxonomy of privacy. *Univ Penn L Rev.* 2006;154:477–559.
22. Schadt EE. The changing privacy landscape in the era of big data. *Mol Syst Biol.* 2012;8:612.
23. Tene O, Polonetsky J. Privacy in the age of big data: a time for big decisions. *Stan L Rev Online.* 2012;64:63.
24. McGraw D. Building public trust in uses of health insurance portability and accountability Act de-identified data. *J Am Med Inform Assoc.* 2013;20: 29–34.
25. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. 1995.
26. U.S. Dept.of Health & Human Services. Standards for privacy of individually identifiable health information, final rule, 45 CFR, pt 160–164. 2002.
27. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc.* 2010;17:169–177.
28. Sweeney L. Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics.* 1997;25:98–110.
29. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *J Am Med Inform Assoc.* 2011;18:3–10.
30. Sweeney L. K-anonymity: a model for protecting privacy. *Int J Uncertain, Fuzz* 2002;10:557–570.
31. Xia W, Heatherly R, Ding X, Li J, Malin B. Efficient discovery of de-identification policy options through a risk-utility frontier. In: Proceedings of the 3rd ACM Conference on Data and Applications Security and Privacy; 2013: 59–70; ACM.
32. Bache K, Lichman M. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. Accessed June 10 2014.
33. Gottesman O, Kuivaniemi H, Tromp G, *et al*. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med.* 2013;15:761–771.
34. Dalenius T. Finding a needle in a haystack or identifying anonymous census records. *J Off Stat.* 1986;2:329–336.
35. Fung BCM, Wang K, Chen R, *et al*. Privacy-preserving data publishing: a survey of recent developments. *ACM Comput Surv.* 2010;42:14:1–14:53.
36. Office for Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) Privacy Rule. U.S. Dept. of Health and Human Services. November 2012.
37. Machanavajjhala A, Kifer D, Gehrke J, *et al*. l-diversity: privacy beyond k-anonymity. *ACM Trans Knowl Discov Data.* 2007;1:1.
38. Li N, Li T, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the IEEE 23rd International Conference on Data Engineering; 2007:106–115.
39. El Emam K, Dankar FK. Protecting privacy using *k*-anonymity. *J Am Med Inform Assoc.* 2008;15:627–637.
40. El Emam K, Dankar FK, Issa R, *et al*. A globally optimal *k*-anonymity method for the de-identification of health data. *J Am Med Informatics Assoc.* 2009; 16:670–682.
41. Mohammed N, Fung BCM, Hung PCK, *et al*. Centralized and distributed anonymization for high-dimensional healthcare data. *ACM Trans Knowl Discov Data.* 2010;4:18:1–18:33.
42. Ciriani V, di Vimercati S, Foresti S, *et al*. *k-anonymity*. In: Yu T, Jajodia S, eds. Secure Data Management in Decentralized Systems. Springer US. 2007; 323–353.
43. Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression. *Int J Uncertain, Fuzz.* 2002;10:571–588.
44. Iyengar VS. Transforming data to satisfy privacy constraints. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002:279–288.
45. Meyerson A, Williams R. On the complexity of optimal *k*-anonymity. In: Proceedings of the 23rd ACM SIGMOD-SIGARCT-SIGART Symp on Principles of Database Systems; 2004: 223–228.
46. Samarati P, Sweeney L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. In: Proceedings of IEEE Symposium on Research in Security and Privacy; 1998.
47. Bayardo RJ, Agrawal R. Data privacy through optimal *k*-anonymization. In: Proceedings of the 21st IEEE International Conference on Data Engineering; 2005: 217–228.
48. Duncan GT, Keller-McNulty SA, Stokes SL. *Disclosure risk vs. data utility: The R-U confidentiality map*. Technical Report Number 121, National Institute for Statistical Sciences: Research Triangle Park, NC. 2001. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.1598&rep=rep1&type=pdf. Accessed June 10 2014.
49. Li T, Li N. On the tradeoff between privacy and utility in data publishing. In: Proceedings of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2009: 517–526.
50. Dewri R, Ray I, Whitley D. On the optimal selection of k in the k-anonymity problem. In: Proceedings of 24th IEEE International Conference on Data Engineering; 2008: 1364–1366.
51. Loukides G, Gkoulalas-Divanis A, Shao J. On balancing disclosure risk and data utility in transaction data sharing using R-U confidentiality map. In: Proceedings of the Joint UNECE/Eurostat Working Session on Statistical Data Confidentiality; 2011: 19.
52. Benitez K, Loukides G, Malin B. Beyond Safe Harbor: automatic discovery of health information de-identification policy alternatives. In: Proceedings of the 1st ACM International on Health Informatics Symposium; 2010: 163–172.
53. U.S. Census Bureau. American fact finder website. http://www.americanfactfinder.gov. Accessed June 18, 2014.

54. North Carolina voter registration. [Online]. ftp://www.app.sboe.state.nc.us/data. Accessed January 27, 2014.
55. LeFevre K, DeWitt D, Ramakrishman R. Incognito: efficient full-domain K-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Mangement of Data; 2005: 49–60.
56. Federal Committee on Statistical Methodology. Statistical Policy Working Paper 22: Report on statistical disclosure limitation methodology. U.S. Office of Management and Budget, Washington, DC, 2005.
57. Truta TM, Fotouhi F, Barth-Jones D. Disclosure risk measures for microdata. Barth-Jones. In: Proceedings of 15th International Conference on Scientific and Statistical Database Management; 2003: 15–22.
58. Zhang X, Yang C, Nepal S, Liu C, Dou W, Chen J. A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud. In: Proceedings of the 3rd International Conference on Cloud and Green Computing; 2013: 105–112.

AUTHOR AFFILIATIONS

[1]Department of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN, USA

[2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

[3]Huazhong University of Science and Technology, Wuhan, China

[4]School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, South Australia, Australia