future internet

MDPI

Article

Analytics on Anonymity for Privacy Retention in Smart Health Data †

Sevgi Arca * and Rattikorn Hewett

Department of Computer Science, Texas Tech University, Lubbock, TX 79409, USA; rattikorn.hewett@ttu.edu
* Correspondence: sevgi.arca@ttu.edu

† This paper is an extended version of our paper published in the Proceedings of the 11th International. In Proceedings of the Conference on Advances in Information Technology, Privacy Protection in Smart Health, Bangkok, Thailand, 1–3 July 2020.

Abstract: Advancements in smart technology, wearable and mobile devices, and Internet of Things, have made smart health an integral part of modern living to better individual healthcare and well-being. By enhancing self-monitoring, data collection and sharing among users and service providers, smart health can increase healthy lifestyles, timely treatments, and save lives. However, as health data become larger and more accessible to multiple parties, they become vulnerable to privacy attacks. One way to safeguard privacy is to increase users' anonymity as anonymity increases indistinguishability making it harder for re-identification. Still the challenge is not only to preserve data privacy but also to ensure that the shared data are sufficiently informative to be useful. Our research studies health data analytics focusing on anonymity for privacy protection. This paper presents a multi-faceted analytical approach to (1) identifying attributes susceptible to information leakages by using entropy-based measure to analyze information loss, (2) anonymizing the data by generalization using attribute hierarchies, and (3) balancing between anonymity and informativeness by our anonymization technique that produces anonymized data satisfying a given anonymity requirement while optimizing data retention. Our anonymization technique is an automated Artificial Intelligent search based on two simple heuristics. The paper describes and illustrates the detailed approach and analytics including pre and post anonymization analytics. Experiments on published data are performed on the anonymization technique. Results, compared with other similar techniques, show that our anonymization technique gives the most effective data sharing solution, with respect to computational cost and balancing between anonymity and data retention.

Keywords: health data anonymity analytics; privacy in smart health; data anonymization

Citation: Arca, S.; Hewett, R. Analytics on Anonymity for Privacy Retention in Smart Health Data. *Future Internet* **2021**, *13*, 274. https://doi.org/10.3390/fi13110274

Academic Editors: Joel J. P. C. Rodrigues and Carlo Blundo

Received: 18 September 2021
Accepted: 22 October 2021
Published: 28 October 2021

1. Introduction

Smart health improves the well-being and quality of lives by providing customized cares and treatments using health data collected from smart health devices (e.g., trackers of movements and heart rates [1], or mobile EKG (electrocardiogram) monitors for heart rhythms [2]). Telemedicine increasingly relies on health devices to treat chronic diseases, e.g., by monitoring glucose [3], blood sugar levels [4], or blood pressure [5] for patients with heart diseases and diabetes. Advancement in wearable technology and Internet of Things enable smart health in self-monitoring and delivery of users' health data to doctors, hospitals, and fitness service providers [6]. Smart health can increase healthy lifestyles, timely treatments, and save lives. Furthermore, collection and sharing of health data can help researchers navigate scientific discoveries. For example, genetic-testing companies collect users' DNA (Deoxyribonucleic Acid) and survey data to gain insights on genetic diseases like Parkinson, Late-onset Alzheimer 's or celiac disease [7].

While smart health brings great benefits, it also poses potential threats to privacy as health data often contains sensitive and disclosed information. Collecting, storing, and sharing these data can put users' privacy at risks of being re-identified (even if personally

identifiable information is removed) or loss of personal information by inferences from overlapping external data, known as side-channel attacks [8]. As health data become larger and more accessible to multiple parties, they become more susceptible to privacy attacks as users may lose control of their personal information. A common practice to safeguard privacy is to increase users' anonymity as anonymity increases indistinguishability, making it harder for re-identification. For example, let us consider a published dataset that consist of people in a health club. Being the only person of age between 30–40 in a published dataset makes Bob easily re-identifiable from the dataset. If an attacker is looking for Bob's home address and knows that Bob is about 35, not only his home address can be found but his genetic disease and other personal information can also be revealed. Compared to Mary being one of the age 30 female health club members, Mary is more anonymous (among female club members) than Bob (among those of age 30–35). Because unlike Bob, Mary is not easily distinguishable since there are other people with the same information as Mary. Since the attacker cannot identify Mary among other age 30 female club members, she is more anonymous. Thus, Mary's privacy is better protected because she has higher anonymity than Bob. Anonymity ensures that each set of 'critical' data values is associated with more than one individual (or a minimum requirement) to protect the individual's identity [8–10]. When the minimum requirement is specified to k, we refer to such a condition as k-anonymity, where each unique 'critical' data values has at least k records (or people) [7,8].

Much work on privacy and anonymity analytics has been studied in two groups: anonymity measures [11–14] and anonymization techniques [8–10,12,14–25]. The former deals with indirect anonymity measures based on relevant information (e.g., average information losses [13,14], likelihood of attacker's correct re-identification given his prior knowledge [12]). The latter deals with transforming a given database into a more anonymous form to better protect privacy. Many anonymization techniques have been researched [8–10,14,15,18–22] using suppression and generalization by either omitting [18,20] or replacing critical data values with more general substitutes (according to its taxonomies) [22]. Some transforms the original data into a generalized table that complies with k-anonymity requirements [8–10,19]. Many techniques have focused on minimizing generalizations to enhance computational efficiency. Still the challenge is not only to efficiently preserve data privacy but also to ensure that the shared data are sufficiently informative to be useful. Excess generalization to provide anonymity can corrupt important information that the data may convey. As a result, the data become less informative. Recent anonymization approaches [16,17,23] aim to increase data information (e.g., by using information gain [17] but they are intended to be used resulting anonymized data for classification.

Our research studies health data analytics focusing on anonymity for privacy protection. This paper presents a multi-faceted analytical approach to (1) identifying attributes susceptible to information leakages by using entropy-based measure to analyze information loss, (2) anonymizing the data by generalization using attribute hierarchies, and (3) balancing between anonymity and informativeness by our anonymization technique that produces anonymized data satisfying a given anonymity requirement while optimizing data retention. Our anonymization technique is an automated Artificial Intelligence search based on two simple heuristics. Part (1) can be viewed as pre-anonymization analytics, whereas part (2) is anonymization with a refinement in part (3). Because there are many anonymization techniques and each may have different objectives, the post anonymization analytics can be performed by applying Part (1) again. The contribution of this paper is not only techniques for each part but an overall methodology for analyzing privacy and anonymity of health data. The proposed method is practical in a sense that it is applicable to any kind of smart health data and moreover balancing the information loss and anonymity makes the approach feasible to use. This paper is an extension from our previous work [15] that only describes part (3) focusing on our proposed anonymization algorithm. The rest of the paper is organized as follows. Section 2 describes related work followed by the

multi-faceted analytical approach in Section 3, which can be viewed as pre-anonymization and anonymization steps. Section 4 describes experiments to evaluate performance and effectiveness of our anonymization technique when compared with similar techniques. Section 5 provides post anonymization analytics and Section 6 concludes the paper.

2. Related Work

Much research in data privacy addresses issues on anonymity [8–10,21,22,24–29]. Many aim to measure anonymity [11–14], whereas some are concerned with the utility of the result [24,25,29]. Majority of the metrics that are concerned with anonymization quality [11,13,14] use Shannon's entropy to quantify average information [13,14]. Work in [13] uses entropy to estimate the average number of correct re-identifications (of individuals) based on binary queries. More correct responses help increase the attacker's information about the individuals in the database and reduce the anonymity of the individuals. Longpre et al. proposed a measure [14] to estimate an average information loss when an attacker acquires additional information through querying. Again, the more information the attacker gains (or average information loss), the less anonymity users have. This makes it easier for the attacker to breach disclosure and identify the users. Our paper suggests a method to analyze the data using this latter measure to pinpoint areas in the data that are susceptible to privacy attacks. Some anonymization techniques consider the utility of the data after the anonymization for evaluation [24,25,29]. Among those, work in [25] considers information loss and calculates utility accordingly whereas some considers classification accuracy to measure utility [24,29].

A large body of research in anonymity concerns with anonymization techniques to transform a given data set into a more anonymous form for privacy preservation [8–10,14–22]. Most of these techniques find anonymized data (via generalization) that complies with *k*-anonymity requirement [8–10,19,20] to guarantee that each group of unique critical attribute values has at least *k* records to prevent individuals from being reidentified easily. Some anonymization uses exhaustive search to find the minimal *k*-anonymization with minimal distortion [9,22]. Although the approach is not practically feasible, it provides a concrete formal model for minimal *k*-anonymization. Work in [9] searches for *k*-anonymizations using a binary search. Since binary search is a blind search, computational cost can still pose a problem when searching for all possible *k*-generalizations as discussed in [15]. Other approaches focus on efficiency rather than minimal k-anonymizations [19,20]. Unlike the approaches that search for minimal generalizations blindly or focusing on efficiency rather than minimal anonymization, our proposed approach [15] aims to efficiently search for anonymized data that strike a balance between satisfying *k*-anonymity requirements and maximizing retention of the original data.

3. Proposed Multi-Faceted Anonymity Analytics Approach

This section describes the proposed approach that analyzes anonymity in multiple aspects to protect data owner's privacy. Figure 1 shows a general overview of the approach. As shown in the figure, the approach identifies the attributes susceptible to information leakages in pre-analytics process. The user can choose to increase anonymity of the vulnerable attributes before the anonymization procedure or directly anonymizes the data using the findings of the vulnerable attributes as guidance. The user can also choose to anonymize the smart health dataset, without applying pre-analytics. Then, the data are anonymized, by our IAB (Intelligent Anonymity Balance) anonymization technique that produces data satisfying a given anonymity requirement while optimizing data retention. The anonymized data are then analyzed in the post-analytics process to see if the critical attributes in the resulting anonymized data are vulnerable to information leakage (e.g., via inference of attackers). If they are further actions can be taken (e.g., alerting data publishers or injecting additional "fake" data to increase indistinguishability of the vulnerable individuals). The first two steps are described in this section whereas the post-analytics step will be described in Section 5. For easy referencing, since we will describe and illustrate

each section with the same data, we briefly introduce them along with common terms and notations below.

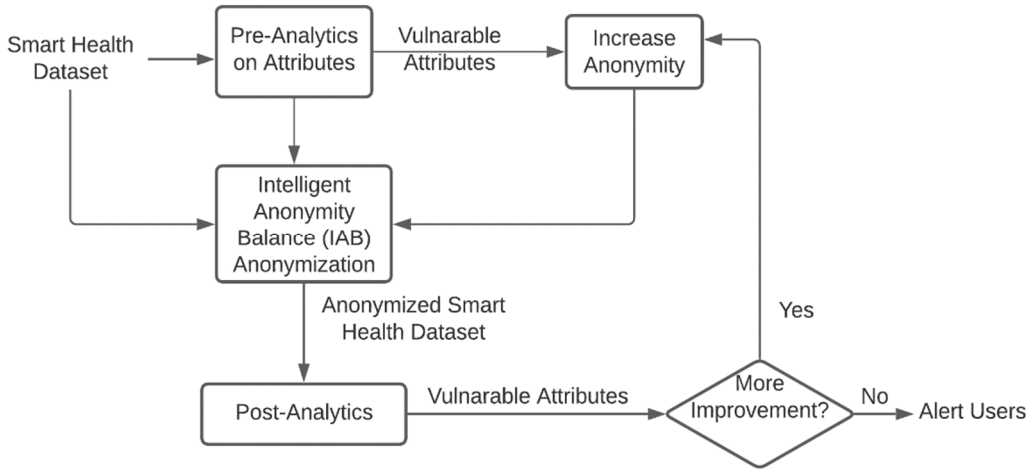

Figure 1. Overview of the Proposed Multi-faced Anonymity Analytical Approach.

Given a (data) table T (or relational database) with A, a set of attributes A_1, A_2, ..., A_n., a data *record* represents an instance of a tuple (a_1, a_2, ..., a_n), where data entry $a_i \in dom(A_i)$, a set of all possible values of A_i. Consider Table 1, each row represents a unique tuple of attribute values where the last column represents the number of records for each row. Here Row 2 represents a unique tuple (F, Low, 35, 52000, 143, Black, No) with three instances of records. As shown in Table 1, Rows 1, 4, 10, 13 are obviously vulnerable to privacy threats since each has one record instance giving low anonymity and easy for re-identification. Next, we will describe the analytical approach.

Table 1. Users' Weight Loss Profiles.

Rows	Sex	Alcohol Cons.	Age	Zip	Weight	Race	Genetic Risk	#Rec.
1	F	Med	35	52000	143	Black	No	1
2	F	Low	35	52000	143	Black	No	3
3	M	Med	50	53000	166	White	No	4
4	M	Low	35	52003	143	White	No	1
5	M	Med	68	52000	190	Hispanic	No	3
6	M	High	68	52000	190	Hispanic	No	4
8	F	Low	75	52002	122	Native Hawaiian	No	4
9	F	Med	35	52003	143	Black	No	2
10	F	High	75	52002	143	White	Hereditary Thrombophilia	1
11	M	High	44	52003	166	Asian	Hereditary Thrombophilia	3
12	F	Med	38	52000	166	White	No	4
13	F	High	35	52000	122	American Indian	Hereditary Thrombophilia	1
14	F	Low	35	52000	122	American Indian	Hereditary Thrombophilia	2
15	M	Med	38	52003	122	White	No	3
16	F	Med	22	54004	180	Native Hawaiian	No	3
17	M	High	20	54000	180	American Indian	L.O. Alzheimer's	3
18	M	Low	21	54001	180	Black	Parkinson's	3
19	F	High	24	54003	180	Asian	Hereditary Thrombophilia	3
20	M	No	25	54000	122	American Indian	Celiac	3
21	M	No	26	54001	166	Asian	L.O. Alzheimer's	3
22	F	High	23	54003	180	Asian	Parkinson's	3

3.1. Assessing Vulnerability to Information Leakages

Before we transform a given health data into a more anonymous form, one may investigate if (and what areas of) the data are susceptible to information loss if an attacker uses some of his information to make inferences. To do this, we propose an analysis on various structures of the data using the Longpre et al.'s entropy-based measure [14] to estimate average information loss in respective areas. The motivation of this pre-anonymization analytics is not simply to apply existing measure in a typical manner but maximizing the measure for systematic use to gain useful information for privacy protection. For example, the finding that certain attribute is vulnerable to information leakage may be linked to low anonymity that can be alleviated by modifying the original data. Next, we briefly describe the measure and its derivations from two sources.

Proposition 1. *Shannon's information quantification.*

Let X be a discrete random variable with outcomes x_1, x_2,...., x_n, $p(x_i)$ be the probability of x_i being the outcome, and $I(x_i)$ be the amount (or value) of information received when learning that x_i is the outcome (sent). Then $I(x_i)$ is $log_2(1/p(x_i))$.

Proof. Since the more probable the information is, the less informative the information becomes. Thus, $I(x_i)$ is inversely proportional to $p(x_i)$. Furthermore, for information of value y, the amount of information is measured by the number of bits to store y, i.e., $log_2(y)$ bits. Thus, Shannon's quantifying information $I(x_i) = log_2(1/p(x_i))$. □

Proposition 2. *Longpre et al.'s entropy-based measure.*

Given a data table of n individuals, where $p(r_i)$ is the probability of individual r_i being re-identified. An attacker makes queries, each of which has m possible answers represented in a sequence $<a_1, a_2,..., a_m>$. All n individuals are partitioned into m partitions, where each partition E_j contains individuals whose attribute value matches the jth answer of the query a_j. Subsequently, the average of information loss is $\Delta S(\{E_j\}) = \sum_{j=1}^{m} p(E_j)(S_0 - S_j)$, where $S_0 = -\sum_{i=1}^{n} p(r_i) log_2 p(r_i)$ and $S_j = \sum_{r_i \in E_j} p(r_i|E_j) log_2 p(r_i|E_j)$ representing an initial average amount of information (before queries) and the average of amount of information after the query answer j, respectively.

Proof. If the attacker knows $p(r_i)$ then the amount or value of the information can be quantified as $log_2(1/p(r_i))$ by Proposition 1. Thus, an average of these information values over all individuals gives an entropy $S_0 = -\sum_{i=1}^{n} p(r_i) \log_2 p(r_i)$. (Note, if an attacker does not have any information about individuals, then everyone in the table is equally likely to be identified with $p(r_i) = 1/n$.).

Now suppose an attacker makes queries as stated. Each individual r_i can belong to one partition. Thus, $\sum_{j=1}^{m} |E_j| = n$. Suppose an individual r_i is found to be in E_j then $p(r_i)$ becomes $p(r_i \mid E_j)$, which is $p(\{r_i\} \cap E_j)/p(E_j) = p(r_i)/p(E_j)$ (since $\{r_i\} \cap E_j = \{r_i\}$), where $p(E_j) = \sum_{k:r_k \in E_j} p(r_k)$. Since $p(r_i)$ is reduced, the information value/amount increases (as less certain is more informative). Thus, an attacker gains more information about the individual and more vulnerable to privacy breach. Thus, the average amount of information when answer j is matched $S_j = \sum_{r_i \in E_j} p(r_i|E_j) \log_2 p(r_i|E_j)$, where $p(r_i)$ is changed to $p(r_i \mid E_j)$. This gives an average loss to be estimated as $\Delta S(\{E_j\}) = \sum_{j=1}^{m} p(E_j)(S_0 - S_j)$. □

Note that $\Delta S(\{E_j\})$ is maximum when S_j is zero and $\Delta S(\{E_j\}) = S_0$ (i.e., no information is lost to the attacker or that he has no information). Hence, the normalized average information loss is $\Delta S(\{E_j\})/S_0$ where its value is in [0, 1].

Analytics on Information Leakages

Instead of applying the Longpre et al.'s entropy-based measure to the entire table, we will analyze which attribute will be most vulnerable to information leaks (i.e., leaks most amount) on the average when an attacker obtains information on the attribute values.

We will use Table 1 to illustrate and explain the concept. Suppose an attacker queries information on attribute Sex. Table 1 has a total of 60 individuals with 27 females (F) and 33 males (M). When an attacker has no information, every individual is equally likely to be identified with $p(r_i) = 1/60$. Therefore, the initial average amount of information S_0 is $-\sum_{i=1}^{60}(1/60)log_2(1/60) = 5.9$. For attribute Sex, there are two possible answers: <F, M>. Thus, we partition 60 individuals into E_1 and E_2 for those who are F and M, respectively.

Based on Proposition 2 and Table 1, $p(r_i \mid E_1) = p(r_i)/p(E_1) = (1/60)/(27/60) = 1/27$, for $r_i \in E_1 = \{r_i \mid i = 1, 2, 8–10, 12–14, 16, 19, 22\}$. This gives $S_1 = -27(1/27)log_2(1/27) = 4.75$. Similarly, $p(r_i \mid E_2) = (1/60)/(33/60) = 1/33$, for $r_i \in E_2 = \{r_i \mid i = 3–7, 11, 15, 17, 18, 20, 21\}$ and $S_2 = -33(1/33)log_2(1/33) = 5$. By Proposition 2, $\Delta S(\{E_j\})$ is $(27/60)(5.9 - 4.75) + (33/60)(5.9 - 5) = 0.99$. Normalizing by a maximum (i.e., when $\Delta S(\{E_j\}) = S_0 = 5.9$), we have the resulting normalized average information loss of $0.99/5.9 = 0.16$ (when the attacker queries on Sex) as shown in the first row of Table 2. Similarly, we can apply the measure to estimate the average information loss given the attacker querying on other attributes except the disclosed one (e.g., genetic risk). Table 2 shows the overall results obtained.

Table 2. Estimated Information Loss.

Query Attribute.	# Query Partitions	Avg. Info. Loss	Norm. Avg. Info. Loss
Sex	2	0.99	0.16
Alcohol Cons.	4	1.86	0.31
Age	13	**3.55**	**0.60**
Zip	8	2.75	0.46
Weight	5	2.24	0.38
Race	6	2.52	0.42

The normalized results show us on average how much information is leaked given attacker knows the attribute value of the person they are looking for. The attribute that discloses more information has a higher value out of the maximum possible value of one. As shown in Table 2, for data Table 1, the Age attribute is the most vulnerable as it leaks the most information. Next is Zip, followed by Race. These are not surprising as they are typical key attributes that lead to identity identification. Although we have not done this, the Longpre et al.'s entropy-based measure can be applied to a combination of attributes at any level to give different insights. Here we apply the measure to each non-disclosed attribute for systematic preliminary findings.

In general, this pre-anonymization analytics can help us decide which attributes we should pay attention to when we try to protect privacy. For example, we may pick a set of most vulnerable attributes to increase anonymity by generalization. In anonymization techniques, a set of such attributes is known as quasi-identifiers or shields that are specified by users. Next section shows more details of basic anonymization techniques.

3.2. Increasing Anonymity by Generalization

The analytics in Section 3.1 show that, once an attacker obtains the query answers, information on some attributes (or set of attributes) can lead to more average information loss than the others. To protect such loss, a common practice to increase anonymity is by generalization and compression [8–10]. This section describes these basic concepts in more details along with the concept of *k*-anonymity that is used in many anonymization techniques including ours (to described in Section 3.3).

Generalization replaces an attribute value by a more abstract form or a more general but semantically consistent value. For example, we can replace the zip "12345" by "123∗∗", or replace a "city" by its "country". The former can be viewed as a suppression of the last two digits of the zip where"∗" represents any non-negative digit. The consistency on semantics of attribute values is governed by its conceptual hierarchy. By doing this, the number of records of each unique tuple will increase and that increases the tuple's degree of anonymity. Consequently, individuals are more indistinguishable, and their identities are better protected. Generalization provides many advantages to preserve data privacy including consistent interpretation, traceability, and minimal content distortion [10].

We will now explain the concepts in more details via illustrations on Table 1. Continuing our analytics from Section 3.1, where we identify that Age, Zip and Race are vulnerable. One can focus on generalizing these attributes to increase their anonymity or exploring other attributes based on domain experts. Here we consider the three attributes: Alcohol Consumption (AC), Age and Zip and their corresponding conceptual hierarchies as shown in Figure 2. For AC, there are four attribute values in the domain although only three appear in Table 1. The Age attribute values are discretized into four ranges and the Zip attribute values are string of numbers where a more general value uses "∗" for any non-negative digit. The Zip hierarchy is general in that it is applicable to any string of digits other than 9's.

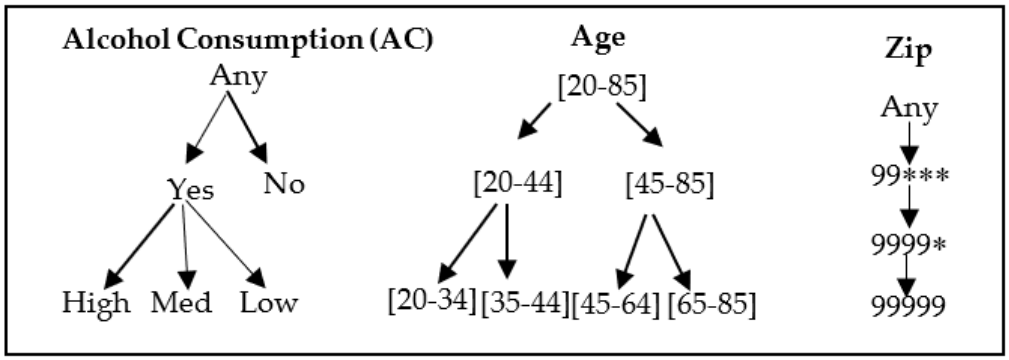

Figure 2. Taxonomy Trees of three attributes in Table 1.

For simplicity and without loss of generality, we will illustrate generalization on parts of Table 1, namely Rows 1, 2, 5, 6, 9 and 10 with four attributes: Sex, Alcohol Consumption (AC), Age and Zip, as shown in Table 3a to be an initial data table.

In Table 3a, Row 1 and Row 5 each has one record. This makes an individual in these two rows vulnerable for re-identification. If an attacker knows that the person he is looking for is a Female (F) having Medium (Med) AC and lives in Zip 52000, he will be able to identify the person and infer his age of 35 (see Row 1). Similarly, Row 5 is the only one record of a Female, Age 75, so this person can be identified and her sensitive information of having High AC can be leaked.

To increase anonymity of individuals in Rows 1 and 5, we generalize on AC cells of all rows of females (i.e., Rows 1, 2, 5, 6) in Table 3a to obtain results as shown in Table 3b where the change and important areas are colored. In this table, individual in Row 1 increases his/her anonymity since Row 1 can be merged with individuals in Row 2 creating a tuple (F, Yes, 35, 5200) with four records. However, this generalization is not enough to increase anonymity of individual in Row 5.

To increase anonymity of individual in Row 5 with the goal to merge with Row 6, we need to further generalize both rows on Age and Zip according to the taxonomies in Figure 2. By generalizing the Age attribute two steps to [20–85] and the Zip to 5200∗, we obtain the results as shown in Table 3c. As shown in this table, Rows 5 and 6 can now be merged. By merging Row 1 with Row 2, and Row 5 with Row 6, we obtain the final table as shown in Table 3d. Here none of the unique tuples of attribute values has a one record.

In fact, the record number indicates the degree of anonymity. Table 3d shows that there are at least three people in each group of the same attribute values and hence their identities and information are better protected.

Table 3. Increasing anonymity by generalization on attribute values.

Rows	Sex	AC	Age	Zip	#Rec.
1	F	Med	35	52000	1
2	F	Low	35	52000	3
3	M	High	68	52000	4
4	M	Med	68	52000	3
5	F	High	75	52002	1
6	F	Med	35	52003	2

(a) Initial table

Rows	Sex	AC	Age	Zip	#Rec.
1	F	Yes	35	52000	1
2	F	Yes	35	52000	3
3	M	High	68	52000	4
4	M	Med	68	52000	3
5	F	Yes	75	52002	1
6	F	Yes	35	52003	2

(b) Generalize some rows on AC

Rows	Sex	AC	Age	Zip	#Rec.
1	F	Yes	35	52000	1
2	F	Yes	35	52000	3
3	M	High	68	52000	4
4	M	Med	68	52000	3
5	F	Yes	[20–85]	5200 *	1
6	F	Yes	[20–85]	5200 *	2

(c) Generalize Rows 5, 6 on Age and Zip

Rows	Sex	AC	Age	Zip	#Rec.
1, 2	F	Yes	35	52000	4
3	M	High	68	52000	4
4	M	Med	68	52000	3
5, 6	F	Yes	[20–85]	5200 *	3

(d) Final result after merging rows

There are many ways to generalize. The above shows generalization at a *cell* level (i.e., a *data entry* of a specific row and column of a table). Another type of generalization is applied to all attribute values of the same level in the hierarchy. Thus, when a table is generalized on attribute A, the generalization is applied only to the table rows whose A's attribute values are either the child or its siblings of the same parent in the hierarchy. For example, generalizing a Table 3a on Age will replace the Age values of Rows 1, 2, and 6 to [20–44] and those of the rest of rows will be replaced by [45–85]. To improve efficiency, many anonymization techniques including ours (Section 3.1) adopt this interpretation when applying generalization. Next, we formally define important concepts for anonymization, namely, k-anonymity requirement and other relevant terminologies.

k-Anonymity Requirement for Anonymization

Anonymity requirement specifies an *anonymity degree* required on a subset of privacy critical attributes, called *shield* (or *quasi-identifiers* [24,25]). Given the degree k and the shield S, the *k-anonymity requirement* on shield S, denoted by <S, k>, is defined to be a set of S-projected tuples, whose each unique tuple is guaranteed to have a minimum of k records. Let [t, n_t] denote an ordered pair of a unique tuple t and its corresponding number of records n_t. We say that <S, k> is violated if there is [b, r_b] such that $r_b < k$, for some S-projected tuple b. The k-anonymity required on shield attributes helps user to protect privacy without over generalizing the tuple. As for example, consider Table 3a with a given anonymity requirement <{AC, Age, Zip}, 3>. Note that each row represents a unique tuple projected on the shield. Rows 1, 5 and 6 violates the given anonymity requirements with the number of records lower than three. However, Table 3d contains four distinct tuples, each of which has three or more records. Thus, Table 3d satisfies the given anonymity requirement.

In general, for a given table, one can define more than one anonymity requirement, each of which can have a different anonymity degree and a shield. In practice, the anonymity requirement is user-specified. If the anonymity degree is too low, the shield may or may not be able to protect the individual identity (e.g., when the projected tuple becomes personally identifiable). On the other hand, if we set the anonymity degree too high, data

may not be informative since almost all tuples would be the same after anonymization [15]. The data privacy is over protected. This k-anonymity requirements are used in many anonymization techniques [8–10,19,20]. Next, we describe our anonymization technique.

3.3. Balancing Generalization with Data Retention in Anonymization

Given a data table and a *k*-anonymity requirements, this section discusses an analytical approach to transforming the data into anonymized data that satisfy the *k*-anonymity requirements and at the same time retains the data from the original as much as possible. In AI (Artificial Intelligence), we can view this problem as a search in a space of all possible generalized tables on all possible attributes. The simplest approach is to search exhaustively for a solution. To improve efficiency, heuristic search can be employed. Our approach relies on two simple heuristics: the number of rows violating the anonymity requirements and the total number of table rows. The interplay between the two heuristics gives a balance between anonymity compliance and optimizing data retention.

3.3.1. Intelligent Anonymity Balance (IAB) Algorithm

We now briefly describe our anonymization algorithm, *IAB* (Intelligent Anonymity Balance) as also discussed in [15]. Given a data table T with a set of attributes A and a taxonomy tree for each shield attribute. Without loss of generality, we assume one anonymity requirements R with shield $S \subseteq A$. The basic overview of the IAB algorithm is shown in Algorithm 1.

Algorithm 1 The IAB Anonymization Algorithm

Procedure *IAB Anonymization*
Inputs: T, a table with a set of attributes A, a set of anonymity requirement R with a set of anonymity shield attributes $S \subseteq A$ and corresponding taxonomy trees of each attribute in S.
Output: a generalized table T' of T where T' has a maximum number of rows among all generalized tables of T satisfying R.
1 **For** each violating row and applicable attribute B in S
2 $T' \leftarrow$ generalized table of T on B
3 Add T' in W;
4 **Endfor**
5 **Repeat**
6 Select from W, table T_k that has a maximum number of rows and a non-zero minimum number of rows that violate R
7 **For** each violating row and applicable attribute B in S
8 $T'_k \leftarrow$ generalized table of T_k on B
9 Add T'_k in W;
10 **Endfor**
11 Remove T_k from W
12 **Until** W is empty or no tables in W has a number of rows > number of rows of a table that satisfies R
13 **Return** T^* that has maximum number of rows over all tables in W that satisfy R

The algorithm iteratively generalizes a table on an appropriate attribute using its corresponding taxonomy tree to increase anonymity degree. In Lines 1–4, a generalized table of T on each attribute in S is generated and maintained in set W. Each generalized table keeps track of two key heuristics: the number of rows that violate R and the total number of rows on the table. The former tells how close we are to finding the table that satisfies the anonymity requirements R while the latter measures how much data is preserved. Among generalized tables in W, the algorithm selects a table that has the highest number of rows with the lowest violation number of rows to be further generalized (Lines 5–10). The selected table is removed from W (as shown in Line 11).

The generalization process repeats until there are no more tables left in W or no tables in W has the number of rows > the number of rows of a table that satisfies R. In other words, we stop expanding the search when we find a table that satisfies R or a table that is smaller than the biggest table that satisfies R found so far (even though it violates R). By monotonicity of generalization, further generalization can never grow the table. Therefore, the algorithm only further generalizes the table that is larger than those found to satisfy R so far. However, if a table violates R but is already smaller than the biggest table found so far to satisfy R, further generalizing it would not result in a larger table that satisfies R. Thus, the algorithm selects the largest table among the tables in W with no anonymity requirements violation.

Note that it is possible to have more than one of such table of the same size. In such a case, the algorithm selects the first one found as it represents the table that has the least number of generalized steps. In other words, it retains most specific data that are closest to the given data table. Since generalization procedure monotonically decreases the number of rows, our approach uses this property to prune the fruitless path of an exhaustive search. Thus, it finds an *optimal* solution. The optimal solution is that maximizes the information preserved (i.e., the table size) from the original table while hiding desired privacy by satisfying anonymity requirements (i.e., zero violation rows). Therefore, the optimal solution has maximum number of rows (maximum information preservation) that satisfies the anonymity requirement (desired anonymity).

3.3.2. Illustration

We apply the algorithm described in Section 3.3.1 to Table 1 with a given anonymity requirement R = <{Zip, Age, AC}, 3>. Based on the number of records of each row, Table 1 contains 6 rows with number of records less than 3. Thus, these rows, namely Rows 1, 4, 9, 10, 13 and 14, violate R. Generalizing these violating rows of Table 1 on attribute Zip (and also generalizing Zip values for the rest of the rows since their Zip values are siblings of those in the violating rows), we obtained a table as shown in Table 4.

Table 4. A generalized Table after generalizing T on the Zip attribute.

Rows	Sex	Alcohol Cons.	Age	Zip	Weight	Race	Genetic Risk	#Rec.
1,9	F	Med	35	5200 ∗	143	Black	No	1 + 3
2	F	Low	35	5200 ∗	143	Black	No	3
3	M	Med	50	5300 ∗	166	White	No	4
4	M	Low	35	5200 ∗	143	White	No	1
5	M	Med	68	5200 ∗	190	Hispanic	No	3
6	M	High	68	5200 ∗	190	Hispanic	No	4
7	M	Med	44	5200 ∗	166	Asian	Hereditary Thrombophilia	3
8	F	Low	75	5200 ∗	122	Native Hawaiian	No	4
10	F	High	75	5200 ∗	143	White	Hereditary Thrombophilia	1
11	M	High	44	5200 ∗	166	Asian	Hereditary Thrombophilia	3
12	F	Med	38	5200 ∗	166	White	No	4
13	F	High	35	5200 ∗	122	American Indian	Hereditary Thrombophilia	1
14	F	Low	35	5200 ∗	122	American Indian	Hereditary Thrombophilia	2
15	M	Med	38	5200 ∗	122	White	No	3
16	F	Med	22	5400 ∗	180	Native Hawaiian	No	3
17	M	High	20	5400 ∗	180	American Indian	L.O. Alzheimer's	3
18	M	Low	21	5400 ∗	180	Black	Parkinson's	3
19	F	High	24	5400 ∗	180	Asian	Hereditary Thrombophilia	3
20	M	No	25	5400 ∗	122	American Indian	Celiac	3
21	M	No	26	5400 ∗	166	Asian	L.O. Alzheimer's	3
22	F	High	23	5400 ∗	180	Asian	Parkinson's	3

As shown in Table 4, Row 1 and Row 9 can be merged to the first row of the resulting table. Rows 4 and 14 can be combined to satisfy R as a unique tuple from Shield attributes, i.e., (Low, 35, 5200∗) has three records. However, the two rows cannot be merged. Therefore,

the resulting Table 4 has reduced number of violating rows to two (i.e., Rows 10, 13) with a total number of rows to be 21.

Let $T(n, m)$ denote a generalized table T, where n is the number of rows violating R and m is the number of rows in T. Tables 1 and 4 are represented by $T(6, 22)$ and $T_1(2, 21)$, respectively. The generalization process repeats. The whole process can be viewed as a search starting from $T(6, 22)$ as a root and as shown in Figure 3.

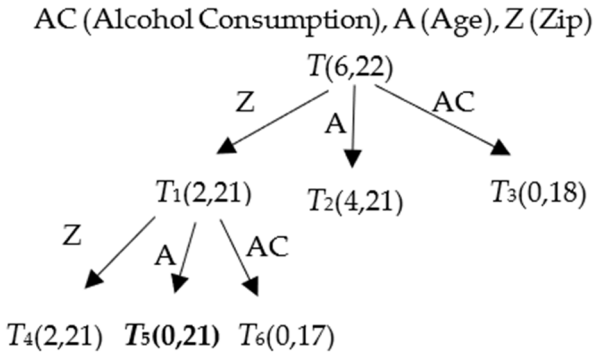

Figure 3. Complete Search Tree of our Approach.

The search starts from the root $T(6, 22)$, i.e., Table 1 (or T) with6 violating rows and a total of 22 rows as shown in Figure 3. We first apply to T, generalization on Zip, Age and AC to obtain tables $T_1(2, 21)$, $T_2(4, 21)$ and $T_3(0, 18)$, respectively. Recall that $T_1(2, 21)$ is actually Table 4.

As seen in Table 4, after merging Rows 1 and 9, we have [(Med, 35, 5200∗), 4]. Hence the violation in these two cases is eliminated. Rows 4 and 14 also have the same shield attribute values after the generalization that is [(Low, 35, 52,000), 3]. Therefore, T_1 has 2 violating rows remained, namely Rows 10 and 13. Moreover, because Rows 1 and 9 merged, the number of rows in T_1 becomes 21. Thus, $T_1(2, 21)$ is obtained. The rest of resulting tables can be obtained similarly.

$T_3(0, 18)$ has zero violations, however we continue to search because there might be a table with more rows and zero violations.

The frontier nodes at this point are $T_1(2, 21)$, $T_2(4, 21)$. They have the same row number, therefore $T_1(2, 21)$ having fewer violating rows is selected to be expanded further. By generalizing $T_1(2, 21)$ on the three attributes we get the tables $T_4(2, 21)$, $T_5(0, 21)$ and $T_6(0,17)$. At this point we stop because, we obtain $T_5(0, 21)$. We do not continue to search even though there are still table with violations such as $T_2(4, 21)$, because none of them have number of rows larger than the current result that is 21. That means we already found the table with the greatest number of rows with zero violations as further generalizing on other tables would only result in a smaller table. Thus, the optimal result of $T_5(0, 21)$ has been found and the algorithm stops.

4. Evaluation and Experiments

This section compares our anonymization approach described in Section 3.3 and evaluate their performances by comparing with two other similar anonymization techniques. Two criteria for evaluating the resulting anonymized table: (1) the table must satisfy a given anonymity requirement with maximum data retention, and (2) the table must be found in timely manner without too much space.

Section 4.1 relates to (1) to evaluate correctness on Table 1, and Section 4.2 relates to (2) by discussion on experiments and results on public datasets. Since the anonymization is viewed as a search problem in this paper, we evaluate our approach by comparing the resulting table(s) with two other search methods: exhaustive search (Method 1) and greedy search (Method 2) [30]. The former is blind search, but the latter is a heuristic search, where the number of violating rows is the heuristic. We will compare results obtained by our approach with Methods 1 and 2.

4.1. On Correctness

Consider Table 1 and the anonymity requirement R is < {Zip, Age, AC}, 3 >. Partial search tree obtained by Method 1 is shown in Figure 4.

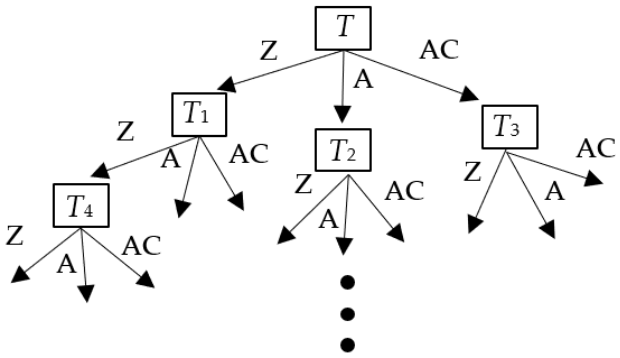

Figure 4. Search Tree obtained from Method 1.

As shown in Figure 4, T corresponds to Table 1 and T_2 is the table obtained from generalizing T on attribute Zip (or Table 4). Method 1 generates all possible generalized tables and choose a table with maximum number of rows among those with zero violation as a solution. The resulting table is shown in Table 5 = T_5 (0, 21) (see Figure 4), which is the same result obtained by our approach. All violations are eliminated (since Rows 4 and 14, Rows 6 and 10, and Rows 11 and 13 has 3, 5, and 4 records, satisfying R, respectively). Furthermore, T_5 has 21 distinct rows as Rows 1 and 9 are merged.

Table 5. Anonymized table from Exhaustive (Method 1) and Our approach.

Row	Sex	Alcohol Cons.	Age	Zip	Weight	Race	Genetic Risk	#Rec.
1,9	F	Med	[35–44]	5200 *	143	Black	No	1 + 2
2	F	Low	[35–44]	5200 *	143	Black	No	3
3	M	Med	[45–64]	5300 *	166	White	No	4
4	M	Low	[35–44]	5200 *	143	White	No	1
5	M	Med	[65–85]	5200 *	190	Hispanic	No	3
6	M	High	[65–85]	5200 *	190	Hispanic	No	4
7	M	Med	[35–44]	5200 *	166	Asian	Hereditary Thrombophilia	3
8	F	Low	[65–85]	5200 *	122	Native Hawaiian	No	4
10	F	High	[65–85]	5200 *	143	White	Hereditary Thrombophilia	1
11	M	High	[35–44]	5200 *	166	Asian	Hereditary Thrombophilia	3
12	F	Med	[35–44]	5200 *	166	White	No	4
13	F	High	[35–44]	5200 *	122	American Indian	Hereditary Thrombophilia	1
14	F	Low	[35–44]	5200 *	122	American Indian	Hereditary Thrombophilia	2
15	M	Med	[35–44]	5200 *	122	White	No	3
16	F	Med	[20–34]	5400 *	180	Native Hawaiian	No	3
17	M	High	[20–34]	5400 *	180	American Indian	L.O. Alzheimer's	3
18	M	Low	[20–34]	5400 *	180	Black	Parkinson's	3
19	F	High	[20–34]	5400 *	180	Asian	Hereditary Thrombophilia	3
20	M	No	[20–34]	5400 *	122	American Indian	Celiac	3
21	M	No	[20–34]	5400 *	166	Asian	L.O. Alzheimer's	3
22	F	High	[20–34]	5400 *	180	Asian	Parkinson's	3

Method 1 (Exhaustive search) and our approach produce the same anonymized table. However, as we will see later that both computational costs are significantly different. A compromising approach between the two is to use a greedy search.

By using the number of rows that violate the anonymity requirement as a heuristic and each time expand on the table with minimum violations (as it has the highest chance

to reach 0 with minimum generalizations) until zero violations. The search tree of Method2 (Greedy approach) is shown in Figure 5.

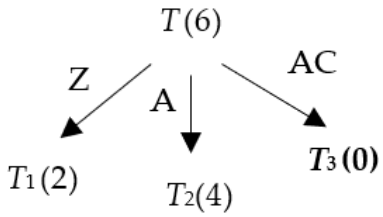

Figure 5. Search Tree for Greedy Approach.

As seen from Figure 5, the result is achieved after generating 3 tables, and when the violations become 0, the search stopped. In Figure 5, each node again is a table annotated by a corresponding heuristic value. Starting from the root node $T(6)$ represents Table 1, with 6 violating rows. We see that after applying a generalization on Zip, Age and AC attribute, the resulting table has 2, 4 and 0 violating rows, respectively.

The greedy solution produces Table 6 = T_3, as a result. As shown in Table 6, there is no violation. However, the number of rows is 19, which is less than our solution which has 21 rows.

Table 6. Resultant table from Greedy Approach.

Rows	Sex	Alcohol Consumption	Age	Zip	Weight	Race	Genetic Risk	#Rec.
1,2	F	Yes	35	52000	143	Black	No	1 + 3
3	M	Yes	50	52000	166	White	No	4
4	M	Yes	35	52003	143	White	No	1
5,6	M	Yes	68	52000	190	Hispanic	No	3 + 4
7,11	M	Yes	44	52003	166	Asian	Hereditary Thrombophilia	3 + 3
8	F	Yes	75	52002	122	Native Hawaiian	No	3
9	F	Yes	35	52003	143	Black	No	2
10	F	Yes	75	52002	143	White	Hereditary Thrombophilia	1
12	F	Yes	38	52000	166	White	No	10
13,14	F	Yes	35	52000	122	American Indian	Hereditary Thrombophilia	1 + 2
15	M	Yes	38	52003	122	White	No	3
16	F	Yes	22	54004	180	Native Hawaiian	No	3
17	M	Yes	20	54000	180	American Indian	L.O. Alzheimer's	3
18	M	Yes	21	54001	180	Black	Parkinson's	3
19	F	Yes	24	54003	180	Asian	Hereditary Thrombophilia	3
20	M	Any	25	54000	122	American Indian	Celiac	3
21	M	Any	26	54001	166	Asian	L.O. Alzheimer's	3
22	F	Yes	23	54003	180	Asian	Parkinson's	3

However, the resulting table from Method 2 (Greedy search) is correct in that it satisfies the R. Therefore, even though Method 2 satisfies (2) it fails (1). Our proposed approach on the other hand satisfies both (1) and (2). Now we will show the performance results (i.e., Section 4.2) as reported in [15].

4.2. On Performances

To demonstrate the effectiveness of our method, we experiment with the public heart disease datasets [31] collected from three different health organizations: Cleveland Clinic Foundation (dataset 1), Hungarian Institute of Cardiology Budapest (dataset2) and V.A. Medical Center, Long Beach, CA (dataset 3). In each data set, we select six most pertinent

attributes for our purpose to illustrate privacy protection of our anonymization approach. For the same reason, we also add the Zip attribute for our experiments giving a total of seven attributes. Figure 6 summarizes the attributes of the three data sets with their corresponding attribute values along with each data set size.

Attributes	Values
Age	Numeric
Smoker	1 = yes, 0 = no
Sex	M, F
Chol (cholestoral)	Numeric (mg/dl)
BS (blood sugar)	1 if BS > 120 (mg/dl), else 0
Heart disease	Present or Abcent
Zip	*ddddd* where *d* is any digit

Data Sets	# Rows
1: Cleveland	303
2: Hungarian	294
3: VA	200

Figure 6. Summary of the Three Data Sets.

The mechanism to anonymize the data relies on data generalization based on the taxonomy of data of each attribute. Here the taxonomy trees for relevant attributes are shown in Figure 7.

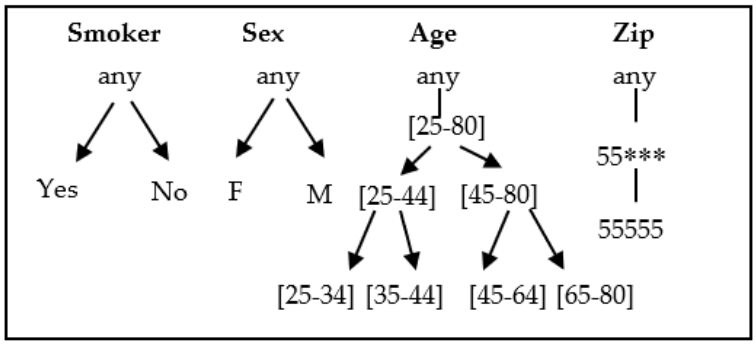

Figure 7. Taxonomy Trees.

Note that a combination of the attributes, selected in Figure 6, can be used to re-identify an individual heart patient. Recall that our method aims to quickly find a solution of an anonymized table that satisfies an anonymity requirement and that it maximally preserves the original data. To better understand how our method performs with respect to the trade-off among each criterion (i.e., data preservation, privacy protection, and efficient solution), we compare our method with two other methods that solve a problem focusing on a single criterion.

Method 1, that is the exhaustive search, aims to find a solution, satisfying anonymity requirements, with maximum information preservation (i.e., retaining the greatest number of data rows), whereas Method 2, that is greedy search, aims to find a solution satisfying anonymity requirements most efficiently. In terms of search, Method 1 exhaustively searches for a solution that has a maximum number of rows in the table, while Method 2 is a greedy search for a table with no rows violating the anonymity requirements. See more details on search algorithms in [30].

4.2.1. Comparisons on Single Shield

Shield attributes in this experiment are Age, Sex, Smoker and Zip and a given anonymity requirement is <{Zip, Smoker, Sex, Age}, 5>. We evaluate in terms of three metrics: number of generalizations, number of table rows, and time. For the number of generalizations, we measure the total number of generalizations applied during the search

for a solution. It indicates the degree of privacy protection. The more generalizations we use, the table becomes more anonymous (but less data preservation). Each generalization transforms a table into a new table. However, when generalizing a table on multiple attributes, the order of the attribute applied for generalization does not affect the resulting table. For example, generalizing a table on attribute *age* then generalizing the resulting table on attribute Sex gives the same table as first generalizing a given table on Sex then generalizing the resulting table on attribute Age. Hence, we label the tables that have the same generalizations as duplicate and only keep one of the tables. The experiment results on total number of generalizations are shown in Figure 8.

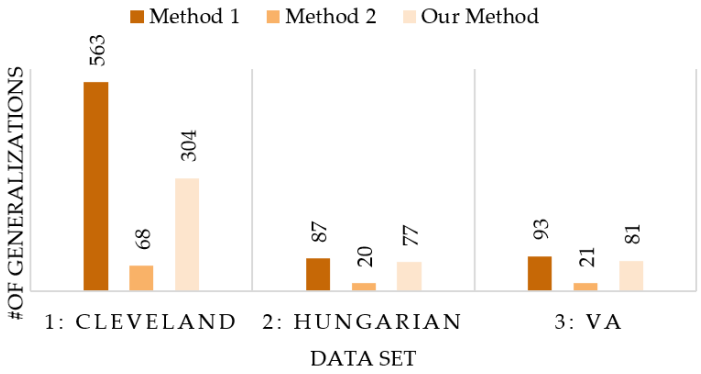

Figure 8. Total Number of Generalizations.

As seen in Figure 8, although each method finds a solution that satisfies anonymity requirements, Method 2 uses optimal number of generalizations in all of the three data sets. This is as expected because the number of generalization steps effect how quickly we can find the solution. On the other hand, Method 1 has the highest number of generalizations in all the three data sets as expected. This can be explained by the fact that Method 1 aims to maximize the data information and thus, it searches over all possible generalized tables for the best solution giving the highest number of rows. On the other hand, the results for our method are in between because it is a trade-off solution that compromises among the three criteria.

The second metric is the number of (distinct) rows that the solution table has. As shown in Figure 6, initially data sets 1–3 has 303, 294 and 200 rows, respectively. Number of rows measures the quality of the result in terms of information preservation. The more distinct row the table has the more original information is preserved. The comparison results are shown in Figure 9. As shown in Figure 9, our method and Method 1 produce the solutions with the same number of rows in all the three data sets. In fact, both obtained an anonymity-complied solution with optimal number of rows. However, as observed in Figure 8, our method uses less effort in terms of the number of generalizations applied. This favors our method in that it takes less work (i.e., number of generalizations) and yet it retains optimal information (i.e., number of rows). The third metric is time that each method takes to find its anonymity-complied solution. Figure 10 shows the comparison results.

As expected, Method 2 has the minimum time as its design (since Method 2 greedily searches for the solution and returns once it finds a solution, see Section 4.1) and Method 1 has the maximum time in finding the solution in all the three data sets. This is because the time is associated with the effort in generalization and thus, the number of generalizations. On the other hand, our method gives a compromised solution in that it is relative fast to find a solution and also retains the high number of data rows.

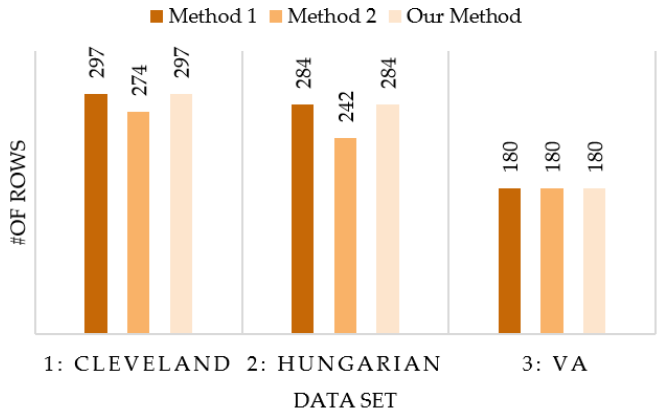

Figure 9. Number of Rows.

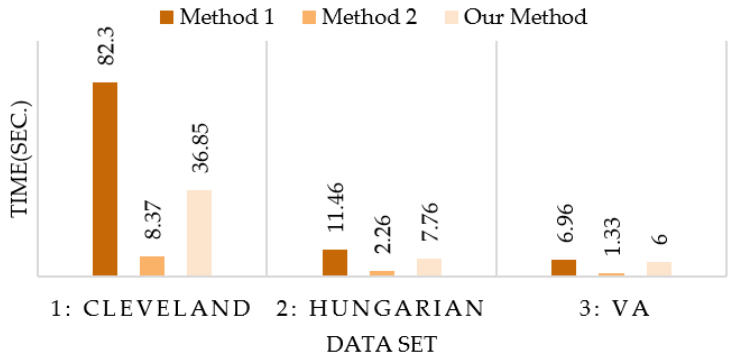

Figure 10. Total Time in Seconds.

4.2.2. Comparisons on Varying Anonymity Requirements

Given a fixed k with varying shield attributes on the anonymity requirement, the biggest factor to both number of generalizations and total time is the selected shield attributes. The taxonomy trees of the attributes and number of selected attributes both effect the results.

Intuitively, the more attributes the shield has the more alternatives for generalization there are. Similarly, if the shield attributes have higher depth of taxonomy trees, there will be more generalizations. As also discussed in [15], to demonstrate that our method still performs well on various shields, we experimented with different shield set sizes and attribute on the same data set, namely dataset 1(Cleveland). The results are shown in Table 7. As shown in the top partition row of Table 7, the anonymity requirement with the greatest number of shield attributes produces the highest number of generalizations in all methods. In the next two partition rows, between the two Anonymity Requirements with three attributes, the Anonymity Requirement (<{Zip, Smoker, Age}, 5>) in the third partition row produces a greater number of generalizations than those produced by the Anonymity Requirement (<{Zip, Smoker, Sex}, 5>) for all methods. This is as expected because the taxonomy tree of Age is larger than that of Sex. In fact, the size of the taxonomy tree of the shield attribute can influence the number of generalizations more than the number of attributes in the shield. As shown in Table 7, the Anonymity Requirement (<{Zip, Smoker, Sex}, 5>) (second partition row) has higher number of attributes than the Anonymity Requirement (<{Zip, Age}, 5>) (last partition row) and yet it produces smaller number of generalizations. This is because the size of taxonomy tree of Age is deeper than those of Sex and Smoker.

Table 7. Experiments with Different Anonymity Requirements.

	Anonymity Requirement	#Rows	#Generalizations
Method 1		**297**	**563**
Method 2	<{Zip, Smoker, Sex, Age}, 5>	**274**	**68**
Ours		**297**	**304**
Method 1		**303**	**11**
Method 2	<{Zip, Smoker, Sex}, 5>	**303**	**6**
Ours		**303**	**6**
Method 1		**298**	**93**
Method 2	<{Zip, smoker, Age}, 5>	**274**	**11**
Ours		**298**	**51**
Method 1		**298**	**14**
Method 2	<{Zip, Age}, 5>	**274**	**12**
Ours		**298**	**14**

In all cases of varying shields on the anonymity requirements, Method 1 (Method 2) generates the highest (lowest) number of generalizations, while ours is in between as it is designed to balance the trade-off between privacy protection (i.e., generalizations) and data preservation (i.e., rows). Our method aims to obtain an anonymized table with maximum information preservation by generating only required amount of generalization. As shown in Table 7, comparing the number of rows of the resulting tables generated by all methods using varying shields on the anonymity requirements, ours and Method 1 generates a maximum number of rows, while results of Method 2 are slightly lower in all but one case.

In general, Method 1 finds the anonymity-complied table that has a maximum number of distinct rows by searching through all possible generalizations. Thus, the search is exhaustive and optimal solution (i.e., an anonymity-compliant generalized table with maximum number of rows) is guaranteed. If there are multiple tables with the same number of rows, the first solution found is selected, as it would have the least generalizations (less time). Even though Method 1 generates a solution that retains maximum information preservation, its exhaustive search that requires many generalizations may not be desirable in practice.

On the other hand, Method 2 finds an anonymized table by greedily searching for a generalized table that has a minimum number of anonymity violations (i.e., zero). Using a heuristic on the number of violating rows, Method 2 finds a solution without going through all possible generalized tables. Thus, its search is more efficient than Method 1. However, finding the optimal solution (i.e., a generalized table with zero violation) is not guaranteed.

Our method combines Methods 1 and 2 by quickly finding a generalized table that has zero anonymity violation as well as being the most informative table (i.e., having maximum number of distinct rows like Method 1). The method is heuristic using the above two evaluation metrics and thus, saves time compared to an exhaustive Method 1. Furthermore, when a generalized table with no violation is found, further generalization is not necessary, as by the monotonicity property of generalization, generalization will not produce a table with a higher number of rows. The reason is that generalization creates rows with common values and therefore it always maintains or shrinks the table size. Our method uses the monotonicity property to reduce search time and guarantees optimal solution (i.e., an anonymity-compliant generalized table with a maximum number of rows). The experimental results obtained are consistent with the design of each of the above methods.

5. Post Anonymization Analytics

After an original data table has been anonymized, the table is ready to be released for public or sharing among appropriate parties. However, in case when the data that are privacy critical, further analyzing the anonymized table can be pursued. In this paper, we examine the resulting anonymized table obtained by our technique as described in

Section 3.3. To illustrate, consider the anonymized Cleveland dataset 1 (as obtained in Section 4). By applying the approach described in Section 3.1 using the Longpre et al.'s entropy-based measure, on the anonymized Cleveland dataset 1, we can further assess the effectiveness of the anonymization.

Table 8 shows the overall results of this post anonymization analytics where each row indicates vulnerability to information leakages (i.e., normalized average information loss) given an attacker obtains information on the corresponding attribute in each column.

Table 8. Post anonymization analytics of Cleveland Dataset 1.

Query Attribute / Avg. Info. Loss	Age	Zip	Smoker	Sex	Chol.	BS
Before Anonymization	0.61	0.99	0.12	0.11	0.85	0.07
<{Age, Zip}, 5>	0	0.4	0.12	0.11	0.85	0.07
<{Age, Zip, Smoker, Sex}, 5>	0	0.4	0	0.11	0.85	0.07

As shown in Table 8, the first row gives the vulnerability "Before Anonymization". We see that the Zip attribute is most vulnerable as it leaks most information of 0.99. Next is Cholesterol and Age that leaks 0.85 and 0.61, respectively. These results can help partially select potential shield attributes although in practice, they are user-specified.

Second row shows the vulnerability on anonymized table that is in compliance with the requirement of 5-anonymity on the shield attribute set {Age, Zip}, as denoted by <{Age, Zip}, 5>. This shield attributes agree with the vulnerability assessment for the most part and omit Cholesterol as it may not be acquired easily through binary query. As shown on second row of Algorithm 1, the Age attribute is now not leaking any information. Each record now has the same Age value as a result of anonymization (i.e., generalization).

Note that only attributes that are on the shield (i.e., Age and Zip) have reduced average leakages (e.g., Age's loss from 0.61 to 0, and Zip's loss from 0.99 to 0.4). This is as expected since the generalization only can be applied to those attributes and causes a value change. The rest of other attribute values stays the same after anonymization. The information disclosure based on that attribute stays the same.

Similarly, on the third row of Algorithm 1, the anonymization satisfies <{Age, Zip, Smoker, Sex}, 5>. Compared with Row 2, two more attributes (i.e., Smoker and Sex) are added to the anonymity requirement, leakages on Age and Zip remain the same (i.e., they are generalized to the same level as previous case). However, leakages on Smoker reduce to 0 but leakages on Sex remain the same. This means that anonymization process generalizes on Smoker attribute.

The overall analytics on leakages after post anonymization indicate that the anonymization is effective since all shield attributes either maintain the same or reduced average information loss after the anonymization. Note that the average information loss is reversed from anonymity. When the information loss is high (i.e., an attacker obtains more information), the anonymity is low because the attacker can use the information to better distinguish individuals for re-identification. Therefore, we can use this measure to link to anonymity.

6. Conclusions

Smart health has significant impacts on healthcare and wellness. However, it also poses privacy threats to users. As health data get larger and become more accessible to multiple parties, users lose more control of their data that increasingly become vulnerable to attacks. Furthermore, the challenge is not only to protect the data but also to ensure that the shared data are sufficiently informative. Increasing users' anonymity is a basic remedy as anonymity increases indistinguishability. The more indistinguishable people are the more anonymous they become and thus, their information and identity are better concealed.

This paper presents an approach to health data analytics focusing on anonymity for privacy protection. The approach is applicable to both data producers (e.g., use of fitness trackers, or glucose and heart rate monitors) as well as data consumers (e.g., weight loss application services, healthcare professionals) to safeguard a given health data set from information leakages and re-identification. A common concept relies on making data anonymous.

An analytical approach is proposed to (1) identifying attributes susceptible to information leakages by using entropy-based measure to analyze information loss, (2) transforming the data into a more anonymous form by generalization using attribute hierarchies, and (3) anonymization that balances anonymity requirements and optimal informativeness by an automated Artificial Intelligence search using two simple heuristics. Unlike existing techniques, our anonymization approach preserves maximum information by avoiding extensive generalizations yet still complies with the anonymity requirements. The proposed anonymization follows *k*-anonymity; therefore, it inherits the limitations of *k*-anonymization as discussed in [21]. We describe and illustrate the detailed approach and analytics including pre and post anonymization analytics. We have conducted experiments to evaluate effectiveness of our anonymization approach. The results obtained show that our approach balances the trade-off between preserving privacy and retaining maximum information with efficient computational cost. Future work includes a framework designed to integrate all different measures to improve anonymization techniques as well as to better increase anonymity and protect privacy. The added metrics will help further the analysis of the anonymized data in terms of privacy. That way, we aim to get a better understanding of what needs to be improved for anonymization or how successful the anonymization is.

Author Contributions: Conceptualization, R.H.; Investigation, S.A.; Methodology, Both; Supervision, R.H.; Validation, R.H.; Data Curation, S.A.; Writing–original draft preparation, S.A.; Writing–review & editing, R.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in [UCI Machine Learning Repository] at [http://archive.ics.uci.edu/ml/index.php], reference number [31].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fitbit LLC. *Innovation Meets Motivation*; Fitbit Official Site for Activity Trackers & More: San Francisco, CA, USA, 2021; Available online: https://www.fitbit.com/global/us/home (accessed on 10 September 2021).
2. Alivecor, Inc. *Kardiamobile*; Alivecor: Mountain View, CA, USA, 2021; Available online: https://www.kardia.com (accessed on 10 September 2021).
3. Dexcom, Inc. Dexcom Continuous Glucose Monitoring, Dexcom. 2020. Available online: https://www.dexcom.com/continuous-glucose-monitoring (accessed on 3 April 2020).
4. Abbott Laboratories. *FreeStyle Libre*; Abbott Laboratories: Chicago, IL, USA, 2018; Available online: https://www.freestylelibre.us/ (accessed on 3 April 2020).
5. Omron Healthcare, Inc. Healthcare Wellness & Healthcare Products, Heartguide Wearable Blood Pressure Monitor, Omron Healthcare. 2020. Available online: https://omronhealthcare.com/products/heartguide-wearable-blood-pressure-monitor-bp8000m/ (accessed on 3 April 2020).
6. Phaneuf, A. Latest Trends in Medical Monitoring Devices and Wearable Health Technology, Business Insider. 2020. Available online: https://www.businessinsider.com/wearable-technology-healthcare-medical-devices/ (accessed on 3 April 2020).
7. *Dna Genetic Testing & Analysis*; 23andMe, Inc.: Sunnyvale, CA, USA, 2021. Available online: https://www.23andme.com/ (accessed on 10 September 2021).
8. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. -Based Systems.* **2002**, *10*, 557–570. [CrossRef]
9. Samarati, P. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 1010–1027. [CrossRef]
10. Sweeney, L. Achieving k-anonymity Privacy Protection Using Generalization and Suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 571–588. [CrossRef]
11. Andersson, C.; Lundin, R. On the Fundamentals of anonymity metrics. In *IFIP International Summer School on the Future of Identity in the Information Society*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 325–341.

12. Arca, S.; Hewett, R. Is entropy enough for measuring privacy? In Proceedings of the 7th Computational Science & Computational Intelligence, Las Vegas, NV, USA, 16–18 December 2020; pp. 1335–1340.
13. Bezzi, M. An entropy based method for measuring anonymity. In *Proceeding of the 3rd International Conference on Security and Privacy in Communications Networks and the Workshops-SecureComm*; IEEE: Piscataway, NJ, USA, 2007; pp. 28–32.
14. Longpr, L.; Kreinovich, V.; Dumrongpokaphan, T. Entropy as a Measure of Average Loss of Privacy. *Thai J. Math.* **2017**, 7–15. Available online: http://thaijmath.in.cmu.ac.th/index.php/thaijmath/article/viewFile/3002/918 (accessed on 21 October 2021).
15. Arca, S.; Hewett, R. Privacy in smart health. In Proceedings of the 11th International Conference on Advances in Information Technology (IAIT 2020), Bangkok, Thailand, 1–3 July 2020; pp. 1–8.
16. Fung, B.C.M.; Wang, K.; Yu, P.S. Anonymizing Classification Data for Privacy Preservation. *IEEE Trans. Knowl. Data Eng.* **2007**, 711–725. [CrossRef]
17. Fung, B.C.M.; Wang, K.; Yu, P.S. Top-down specialization for information and privacy preservation. In Proceedings of the 21st International Conference on Data Engineering; IEEE Computer Society: Piscataway, NJ, USA, 2005; pp. 205–216.
18. Hundepool, A.; Willenborg, L. μ-and τ-argus: Software for statistical disclosure control. In Proceedings of the Third International Seminar on Statistical Confidentiality, 1996. Available online: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=118A75C0CCF39B4AF2BDB65E7B52C147?doi=10.1.1.132.3621&rep=rep1&type=pdf (accessed on 21 October 2021).
19. LeFevre, K.; DeWitt, D.; Ramakrishnan, R. Incognito: Efficient full-domain k-anonymity. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 14–16 June 2005; pp. 49–60.
20. LeFevre, K.; DeWitt, D.; Ramakrishnan, R. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06); IEEE: Piscataway, NJ, USA, 2006; p. 25.
21. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data TKDD* **2007**, *1*, 3-es. [CrossRef]
22. Sweeney, L. *Datafly: A system for providing anonymity in medical data. Database Security XI*; Springer: Boston, MA, USA, 1998; pp. 356–381.
23. Wang, K.; Yu, P.S.; Chakraborty, S. Bottom-up Generalization: A Datamining Solution to Privacy Protection. In Proceedings of the 4th 22nd International Conference on Data Mining, Brighton, UK, 1–4 November 2004; pp. 249–256.
24. Majeed, A. Attribute-centric Anonymization Scheme for Improving User Privacy and Utility of Publishing e-health Data. *J. King Saud Univ.-Comput. Inf. Sci.* **2019**, *31*, 426–435. [CrossRef]
25. Liang, Y.; Samavi, R. Optimization-Based k-anonymity Algorithms. *Comput. Secur.* **2020**, *93*, 101753. [CrossRef]
26. De Montjoye, Y.; Radaelli, L.; Singh, V.K. Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata. *Science* **2015**, *347*, 536–539. [CrossRef] [PubMed]
27. Rocher, L.; Hendrickx, J.M.; de Montjoye, Y. Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models. *Nat. Commun.* **2019**, *10*, 1–9. [CrossRef] [PubMed]
28. Al-Zubaidie, M.; Zhang, Z.; Zhang, J. PAX: Using Pseudonymization and Anonymization to Protect Patients' Identities and Data in the Healthcare System. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1490. [CrossRef] [PubMed]
29. Dam, T.; Kieseberg, P.; Zeppelzauer, M. k-Anonymity in Practice: How Generalisation and Suppression Affect Machine Learning Classifiers. *Comput. Secur.* **2021**. [CrossRef]
30. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice Hall: Hoboken, NJ, USA, 2010.
31. Detrano, R.; Janosi, A. UCI Repository of Machine Learning Databases. Available online: https://archive.ics.uci.edu/ml/datasets/heart+Disease (accessed on 3 March 2020).