TRANSACTIONS ON DATA PRIVACY 10 (2017) 27–59

Enhancing the Utility of Anonymized Data by Improving the Quality of Generalization Hierarchies

Vanessa Ayala-Rivera*, **Patrick McDonagh****, **Thomas Cerqueus*****, **Liam Murphy***, **Christina Thorpe***

*Lero@UCD, School of Computer Science, University College Dublin, Ireland.

**CTO Reliability and Eco-Environmental Engineering Team, Alcatel-Lucent, Blanchardstown, D15, Ireland.

***R&D team, Lengow, Nantes, France.

E-mail: vanessa.ayala-rivera@ucdconnect.ie, patrick.mcdonagh@alcatel-lucent.com, thomas.cerqueus@lengow.com, liam.murphy@ucd.ie, christina.thorpe@ucd.ie

Received 09 August 2016; received in revised form 26 January 2017; accepted 04 March 2017

Abstract. The dissemination of textual personal information has become an important driver of innovation. However, due to the possible content of sensitive information, this data must be anonymized. A commonly-used technique to anonymize data is generalization. Nevertheless, its effectiveness can be hampered by the Value Generalization Hierarchies (VGHs) used as poorly-specified VGHs can decrease the usefulness of the resulting data. To tackle this problem, in our previous work we presented the Generalization Semantic Loss (GSL), a metric that captures the quality of categorical VGHs in terms of semantic consistency and taxonomic organization. We validated the accuracy of GSL using an intrinsic evaluation with respect to a gold standard ontology. In this paper, we extend our previous work by conducting an extrinsic evaluation of GSL with respect to the performance that VGHs have in anonymization (using data utility metrics). We show how GSL can be used to perform an a priori assessment of the VGHs' effectiveness for anonymization. In this manner, data publishers can quantitatively compare the quality of various VGHs and identify (before anonymization) those that better retain the semantics of the original data. Consequently, the utility of the anonymized datasets can be improved without sacrificing the privacy goal. Our results demonstrate the accuracy of GSL, as the quality of VGHs measured with GSL strongly correlates with the utility of the anonymized data. Results also show the benefits that an a priori VGH assessment strategy brings to the anonymization process in terms of time-savings and a reduction in the dependency on expert knowledge. Finally, GSL also proved to be lightweight in terms of computational resources.

Keywords. Privacy, Data Publishing, Data Quality, Generalization Hierarchies, Data Semantics

1 Introduction

Nowadays, information systems collect considerable amounts of data about individuals. Within this data, there is a large amount of categorical personal information, such as users' preferences, physical conditions, and leisure activities. This data has become a valuable resource for businesses in various sectors, such as insurers or retailers. By mining and analyzing microdata (i.e., records containing information about individuals) companies have not only unlocked new sources of economic value but also have provided fresh insights

into science. In order to obtain benefits from this data, companies typically must share it with third parties. However, microdata may contain sensitive information about individuals (e.g., religious beliefs, medical conditions) that could potentially cause harm to the involved parties if disclosed. For example, organizations may suffer from fines, negative publicity, or other sanctions. Similarly, individuals may suffer from identity theft or discrimination [60]. Thus, this data must be anonymized before being shared for analysis. To achieve this goal, Privacy-Preserving Data Publishing (PPDP) provides methods for sharing data without compromising the confidentiality of individuals, while also trying to retain the utility of the data for a variety of tasks [12]. As there can be diverse data recipients, the anonymized data should be useful enough to perform different analysis tasks (e.g., data mining, decision-support systems, query answering), and not be tailored to a specific one. One commonly-used technique for anonymizing data is generalization [35, 57, 77, 79]. It consists in replacing the values of an attribute in a dataset with less precise, but semantically consistent, alternative values. This transformation reduces the probability of reidentifying the individuals contained in the published anonymized datasets (e.g., generalizing *"pediatrician"* with *"doctor"* to safeguard the profession of a person).

A common prerequisite of generalization algorithms is the use of Value Generalization Hierarchies (VGHs) [61], which are tree-like structures that drive the anonymization process. They contain the set of transformations that an attribute can undergo. Traditionally, VGHs are created and evaluated by data publishers. In this context, a data publisher is any individual or organization (e.g., practitioners, researchers, data sanitizers) who is involved in the sharing of data and seeks to disseminate it in a safe and useful manner (hereinafter referred to as *users*). Users usually follow an iterative process for assessing the right coverage of the concepts and appropriate details to represent in the VGHs. This process can yield multiple candidates of VGHs (per attribute) which can be used for anonymization. The users must then decide which VGH will be used to anonymize each attribute. This decision plays a crucial role in the utility remaining in the anonymized data and ultimately in the precision of the analysis performed, as both can diminish if poorly-specified VGHs are used [8]. All these tasks are often performed manually, relying on the users' judgment.

A key problem of this practice is that the quality of the VGHs is assessed in an informal and subjective way. Despite this, the "correctness" (in terms of semantic consistency) of the VGHs that feed generalization algorithms is an aspect that has been rarely questioned in the PPDP literature [11, 19]. This is because it is normally assumed that users are fully capable of providing the adequate domain expertise to the VGHs based on their own knowledge and experience [45, 78]. To mitigate possible issues, knowledge engineers are often involved in the evaluation process. Nevertheless, the process may become expensive due to the limited availability of subject-matter experts [28, 80], and the intensive and time-consuming manual labor it requires. In any case, the decision about the quality of VGHs is normally based on the subjective opinion of a single individual, therefore corresponds to only a single interpretation of a domain. Clearly, the design of VGHs for anonymization is a challenging task and the current practices are not effective [11, 72].

Considering the above challenges, our research work has centered on developing techniques to evaluate in a quantifiable, objective, and automatic way, the quality of VGHs for categorical data with the aim of improving their effectiveness for anonymizing data. In our previous work [11] we presented the Generalization Semantic Loss metric (GSL), which assesses the quality of a VGH with respect to its taxonomic organization and semantic consistency. In that initial work, we performed an intrinsic evaluation of GSL by judging the quality of single-attribute VGHs with respect to a reference ontology. In this paper, we extend our previous work by conducting an extrinsic evaluation. That is, the quality of

the VGHs is evaluated with respect to their effectiveness in anonymization and the accuracy of the analysis results that can be obtained from the anonymized data. We also assess the feasibility of using GSL to perform an a priori evaluation of VGHs for anonymization. This strategy would assist data publishers to make informed decisions about which VGHs would produce more useful anonymized solutions.

In particular, the research contributions in this paper are as follows:

- An extended description of our proposed metric (GSL) for the quality evaluation of categorical VGHs, including its extension to be suitable to the multi-attribute VGH scenarios.
- A comprehensive extrinsic evaluation of GSL, consisting of a prototype and a set of experiments to assess: (1) the accuracy of using GSL to estimate the effectiveness of the VGHs for producing good quality anonymized data, measured in terms of task-independent (general-purpose utility metrics) and task-dependent metrics (clustering); (2) the benefits that an a priori evaluation of categorical VGHs can bring to the anonymization process, in terms of time-savings; and (3) the costs of using GSL, in terms of computational resources.
- Key findings that could serve as guidelines for data publishers to use GSL, as well as the conditions under which the metric can be more useful.

The paper is organized as follows: Section 2 provides the background and related work. Section 3 explains the internal workings of GSL. Section 4 describes the experimental setup. Section 5 explains the evaluation criteria used in our experiments, while Section 6 discusses the experimental results. Finally, Section 7 presents our conclusions and future work.

2 Background and Related Work

In this section, we first recall some of the most relevant concepts in PPDP which are necessary to understand the rest of the paper. Next, we review the state-of-the-art work in the area of creation and evaluation of generalization hierarchies, with a special emphasis on their application in data anonymization.

2.1 Privacy-Preserving Data Publishing

In the most basic form of PPDP, a microdata table is composed of tuples defined over a set of attributes, where each tuple corresponds to a record that is associated with an individual (e.g., healthcare, educational, criminal records). The attributes are typically classified in four categories according to the information they contain [27]: Identifiers (IDs), which are attributes that explicitly identify individuals (e.g., name, unique identifying numbers); Quasi-Identifiers (QIDs), which are attributes that could be linked to external information and lead to the re-identification of people in published anonymized datasets (e.g., profession, nationality); Sensitive (SAs), which are attributes that represent sensitive information (e.g., disease, salary); and Non-Sensitive, which are attributes that do not fall into the previous categories. In order to protect the data from disclosure, the IDs are removed and the QIDs are modified with the aim of decreasing the probability of re-identification.

Many anonymization techniques proposed in the literature use generalization as the operation to transform the QIDs. One advantage of generalization is that (unlike perturbation techniques that apply noise to data), it preserves the truthfulness of the data [68].

Furthermore, this technique enables the definition of boundaries during the anonymization process. That is, users can ensure that generalization intervals are disjoint, which is a desirable property, as overlapping intervals can increase the difficulty of analyzing the anonymized data. These properties (truthfulness and interval disjointness) are guaranteed to be achieved by using well-defined VGHs. An example VGH is shown in Fig. 1. The leaves at level 0 (L0) represent the distinct values of an attribute in the original dataset. The ancestors at upper levels (L1 to L3) correspond to the candidate values used for the generalizations. The root node (at L3) corresponds to the maximum generalization (or full suppression) of a value.

Even though the aim of PPDP is to share anonymized data for legitimate (non-privacy-violating) purposes, a key assumption in this area is that attackers can also be found among the data recipients, who will intend to uncover sensitive information about individuals. Thus, generalization is used in conjunction with a privacy model (e.g., k-anonymity, ℓ-diversity, t-closeness) [27] with the aim of providing formal privacy guarantees. In our work, we have focused on k-anonymity, a widely-adopted privacy model that has proven to be practical in real-world systems [10, 68] (despite having some limitations [40, 46]). Moreover, it constitutes the basis of newer anonymization techniques and enhanced privacy models in diverse contexts [12, 33, 69]. k-Anonymity works by altering the QID attributes to form equivalence classes (EQs), which are groups of records sharing the same QID values. The aim is to make each record indistinguishable from a group of at least k-1 other records [68].

Example 1. As an illustrative example of achieving k-anonymity by generalization, let us consider Table 1 showing a table with socio-economic records. Among the attributes, *name* is the ID, *occupation* is the QID, and *salary* is the SA. Suppose the desired privacy goal is k=3. In order to achieve it, the ID is removed and the QID is generalized two levels of the VGH shown in Fig. 1. Table 2 shows a 3-anonymous version of Table 1. Note that the generalization created two EQs (Doctor, Educator). Within each EQ, individuals are indistinguishable from each other.

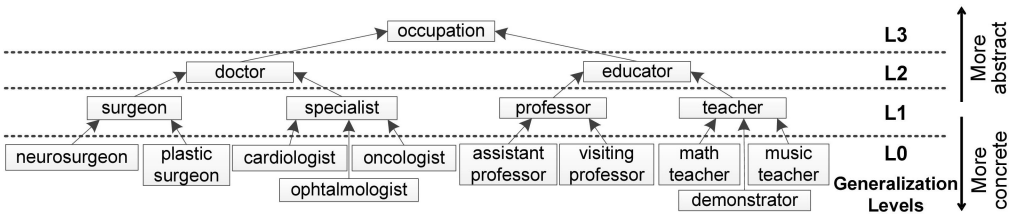


Figure 1: VGH for *occupation*

Table 1: Example socio-economic dataset

	ID	QID	SA
#	Name	Occupation	Salary
1	Ben	Neurosurgeon	300K
2	Emma	Music Teacher	49K
3	Eve	Plastic Surgeon	85K
4	Meg	Oncologist	102K
5	Jim	Assistant Profesor	81K
6	Leo	Math Teacher	46K

Table 2: A 3-anonymous version of Table 1

		QID	SA
#	EQ	Occupation	Salary
1	1	Doctor	300K
3		Doctor	85K
4		Doctor	102K
2	2	Educator	49K
5		Educator	81K
6		Educator	46K

2.2 Creation and Evaluation of VGHs for Data Anonymization

The creation and evaluation of VGHs for numerical attributes have been well-studied in the literature [19, 72]. For numerical attributes, it is relatively easy to evaluate if the data quality has been preserved after anonymization (e.g., by minimizing the size of an interval, or by retaining their statistical properties). In contrast, little research has been done to study the quality of VGHs for categorical data. Previous studies have discussed the role that VGHs play in the utility of anonymized data [19, 41, 54, 72]. They indicate that a good VGH would improve the usefulness of the data, whereas a poor VGH would reduce the quality of the data; hence, the precision of the analysis results obtained from it. Although these works have helped to understand a set of desired properties in VGHs, formal methodologies to assess VGHs are still scarce as VGHs continue to be judged by the users based on their own knowledge and experience [34, 45].

A closely related research area is ontology evaluation, as VGHs could be seen as particular cases of ontologies in which only the *is-a* semantic relationships are considered. Several valuable works have been proposed in this field [16, 29, 65]. However, the direct applicability of those techniques in PPDP is limited as they do not consider the particular characteristics needed by a VGH in the context of data anonymization. For example, those techniques usually validate how well the domain of interest has been covered (i.e., granularity). Nevertheless, in anonymization, a trade-off exists between the granularity and the privacy vulnerability that a VGH should have. This is because, the finer the granularity, the more useful the anonymized data is, but also the more vulnerable it could be to inferences.

More recently, some data privacy works have proposed to use ontologies (instead of VGHs) to anonymize data [25, 49, 59]. However, ontologies' applicability may be limited as they can bring significant restrictions to anonymization. For example: (1) The size of the solution space increases with respect to the number of QIDs and the height of their corresponding VGHs. Due to the complexity of ontologies' graph model, the solution space would substantially increase. Thus, existing anonymization algorithms would not be able to efficiently handle such deep and broad taxonomies (hence, becoming impractical for real-world applications); (2) The fine granularity of ontologies can overexpose information to an adversary such that the anonymized data could still be vulnerable to inference attacks; (3) Ontologies cannot be easily customized to the requirements of the data recipients, whereas VGHs are more flexible and can be adapted to different use cases (e.g., eliminating undesirable generalizations, controlling the level of explicitness). For these reasons, our work only uses ontologies as an external source of knowledge for the evaluation of VGHs; leveraging the fact that many large and consensus ontologies have been made available [22, 50]. Moreover, multiple ontologies can be integrated to complement each other and have a more complete source of knowledge [62, 63] (hence, overcoming the limitation of exploiting a single ontology).

3 Evaluating the Quality of VGHs

In this section, we present the contextual view of our solution applied to the traditional anonymization process. We also motivate the use of a data semantics-oriented approach to evaluate VGHs for categorical data. Finally, we describe the elements involved in the VGH evaluation method including the details about how to compute the GSL score and how it can be used to perform an a priori evaluation and selection of VGHs.

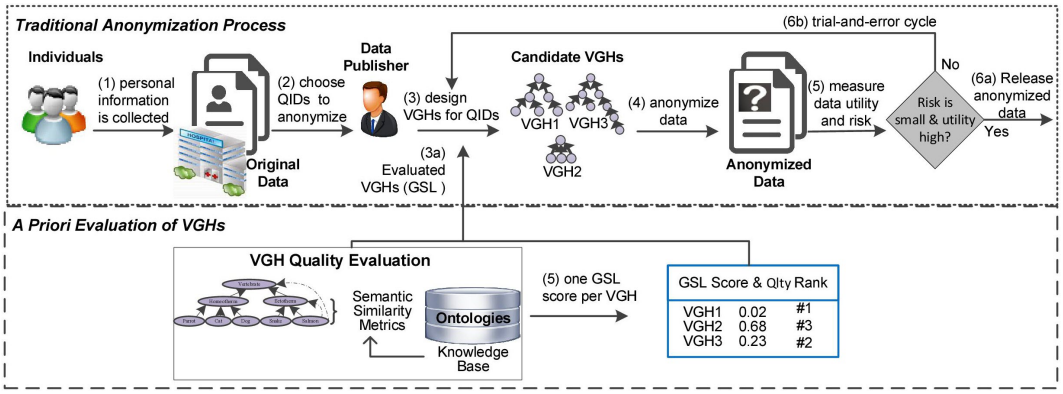


Figure 2: Contextual View of APES in PPDP

3.1 Overview

Our proposed solution consists of a method to quantitatively and objectively evaluate the quality of VGHs for categorical data with the aim of improving their effectiveness for anonymizing data. Fig. 2 depicts the contextual view of our solution in PPDP: (1) An organization (e.g., hospitals, government agencies) collects personal information about individuals and is required to publish it under different circumstances (e.g., research, commercialization). Therefore, this data must be anonymized before being disseminated. (2) Traditional anonymization consists in transforming the attributes that an adversary could link to external information (i.e., QIDs), and which could result in the re-identification of an individual in an anonymized dataset. Thus, the user classifies each attribute of the dataset into IDs, QIDs, or SAs. The QID attributes would then be generalized according to a privacy requirement, an anonymization algorithm, and their associated VGHs. (3) For each QID, the user creates a set of candidate VGHs modeling the given domain. Several candidate VGHs can be considered depending on how the user wants to refine the anonymizations (e.g., to balance the potential utility and privacy of the data). The quality of the VGHs is then evaluated by the user. Traditionally, this task is performed manually and based on the user's own knowledge and experience, so it is time-consuming and error-prone. (4) Once the user is confident about the chosen VGHs, they are used to anonymize the data. (5) After anonymization, the quality of the resulting data is evaluated either by comparing it to the original data or based on the accuracy obtained from a particular application task. Likewise, the disclosure risk is also assessed. (6a) If the utility is acceptable high and the risk is small enough, the data is released. (6b) Otherwise, a new cycle of anonymization starts (Step 3) in which the anonymization settings (including the VGHs) can be tailored depending on the user's objectives.

Considering this, users need to have an efficient and effective manner of assessing the VGHs and make an informed decision about which VGH to use for their publishing scenario. Our proposed solution fits into Step 3, where the VGHs are designed. (3a) We provide users with a method that evaluates the quality of VGHs by capturing (in a score denoted as GSL) the degree of data semantics that VGHs lose in their specification. The lower the GSL score, the less information loss incurred in a VGH. Our method exploits two main elements to carry out the VGH evaluation: a knowledge base and semantic similarity metrics (discussed in Section 3.4.1). The output of our method is one GSL score per evaluated VGH, which enables the users to quantitatively compare the quality of the candidate

VGHs. Based on GSL, the users can select (a priori) the best VGHs (e.g., those ranked #1) to feed the algorithm that anonymizes the data with more guarantees that the chosen VGHs will help to retain the desired level of data usefulness (hence eliminating the need of costly trial-and-error anonymization cycles).

3.2 A Priori vs. A Posteriori Evaluation

As discussed in Section 3.1, the Traditional Anonymization Process (denoted as TAS) is usually iterative. In TAS, several configurations can be tested by the users before selecting an anonymized solution that satisfies their privacy and data utility requirements. Two examples of possible tested scenarios in TAS are:

1. **Evaluating multiple candidate VGHs.** This scenario typically involves various candidate VGHs which have been defined to model the domain of one attribute in the dataset. In order to identify the VGH that would produce the best anonymized solution (e.g., highest utility), the user needs to perform the anonymization of the dataset with each of the candidate VGHs and calculate the utility of the resulting datasets (e.g., using an information metric).

2. **Evaluating different levels of privacy.** In this scenario, different levels of privacy (e.g., k-values) often need to be tested as there is no fixed standard of privacy to perform the anonymization (e.g., a minimum size for a disclosed group). In this case, the user must anonymize the data using each combination of candidate VGHs and privacy goals to find an appropriate solution that achieves a good trade-off between the privacy and the utility of the dataset.

The above scenarios describe situations in which the effectiveness of the VGHs is evaluated using an a posteriori approach. This involved significant effort to test various configurations. To tackle this problem, users can leverage on our proposed metric GSL to perform an A Priori Evaluation and Selection of VGHs. Our strategy (denoted as APES) allows users to estimate the effectiveness of the VGHs (before anonymization) and identify/select the ones that better preserve the semantics of the original data. As a consequence, not only the quality of the resulting anonymized data can be improved but also the efficiency of TAS, as some trial-and-error cycles can be eliminated.

3.3 Preservation of Data Semantics in Categorical Data

In the following paragraphs, we motivate the importance of preserving the data semantics in categorical data, and why our solution uses a data semantics-oriented approach to evaluate the quality of VGHs for categorical attributes.

Motivation 1. *Traditional task-independent metrics do not consider semantics, thus they are limited in reflecting the data distortion for categorical values.*

Data semantics is an implicit property of categorical data. Despite this, traditional data utility measures in PPDP do not usually take it into account. This is because the quantification of the usefulness remaining in the anonymized data is usually performed using metrics that are better suited for numerical attributes than categorical. In the scenario of general-purpose data publishing, two commonly-used types of metrics are: distance-based metrics (e.g., general loss metric [32, 54], generalization height [61, 68]), and distributional

metrics (e.g., discernibility metric [15], average equivalence class size metric [39]). For instance, a typical approach is to transform each categorical value to a numeric one, then the amount of data distortion due to anonymization is determined by the size of the interval in which the original values have been grouped [32]. Even though these types of metrics capture a certain level of data distortion, they do not capture the loss in the meaning of data for categorical values as the same degree of ambiguity is assigned to all values. For example, consider that "heart disease" is anonymized to "cardiopathy" (its synonym). A semantic-based approach would correctly capture that both terms are semantically equivalent, thus the information loss would be zero, as the meaning of the original value is preserved.

Motivation 2. *Classical data mining applications treat categorical data at a syntactic level.*

Some task-specific metrics also suffer from similar issues as those discussed in Motivation 1. For example, data mining techniques (e.g., clustering) typically ignore the inherent semantics of categorical data, which can make the interpretation of the results difficult. Consider classical clustering, where the comparison of the objects is performed by computing the distance over the raw attributes of the objects. This is often performed using measures that are based on equality comparisons and/or frequency tables with the number of occurrences of the terms (e.g., Hamming distance, Chi-Squared) [14]. However, those methods do not consider the semantics of the concepts being compared. For example, consider the terms: *"bus"*, *"car"* and *"bicycle"*. Traditional approaches would consider them as equally dissimilar. However, a semantic-based approach would capture more precisely the degree of similarity between the terms by taking into account their meaning. In this case, that *"car"* and *"bus"* are more similar than *"car"* and *"bicycle"* (or *"bus"* and *"bicycle"*), as they are motorized vehicles.

Motivation 3. *The loss of information might not be monotonic for poorly-specified VGHs.*

The "correctness" of the VGHs that feed generalization algorithms is a property that most of the anonymization algorithms and data utility metrics take for granted. That is, they assume that the specification of the VGHs is semantically consistent (i.e., concepts properly positioned in the VGH based on their semantic proximity and level of abstractness) [46, 61, 68]. Considering this, the loss of information (due to generalization) should increase monotonically as one goes up in the VGH until reaching the root node, where the maximum generalization, and thus the maximum loss of information occurs. However, as it has been motivated in this paper (Section 1), the semantic consistency of a VGH is not always guaranteed (e.g., lack of expertise of users, limited availability of subject-matter experts, operational costs). As a result, poorly-specified VGHs can be used in the anonymization process. In these cases, the information loss (from a semantic point of view) at a given level i can be lower than the information loss at the level below it $i-1$. This means that the maximum loss of information may occur at any level of the VGH (not necessarily at the root node). In such scenario, a semantic-based VGH evaluation approach would help to correctly position the concepts in the VGH such that the fundamental property of specialization/generalization is not infringed.

Considering the above reasoning, we have integrated the use of data semantics in our VGH evaluation approach with the purpose of estimating more accurately the loss of information incurred in the VGH. As a result, well-defined VGHs can be used in data anonymization which would help to achieve a higher utility in the anonymized data.

3.4 GSL: Generalization Semantic Loss

In this section, we describe our approach to evaluate the quality of VGHs relying on a knowledge base and semantic similarity metrics applied to an anonymization context.

3.4.1 Knowledge Base

In our work, the knowledge base is a crucial element in the evaluation of VGHs. It acts as a gold standard in which the domain expert knowledge is encapsulated. The knowledge base is represented in the form of ontologies. This approach provides several advantages: Firstly, ontologies often incorporate the consensus opinion of a panel of experts, thus it mitigates the risk of having partial interpretations and single judgments over the domains represented in the VGHs. Secondly, the use of ontologies facilitates the interpretability of the terms in a VGH, hence, facilitating the communication among users working on the same domain. Thirdly, ontologies represent an efficient and scalable manner to perform an automatic evaluation of VGHs, in contrast with manual expert-based evaluations.

The semantic content of the ontologies is exploited by semantic similarity metrics to quantify the information loss incurred in a VGH. This is done by measuring the proximity between the actual values of the original dataset and their generalizations.

In this work we use WordNet [26] as knowledge base, and a set of path-based metrics as semantic similarity metrics [17]. The richness of WordNet provides a broad coverage of general concepts (due to its domain-independent knowledge). Furthermore, as it is manually assembled by experts, it provides a high-quality knowledge source [37]. However, the ontology used as the knowledge base is configurable as it can be tailored to the domain of the data being anonymized. There are many large and consensus ontologies available [22, 50]. For example, general-purpose (e.g., Yago [67]) or domain-specific sources (e.g., MeSH [3] and UMLS [43] for medical concepts; UNSPSC and NAICS [23] for e-commerce; geoNames [5] for geospatial information). Additionally, the exploitation of a single ontology does not represent a limitation, as much research has been done to support the integration of multiple ontologies [63].

3.4.2 Computing the GSL Score

The quality of a VGH is quantified by GSL, which is a score that indicates how good a given VGH would be to perform the anonymization of a dataset. Note that GSL is not a data utility metric, which measures the quality of the data after the anonymization occurs. In contrast, GSL aims to capture, a priori, the degree of semantic information loss that would occur in the anonymized data, due to generalization using a particular VGH.

From a data utility point of view, lower values of GSL are desirable, as it would suggest a higher retention level of the meaning of the original data in the specification of the VGH. The GSL can be quantified at four different degrees: TransGSL, LevelGSL, VghGSL and VghSetGSL. In the following paragraphs, we provide the definitions of those scores.

Definition 1. The GSL for a transition edge in a VGH (denoted by $TransGSL$) captures the semantic loss caused by a single generalization that transforms an original term (located at a leaf node) to any of its possible anonymized states (located at the ancestor nodes), as depicted in Fig. 3. TransGSL is given by (1) and is defined as the semantic dissimilarity between a term in a leaf node, l, and a term in an ancestor node, a.

$$TransGSL(l, a) = 1 - Sim(l, a) \tag{1}$$

In Eq. 1, the value of 1 represents the maximum similarity value for the WuP metric (Eq. 9 in Appendix A). This value also corresponds to the upper limit for $Sim(l, a)$ in Eq. 1 (i.e., $Sim(l, a) \leq 1$). If an alternative semantic similarity metric is used, this 1 must be replaced by the maximum value produced by the chosen metric (i.e., the similarity score between a concept and itself).

Definition 2. The GSL for the level i in a VGH (denoted by $LevelGSL$) is defined as the overall semantic loss occurring at level i, which represents an anonymized state for the modeled QID. LevelGSL is determined by applying an aggregation function, $\mathcal{F}$, to all the TransGSL scores T_j^i that go from the jth leaf node to its ancestor in level i. LevelGSL is given by (2):

$$LevelGSL(i) = \mathcal{F}(T_1^i, \ldots, T_n^i) \tag{2}$$

The function $\mathcal{F}$ combines the T_j^i scores into a representative value for each level of a VGH. Thus, diverse aggregation and fusion techniques can be used [71]. The choice depends on the intended analysis to be performed by users. For example, considering the maximum function (see 3a) would help to identify the transformations causing the highest semantic losses in the VGH; whereas the average function (see 3b) can serve to compare the overall quality of the VGH levels.

$$\mathcal{F}(s_1, ..., s_n) = \begin{cases} \max\limits_{i \in [i,n]} s_i, & \text{if } \mathcal{F} \text{ is maximum} \quad (3a) \\ \dfrac{1}{n} \sum\limits_{i=1}^{n} s_i, & \text{if } \mathcal{F} \text{ is average} \quad (3b) \end{cases}$$

Definition 3. The GSL for a VGH (denoted by $VghGSL$) is the main score in our evaluation scheme, as it represents the overall quality of a VGH. VghGSL is defined as the weighted sum of the LevelGSL scores. It is given by (4):

$$VghGSL(V) = \sum_{i=1}^{h} w_i \cdot LevelGSL(i) \tag{4}$$

where h is the height of the VGH, i is the index of a level in the VGH, and w_i is a predefined weight associated to the i level. The weights are chosen such that, for $1 \leq i \leq h$, $0 < w_i < 1$, and $\sum_{i=1}^{h} w_i = 1$. Different weighting schemes can be used, as the ones given in (5):

$$w_i = \begin{cases} \dfrac{1}{h} & \text{if weighting scheme is uniform} \quad (5a) \\ \dfrac{(h+1-i)}{\sum_{j=1}^{h} j} & \text{if weighting scheme is level-based} \quad (5b) \end{cases}$$

Weights are the manner in which the taxonomical structure of the VGH (e.g., height) is also considered into the evaluation score. For example, a generalization occurring in a flat VGH would produce a higher loss than one in a taller VGH. Furthermore, weights can be used to penalize the abstraction/specificity of the terms in the VGH. In the following paragraphs, we provide more details about the presented weighting schemes.

Uniform Weight. In this scheme, all the levels of the VGH obtain the same penalty (as given by Eq. 5a). It is defined with the aim of using the arithmetic mean in the computation of VghGSL. The reasoning is that, as it is not known how many generalizations will be applied to the data, all the levels have the same probability of satisfying the privacy goal.

Level-Based Weight. In this scheme, the weight is determined by the VGH level under consideration. The idea is to magnify the differences in semantics for the generalization at the lower levels of the VGH, where the more specific levels are found (as given by Eq. 5b). This follows a basic principle behind most semantic similarity metrics which is that the concepts at the lower levels are more similar than those at the upper levels (as those are more abstract). However, in our case, this is applied with respect to the VGH structure.

Definition 4. The VghGSL score definition (Eq. 4) can be extended to the multi-attribute QID scenario as follows: Let $QIDs = \{A_1, \ldots, A_n\}$ be the set of n attributes composing the QID set in a dataset, whose domains are represented by a set of n individual VGHs, $VghSet = \{V_1, \ldots, V_n\}$. An aggregated quality score can be obtained for the VghSet (denoted by $VghSetGSL$) as given by (Eq. 6):

$$VghSetGSL(V_1, ..., V_n) = \frac{1}{n} \sum_{i=1}^{n} p_i \cdot VghGSL(V_i) \tag{6}$$

where V_i is a VGH in the VghSet and p_i is a predefined user preference to tailor the anonymization to a specific use case. Such preferences are specified as weight values representing the importance of certain attributes for an analysis task. For example, the users may prefer to discourage generalizations on specific attributes that can decrease the utility of the anonymized data.

Example 2. To illustrate the different degrees of GSL and the benefits of leveraging data semantics to carry out VGH evaluation, let us consider a univariate dataset where the QID attribute refers to *occupation* values. The user has created a VGH (based on her own knowledge) for this QID as shown in Fig. 3. The leaf nodes correspond to the original values in the dataset. The ancestor nodes correspond to the candidate values used for the generalizations. To demonstrate our approach, we use the WuP metric (Eq. 9 in Appendix A) to compute the semantic dissimilarity between two terms, and WordNet as the knowledge base. The senses (in WordNet) for each concept are shown next to each word in the VGH.

Applying Eq. 1 to all the leaf-ancestor transitions, we obtain the TransGSL scores shown in Table 3. Ideally, more general terms are located at higher levels in the VGH and more specialized terms are lower in the VGH in order to be semantically consistent. However, there are imprecisions in the VGH which might not be easy to identify at first sight without the help of the TransGSL scores. For example, the semantic loss caused by the transformation *"ophthalmologist"→"medical practitioner"* in L1 (i.e., 0.1111) is higher than the one caused by *"ophthalmologist"→"doctor"* in L2 (i.e., 0.0714). This indicates that the ancestors of *"ophthalmologist"* have not been correctly positioned in the VGH (based on their level of abstractness and coverage of their domain), as *"doctor"* is more specific than *"medical practitioner"*, which is a broader concept that can also include *"nurses"* or *"pharmacists"*, who are not *"doctors"*. Thus, *"doctor"* should be a subclass of *"medical practitioner"*, not vice versa. Another case of semantic inconsistency can be observed in the transformations specified for the terms *"optician"* and *"optometrist"*, as a moderate loss is incurred (i.e., 0.3043 and 0.3333). This is because although these terms refer to people involved in eye caring, they are not *"medical practitioners"* or *"doctors"*. Finally, by inspecting the TransGSL scores the users can also detect inconsistencies in the *is-a* taxonomic relationships defined in the VGH. For example, the *"math/music teacher"→"education"* transformations, where a more appropriate ancestor term would be *"educator"*.

Table 3 also shows the LevelGSL scores (Eq. 2), which have been computed using average as the aggregation function (Eq. 3b). The LevelGSL values indicate that the semantic loss is

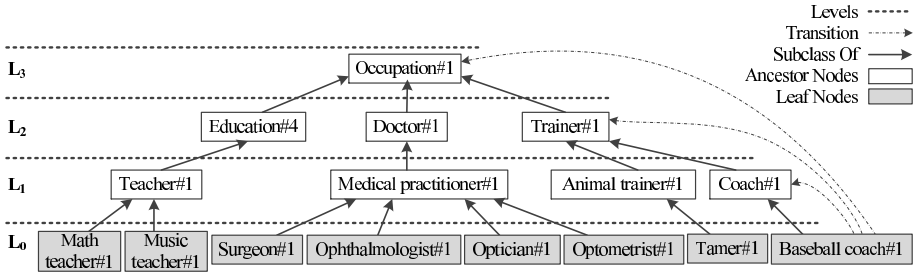


Figure 3: A VGH specified for *occupation*

Table 3: GSL scores for the *occupation* VGH

Leaf Nodes	**Level 1**				**Level 2**			**Level 3**
	Teacher	Medical practitioner	Animal trainer	Coach	Education	Doctor	Trainer	Occupation
Math teacher	0.0400	-	-	-	0.8261	-	-	0.8095
Music teacher	0.0400	-	-	-	0.8261	-	-	0.8095
Surgeon	-	0.0769	-	-	-	0.0370	-	0.8182
Ophthalmologist	-	0.1111	-	-	-	0.0714	-	0.8261
Optician	-	0.3043	-	-	-	0.3333	-	0.7895
Optometrist	-	0.3043	-	-	-	0.3333	-	0.7895
Tamer	-	-	0.0435	-	-	-	0.0909	0.8000
Baseball coach	-	-	-	0.0435	-	-	0.0909	0.8000
LevelGSL				0.1205			0.3261	0.8053
VghGSL								0.4173

higher as one moves up higher in the VGH, which is in general correct, as the most specific terms should be located lower in the VGH. This table also shows the VghGSL score (Eq. 4), which has been computed using a uniform weighting scheme (Eq. 5a). VghGSL represents the overall quality for the VGH.

In the above example, we illustrated some of the benefits that our approach offers to the users in terms of facilitating the evaluation of VGHs, such as: (1) Inconsistent specifications introduced in the VGH can be easily identified (e.g., misclassifications, redundancies, incorrect term positioning); (2) A clearer differentiation can be made between terms that look similar; (3) Users do not have to depend on the limited availability of knowledge engineers and the associated cost as the required expert knowledge is reduced; (4) Users mitigate the risk of relying on the subjective judgment of a single individual expert as the knowledge base is represented by consensual ontologies often created by a panel of experts.

4 Experimental Setup

In this section, we describe the experimental methodology, the developed prototype, the test environment, and the parameters that defined the evaluated experimental configurations: the selected datasets, VGHs, anonymization algorithms, and privacy settings.

4.1 Experimental Methodology

To evaluate our proposed approach, we conducted a series of experiments that pursued the following objectives: (1) to investigate how the utility of anonymized datasets is impacted by the quality of the VGHs used; (2) to demonstrate how GSL offers a reliable and objective mechanism to identify well-specified VGHs; (3) to show how GSL helps to improve the

quality of anonymized datasets and to reduce the overall effort spent on the anonymization process; and (4) to assess the costs of using our approach.

For this purpose, a set of VGHs (that served as the candidate VGHs to be compared) was created modeling the domains of various datasets' attributes (see Section 4.3). We considered two scenarios: single-attribute and multi-attribute, which consisted in using two sizes of QIDs (denoted by $|QIDs|$). We firstly used $|QIDs|$=1 to study the influence that each individual VGH had in the anonymization process. Secondly, we used $|QIDs|$=4 to evaluate the applicability of the GSL metric when multiple attributes are combined. The quality of the candidate VGHs was measured using the GSL metric, which allowed to compare them in a quantitative way. Later, we tested the effectiveness of the VGHs by using them in the anonymization of various datasets using two well-known anonymization algorithms and different levels of privacy (see Section 4.4). Next, we measured the quality of the resulting datasets using both task-independent and task-dependent data utility metrics (see Section 5). Finally, we performed different types of analysis to assess the benefits, costs, and reliability of our proposed solution (see Section 6).

4.2 Prototype and Environment

The experiments were performed in a computer with an Intel Core i5-4310U CPU at 2.00Ghz, 8GB of RAM, Windows 7 64-bit, and the Oracle HotSpot Java Virtual Machine version 7 with a 1GB heap. Our prototype was implemented in Java; internally it uses the WordNet Similarity for Java (WS4J) library [6] for all the semantic similarity computations.

4.3 Evaluation Data

Our evaluation data consisted of a set of datasets and VGHs. The following paragraphs briefly describe them:

- **Evaluated Datasets.** We used four publicly available datasets. The first two datasets were *Adult* and *German Credit* from the UCI Repository [42]. Both datasets have been widely used by the data privacy community in the past to evaluate anonymization algorithms, in particular, k-anonymity techniques [12, 32, 34, 39, 40]. The Adult dataset consists of 30,162 records (after removing those with missing values) with census information. The German Credit consists of 1,000 records with credit applicants information. The third dataset was the *Chicago Homicide* dataset [1]. This dataset is particularly interesting due to the diversity of categorical values that it contains (in comparison to Adult and German Credit). It consists of 23,817 records with information about homicides filed by the Chicago Police Department for the years 1965-1995. The fourth dataset was the *Insurance* dataset [2], which contains 10,000 records with personal information that can be of interest to an insurance company for carrying out a risk assessment on potential clients.

  For each dataset, we considered a set of categorical attributes as QIDs whose domain must be modeled in a VGH in order to be anonymized. We considered those attributes having the most distinct values, as it allowed us to have a diverse set of candidate VGHs to evaluate our proposed metric. For the single-attribute scenario, all the attributes shown in Table 4 were used. For the multi-attribute scenario, four attributes (highlighted in bold in Table 4) from the Chicago Homicide and the Insurance datasets were used.

Table 4: QIDs considered for VGH evaluation and anonymization

Dataset	*Attribute*	*Cardinality*	*Attribute Index*
Adult	Occupation	14	7
GermanCredit	Purpose	12	3
ChicagoHomicide	Location	96	46
	PHome	11	48
	POutdoor	33	56
	CausalFactor	47	59
	VicRelation	95	71
	OffRelation	95	72
	WClub	57	106
	WKnife	25	109
Insurance	**Occupation**	60	3
	Workplace	29	4
	Hobby	40	5
	PlaceOfHobby	32	6

- **Evaluated VGHs.** For the single-attribute scenario, our testbed consisted of 100 candidate VGHs for each QID attribute. Those VGHs were created by perturbing the semantic content of an *ideal* VGH which was constructed by extracting the minimal taxonomy from WordNet for each in-scope attribute of our evaluated datasets. In order to obtain VGHs that yielded a varied range of quality scores for our tests, we applied diverse transformations to the ideal VGH, such as: removing levels in the VGH, aggregating nodes in different groups, and replacing the terms of the ancestors with another one that is within a semantic similarity boundary. For the multi-attribute scenario, three sets of VGHs were created based on the involved attributes. Each VGH set consisted of those VGHs having the best (ranked #1), the worst (ranked #100), and a medium (ranked #50) quality according to their individual GSL scores.

Finally, it is worth mentioning that most of the attributes' values in the evaluation data corresponded to concepts that were directly found in WordNet. Only a few values were not found directly (due to their ad-hoc linguistic labels). In such cases, the values were mapped to a closely-related WordNet concept (this strategy has been used in previous works [13, 47]). In our work, it is assumed that the correct sense for a noun is provided along with the VGH. This is because the area of automatic word-sense disambiguation is a broad research field on its own [53] and is beyond the scope of our paper. As the users are usually involved iteratively in the whole anonymization process, a manual disambiguation is a reasonable precondition to perform the evaluation of VGHs for the time being.

4.4 Anonymization Algorithms and Privacy Settings

To evaluate the effectiveness of the VGHs and the accuracy of GSL, we used Datafly [68] and Incognito [38], which are two anonymization algorithms widely-cited and highly-known in the data privacy field [9, 27]. Moreover, as they use distinct strategies of anonymization, we were able to further broaden the tested scenarios. Datafly is a greedy heuristic algorithm that generalizes the data until every combination of QID values appears at least k-times. Although this algorithm guarantees a k-anonymous solution, it does not provide the minimal generalization [61]. Incognito is an algorithm that constructs a generalization lattice representing the solution search space. It traverses the solution space performing a bottom-up breadth-first search and pruning parts of it by using predictive tagging. Incognito produces globally optimal results. In our work, we used the implementations publicly available in the UTD Anonymization Toolbox [4] as the core versions of the algorithms. Since the aim of our evaluation is to assess the semantic usefulness of the anonymized

datasets, we must employ semantic heuristics (to share the same principle of data utility [27]). Hence, the original distribution-based approach of both algorithms was replaced with semantic-aware versions (following a strategy similar to those of previous works in the literature [44, 47]).

Finally, in terms of privacy settings, we adopted k-anonymity [61] as the chosen privacy model. This is because, as discussed in Section 2.1, it is a fundamental and representative anonymity principle in the privacy area [12]. We varied the tested k-values as $k \in [2..100]$. The broad range of the values used in these experiments was selected on the basis of those commonly used in the literature [15, 33, 38, 39, 76].

5 Evaluation Criteria

In this section, we describe the metrics used in our experimental evaluation in terms of the VGH quality, data utility, efficiency of the anonymization process, and the different costs (i.e., resource utilizations) of using our solution.

5.1 VGH Quality

The quality of the VGHs is expressed in terms of our GSL metric (presented in Section 3.4). For each VGH, we calculated the VghGSL score using the average as the aggregation function (Eq. 3b). Since GSL leverages semantic similarity measures, we used two widely-used path-based metrics that are part of the WS4J library [6]: Wu and Palmer (WUP) [75] and Leacock and Chodorow (LCH) [36]. This decision was made in order to assess the generality of GSL with respect to the used semantic similarity metric. These metrics capture the similarity of two concepts by measuring how closely they are related in a taxonomy (i.e., their structural relations). WUP takes into account the depth of the concepts in a taxonomy and the depth of their least common subsumer. LCH considers a similar approach but also takes into account how deep each of the concepts is in the taxonomy. Readers can refer to Appendix A for the WUP and LCH metrics calculation.

5.2 Data Utility

To evaluate the effectiveness of the VGHs, we measured the utility of the data after anonymization. This involved using a mixture of task-independent and task-dependent metrics. This hybrid strategy was chosen in order to perform a more robust assessment of GSL with respect to its capability to estimate the quality of the anonymized data. The following paragraphs briefly describe the selected metrics:

- **Task-Independent Metrics.** In our work, we firstly focused on applying two well-known task-independent metrics as not knowing in advance the analysis task that the data recipients will perform over the anonymized data is an essential premise of PPDP. Thus, the metrics considered were: Semantic Information Loss (SemILoss) [49] and Semantic Sum of Squared Errors (SSE) [25]. SemILoss measures how semantically different the anonymized values are, on average, compared to the original ones. Semantic SSE is an adapted version of the traditional SSE metric, which has been widely-used in the area of statistical disclosure control with microaggregation methods [24]. Semantic SSE integrates semantic similarity metrics to calculate the distance between the original records and their anonymized version (which acts as the centroid of an anonymized group) [7, 47, 48].

- **Task-Specific Metrics.** We also evaluated the utility of anonymized solutions in a task-specific context (from a data mining point of view) by performing clustering over the anonymized datasets. For this purpose, we compared the similarity between the clusters obtained from the original data against the ones obtained from the data that was anonymized using each of the candidate VGHs. In this manner, we could determine which VGHs produced the best solutions. In the clustering process, we used the Ward's method [74] to perform a hierarchical clustering and the Calinski-Harabasz variance ratio criterion [18] to determine the appropriate number of clusters in each dendrogram to be compared (i.e., final partition). To compare the quality of the clusters we used two representative clustering evaluation metrics: the Rand index (Rand) [58] and the Normalized Mutual Information (NMI) metric [66]. Rand is a pairwise agreement metric that assesses whether each pair of data points are either clustered together or separated into different clusters. NMI is an entropy-based metric that relies upon concepts from information theory to measure how much information is shared between partitions of clusters. Both metrics range between 0 and 1; larger values indicate a higher similarity between the partitions. These metrics were computed using the framework presented in [20].

Readers can refer to Appendix B for the equations of the previously discussed data utility metrics, as well as the detailed methodology used for performing the cluster comparison.

5.3 Anonymization Efficiency

Selecting a priori the best VGHs to perform the anonymization of data cannot only improve the quality of the resulting data but also the efficiency of the anonymization process. In particular, this is desirable in scenarios of continuous data publishing. These improvements can be achieved by reducing the overall effort spent by the users. For this reason, we also assessed the time-saving benefits that APES can bring to the TAS scenarios (described in Section 3.2). For this purpose, we compared the time taken to anonymize data using TAS, against the one using APES. For this part, all the associated stages of the corresponding anonymization process (from pre-processing the dataset until saving the anonymized dataset), for each of the candidate VGHs, were considered. This strategy is similar to the one used in [12]. The following paragraphs briefly describe the used metrics:

- **TAS Efficiency.** The execution time taken for an experimental configuration (e.g., one evaluated k-value) using TAS (t_{TAS}) involves the anonymization time using the n candidate VGHs. For each VGH, its anonymization time includes the original dataset reading/uploading (r_i), the anonymization (a_i), the anonymized dataset writing (w_i), and the time taken to evaluate the utility of the anonymized solution according to the SSE metric (u_i). This is given by (7):

$$t_{TAS} = \sum_{i=1}^{n} (r_i + a_i + w_i + u_i) \tag{7}$$

- **APES Efficiency.** The execution time taken for an experimental configuration using APES (t_{APES}) includes the time spent on the quality evaluation of the n candidate VGHs using the GSL score (e_i), plus the elapsed time of the anonymization process using the best VGH (B) according to its GSL score. That is, the original dataset reading/uploading (r_B), the anonymization (a_B), the anonymized dataset writing (w_B),

and the time taken to evaluate the utility of the anonymized solution according to the SSE metric (u_B) from the best VGH. This is given by (8):

$$t_{APES} = \sum_{i=1}^{n}(e_i) + r_B + a_B + w_B + u_B \tag{8}$$

It is worth noting that when using TAS, the users generally face the overhead of uploading a dataset m times (where m is the number of configurations to be tested). This constraint is usually dependent on the framework used by users to perform the anonymization [4, 21, 56]. In our case, the framework used in our experiments exhibited that overhead behavior [4]. This behavior exemplifies how anonymization can be very time-consuming, as it is a process influenced by multiple factors. To isolate the overhead of uploading the data many times, we equalized the reading times for both approaches (i.e., TAS and APES). That is, the reading time was considered only once in the comparison, thus favoring to some degree the competitor approach TAS versus ours.

5.4 Resource Utilization

Finally, we also studied the costs of using our proposed solution. That is, we also assessed the computational resources required to perform the quality evaluation of VGHs (i.e., GSL computation). The main monitored metrics were: memory consumption (MB), CPU usage (%), and execution time (ms). The garbage collection (GC) was also monitored because it is an important performance concern in Java [55]. For all metrics, lower values are better.

6 Experimental Results

In this section, we present the results obtained from our evaluation in terms of the metrics relevant to each experiment. We conclude with a final discussion for data publishers.

6.1 Data Utility Results

Our analysis initially focused on assessing the capacity of the GSL metric to capture (a priori) the effectiveness of the VGHs for anonymizing data. This was done from a task-independent and a task-specific perspective. The aim was to evaluate the improvements that GSL can bring to the utility of the anonymized data for single- and multi-attribute scenarios.

Task-Independent Perspective. For this analysis, we firstly investigated the degree of association between the scores representing the quality of the VGHs and the utility of the anonymized datasets. In this manner, we could investigate the accuracy of GSL. For this purpose, we used the Spearman's rank order correlation (r_{Spm}) which measures the strength of a monotonic relationship (but not necessarily linearly related) between paired data [64]. r_{Spm} can take values from -1 to +1; the closer the value is to ± 1, the stronger the relationship. For the sake of clarity, the QIDs in the figures are labeled using their corresponding dataset name and attribute index (e.g., H46 corresponds to the attribute 46 of the Homicide dataset).

Fig. 4 shows the correlations obtained between the anonymized data utility (quantified with SSE and SemILoss across all the tested k-values) and the VGH quality scores (GSL

Table 5: Rating Scale for the Spearman's rank order correlation

Ranges	Category
$0.90 \leq r_{Spm} \leq 1.00$	Very Strong
$0.70 \leq r_{Spm} < 0.90$	Strong
$0.50 \leq r_{Spm} < 0.70$	Moderate
$0.30 \leq r_{Spm} < 0.50$	Weak
$0.00 \leq r_{Spm} < 0.30$	Very Weak

computed with WUP and LCH) per dataset attribute. The strengths of the correlations obtained across all datasets were in the category of *strong* and *very strong* (according to the commonly accepted qualitative interpretation of r_{Spm} [52], shown in Table 5). The correlations for SSE and GSL_{WUP} ranged between 0.72 and 0.99 (with an average across all tested datasets/QIDs of 0.93 and a standard deviation of 0.07), whereas the ones for SemILoss and GSL_{LCH} ranged between 0.77 and 0.96 (with an average of 0.91 and a standard deviation of 0.06). These results proved that GSL worked well, as it was able to capture with a good degree of precision the effectiveness of VGHs as the VGHs qualified as "well-specified" by our metric were the ones that gave the best results. Moreover, the results obtained with GSL_{WUP} and GSL_{LCH} were similar (as they achieved comparable trendings), demonstrating the generality of GSL with respect of the chosen semantic similarity metric. It can also be noticed how the accuracy of GSL was relatively similar across the four tested datasets (with most r_{Spm} correlations above 0.90), meaning that GSL worked well irrespectively of the tested dataset. The lowest correlation was experienced with H48. This behavior is explained by the fact that this QID belongs to a relatively homogeneous domain (i.e., home) and it also has a very low cardinality (i.e., 11). These two characteristics combined provoked that the VGHs for this QID were significantly flat (e.g., their average depth was 4 levels). As a consequence, the anonymized solutions for this QID reached the root node considerably more frequently than the other QIDs (hence, experiencing its maximum data distortion and making the VGHs indistinguishable from each other in terms of their quality). It is worth noting that, even in such conditions, the correlation of GSL remained within the strong category.

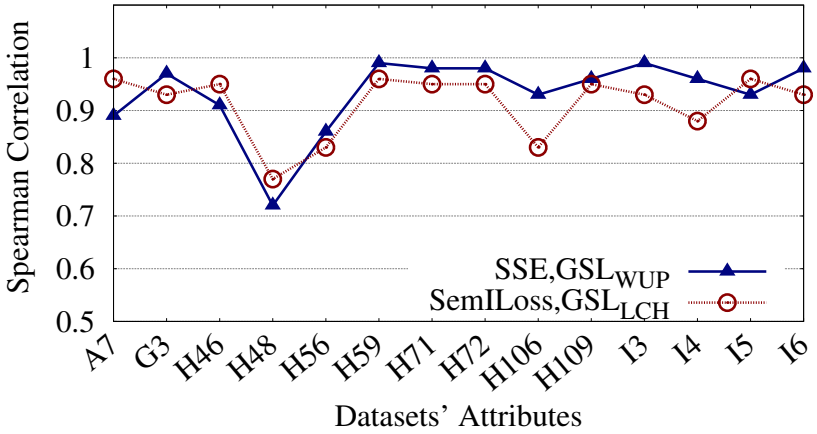
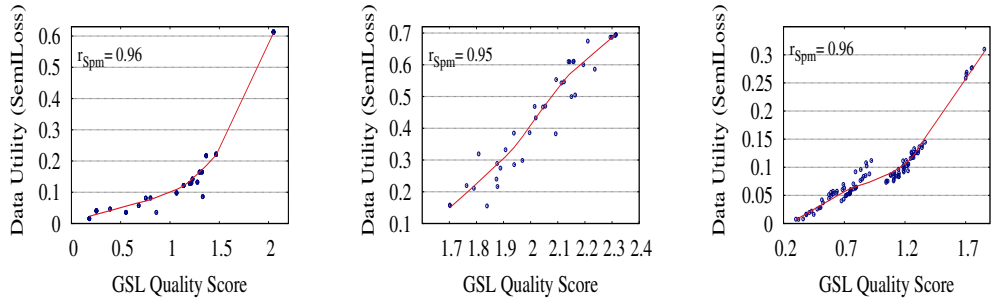
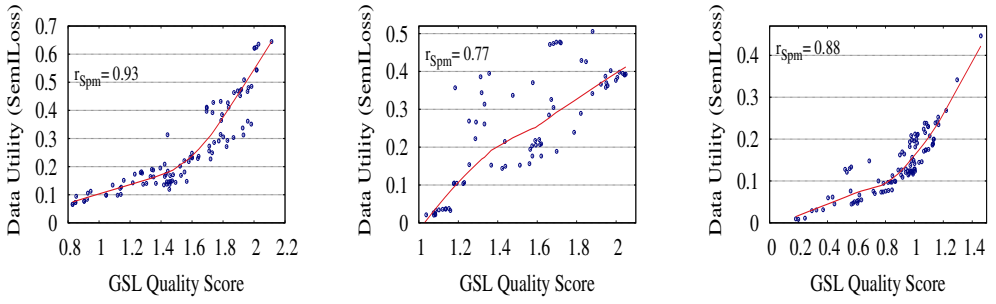


Figure 4: Corr. Comparison between VGH Quality vs Data Utility Scores for $|QIDs|$=1



(a) A7 (b) H59 (best) (c) I5 (best)



(d) G3 (e) H48 (worst) (f) I4 (worst)

Figure 5: Individual Correlations between VGH Quality vs Data Utility Scores

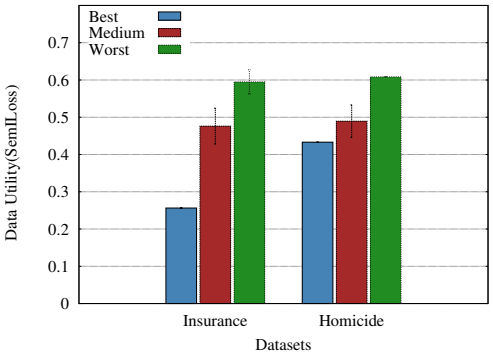


Figure 6: SemILoss for $|QIDs|$=4

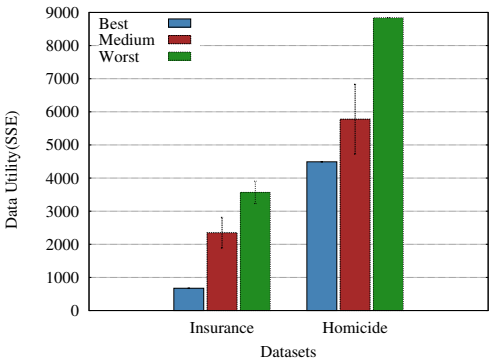


Figure 7: SSE for $|QIDs|$=4

To complement the discussion of this analysis, the highest and the lowest correlations obtained for each dataset are shown in Fig. 5. For the sake of brevity, we only present the results for SemILoss and GSL_{LCH}. However, similar trends were obtained for the other configurations. The upward trends observed in the figures indicate that the effectiveness of the VGHs in the anonymization positively correlates with their quality. In this case, an a priori selection strategy brought improvements in the utility of all the anonymized datasets between 16% and 98% for SemILoss; and between 22% and 99% for SSE. In this context, these improvements corresponded to the differences in data utility that can be achieved by using the best VGHs, compared to the worst ones, for anonymizing the data.

As next step, we analyzed the results for the multi-attribute scenario. They are depicted in Figs. 6 and 7, which show the performance that each VGH set obtained, on average, across all the tested k-values. Similar to the single-attribute scenario, upward trends are observed

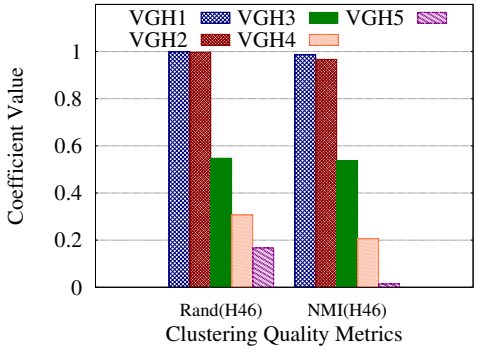


Figure 8: Clustering Comparison ($|QIDs|$=1)

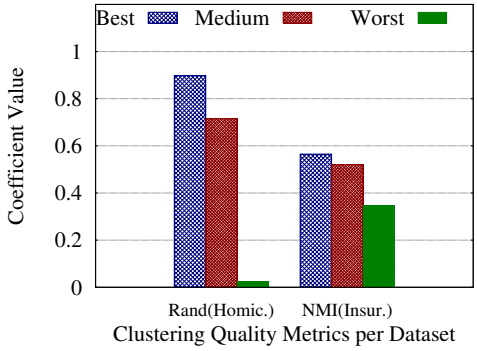


Figure 9: Clustering Comparison ($|QIDs|$=4)

in the figures, indicating that the effectiveness of the VGH sets in the anonymization positively correlated with their quality. When comparing both datasets, it can be noticed how a lower information loss was obtained for the Insurance dataset. This is because the quality of its VGHs was better than those defined for Homicide. For example, the GSL quality scores for the best, medium, and worst VGH sets of Insurance were 0.0491, 0.2027, and 0.4397, respectively. Whereas for Homicide they were 0.1810, 0.3583, and 0.6856, respectively. For the multi-attribute scenario, the a priori VGH selection strategy brought improvements in the utility of the anonymized datasets between 29% to 57% for SemILoss; and from 49% to 81% for SSE. Similarly to the single-attribute scenario, these improvements corresponded to the differences of using the best VGH set, versus the worst one for anonymizing data.

Task-Specific Perspective. In the following paragraphs, we present the results of the comparison done between the clusters obtained from the original data and those obtained from the data anonymized with the candidate VGHs. Figs. 8 and 9 depict the results for the single- and multi-attribute scenarios, respectively. For the sake of brevity, in the single-attribute scenario we only present the results for five representative VGHs belonging to the *location* domain of the Homicide dataset (i.e., H46) whose quality score (GSL_{WUP}) ranged between 0.26 and 0.6957. This domain was chosen to illustrate the obtained results due to its cardinality (i.e., 96), which was the highest among the tested domains. For this experiment, we anonymized the Homicide dataset using a value of k=10 when $|QIDs|$=1 and k=2 when $|QIDs|$=4, as they offered a moderate degree of privacy and a good level of data utility preservation (avoiding over-generalization of the data).

The downward trend observed in Figs. 8 and 9 indicate that highly-ranked VGHs (based on their GSL score) preserved the utility of the original data better than the lower-ranked VGHs, therefore, allowing a better interpretation of the data. Some poorly-defined VGHs even obtained a clustering level agreement close to zero. This is because the anonymization using such VGHs reached the root node (i.e., maximum generalization), which means that all the records were clustered together making the records indistinguishable from each other. Finally, it is worth mentioning that a similar behavior (i.e., a downward trending) was exhibited by the rest of the attributes in the evaluated datasets. The main difference among them was that the slopes of the trends varied.

Summary. In conclusion, this analysis demonstrated that GSL offers a reliable mechanism to identify well-specified VGHs as it estimated well the effectiveness of the VGHs. As a consequence, the utility of the anonymized data was improved (as showed by task-independent and task-dependent criteria). We also validated that our evaluation approach correlated well regardless of the semantic similarity metric used to calculate the GSL score.

6.2 Anonymization Efficiency Results

Our next analysis focused on assessing the benefits of using APES to help to improve the efficiency of the anonymization process by reducing the effort spent on the design of VGHs. In the following paragraphs, we present an effort comparison between anonymizing the data using TAS versus APES.

To offer a more comprehensive analysis, we discuss our results from two perspectives (presented in Section 3.2): (1) when a user needs to test multiple candidate VGHs to choose the best one for a predefined configuration, and (2) when a user needs to test different levels of privacy (e.g., k-values) with each of the candidate VGHs. For the sake of brevity, we present the results corresponding to the Homicide dataset. This is because they are representative of our identified findings and observations, which are equally applicable to the other datasets.

(1) Evaluating multiple candidate VGHs. Figs. 10 and 11 depict the total time taken by the anonymization process (using TAS and APES) as the number of candidate VGHs increases. Fig. 10 shows a single-attribute scenario, in which 100 candidate VGHs were available for modeling the domain of one attribute (i.e., H46). An upward trend over time can be observed for TAS. This is because, as more candidate VGHs needed to be evaluated a posteriori (to find the best one), the time of the overall anonymization process increased substantially (due to the need to perform multiple trial-and-error anonymization cycles). On the contrary, the trend for APES remained practically steady (with marginal increases) having an average anonymization time of 19.54 mins (with a standard deviation of 0.17 mins). This is because with APES the best VGH could be known before anonymization. Hence, it was only one VGH which was used to perform the anonymization of the data. Moreover, the time differences within APES among the number of candidate VGHs were small, as they only corresponded to the time taken to compute the GSL score for the candidate VGHs (as reflected in the low standard deviation). In this scenario, by using an a priori strategy (depicted by APES), we obtained performance improvements that ranged between 54% and 93% which corresponded to time-savings of 22.62 mins and 264.65 mins (times corresponding to the cases of evaluating 10 and 100 candidate VGHs), respectively. A similar behavior can be observed in Fig. 11, which shows a multi-attribute scenario in which 3 candidate VGH sets were available for modeling four domains of the Homicide dataset. In this scenario, the average anonymization time using APES was 133.48 mins (with a stan-

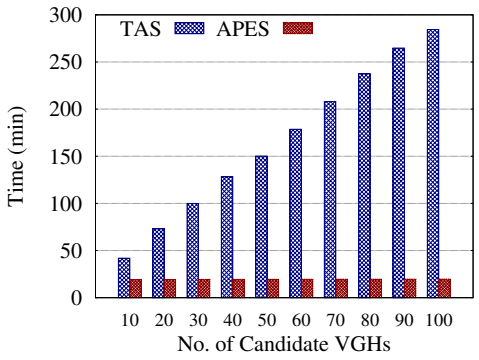


Figure 10: Time vs No. of VGHs ($|QIDs|$=1, H46, k=10)

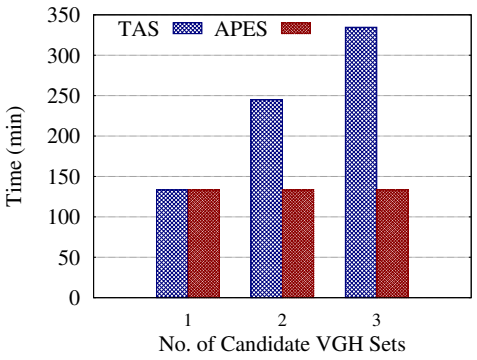


Figure 11: Time vs No. of VGH Sets ($|QIDs|$=4, Homic., k=2)

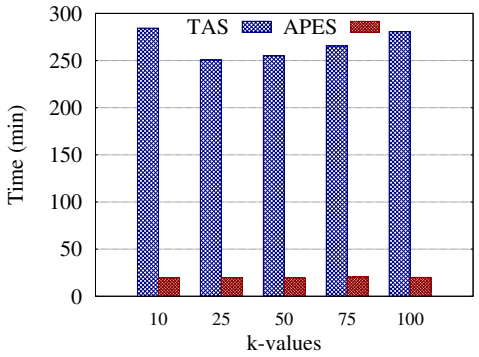


Figure 12: Time vs k ($|QIDs|$=1, H46)

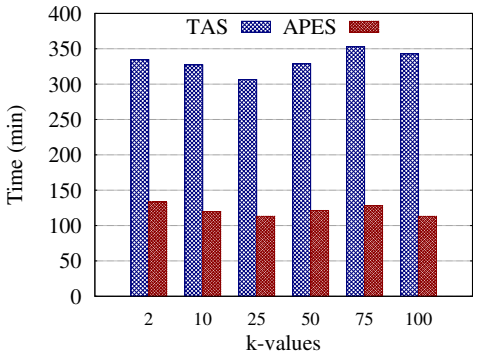


Figure 13: Time vs k ($|QIDs|$=4, Homic.)

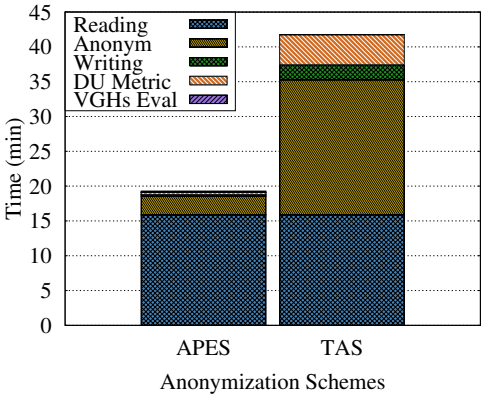


Figure 14: Drilldown for $|QIDs|$=1, k=10, H46

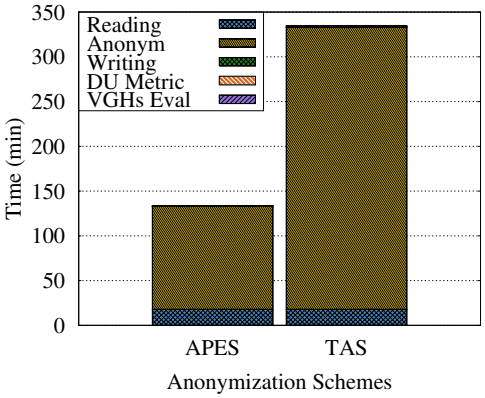


Figure 15: Drilldown for $|QIDs|$=4, k=2, Homicide

dard deviation of 0.02 mins) and the achieved performance improvements ranged between 45% and 60% which corresponded to time-savings of 111.39 mins and 201.15 mins, respectively. Similar performance improvements were obtained for the Insurance dataset, where the time-savings ranged between 62.09 mins and 113.08 mins.

(2) Evaluating different levels of privacy. Based on the previously observed behaviors, our hypothesis was that improvements should also be obtained when testing different levels of privacy. This is because, to fulfill this usage scenario, the TAS process also required the anonymization of the datasets several times (one per candidate VGH/k-value combination). Therefore, offering potential time-savings to exploit. This was confirmed by the results of this experiment. Even though there were some minor variances in the percentage of improvements that APES achieved, the improvements were closely similar, across the different k-values, per tested dataset/QID combination. This is visually shown in Figs. 12 and 13, which depict the total time taken by the anonymization process as the value of k increases for $|QIDs|$=1 and $|QIDs|$=4, respectively. For example, in Fig. 12 it can be observed that the overall improvements of using APES were more than 91% compared to using TAS; this represented time-savings of more than 3.80 hrs. Larger benefits can be observed for the multi-attribute scenario shown in Fig. 13, where the improvements achieved up to 67% (approximately 3.83 hrs). Similar time-savings were obtained for the Insurance

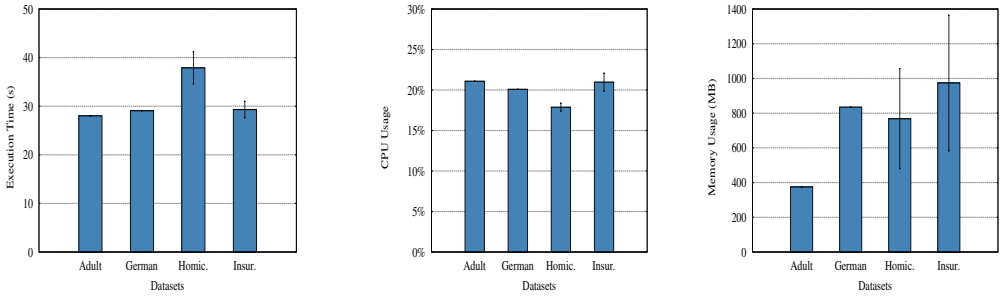
dataset, where the improvements achieved up to 66% (approximately 2.71 hrs). The benefits obtained for the multi-attribute scenario were significant, in particular considering that only 3 VGH sets were involved. This is because the number of QIDs was larger, which considerably increased the size of the anonymization solution space. As a consequence, the anonymization time using TAS significantly increased too. To complement this analysis, by having a better understanding of the phases which were benefited by the usage of GSL, Figs. 14 and 15 show the time that both schemes (TAS and APES) spent in each phase of the anonymization process. These figures present a breakdown of the experimental configurations when k=10 and $|QIDs|$=1, and when k=2 and $|QIDs|$=4. It can be observed that APES offered time-savings benefits in all the phases of the anonymization process. In particular, to the anonymization phase, which is usually the most time-consuming phase when the number of QIDs is large. Finally, it is worth remarking that our time analysis assumed that the reading time of both APES and TAS approaches were equal, even though it was not the case (as explained in Section 5, the dataset reading actually occurred every time the anonymization process happened). This decision was taken in order to offer a more conservative perspective of the time-savings gained by the usage of GSL. However, bigger time-savings can be expected (compared to the reported ones) if the real reading time of using TAS is considered (as it experienced a linear growth with respect to the number of candidate VGHs used for anonymizing the datasets).

Summary. In conclusion, the results of this experiment showed how the potential time-savings that GSL can bring to the anonymization process are significant. This is because GSL enables users to perform an a priori evaluation of the candidate VGHs in order to identify the best one and use it for anonymization (i.e., APES). As a consequence, the need of costly trial-and-error anonymization cycles (i.e., TAS) is eliminated.

6.3 GSL Costs

Finally, we also assessed the costs associated with using GSL to evaluate the quality of VGHs. These results are shown in Figs. 16a, 16b, and 16c, which depict the execution time, CPU, and memory utilizations of the evaluation process of the 100 VGHs in our testbed, per dataset.

The computation of GSL proved to be lightweight in terms of CPU and execution time. For example, the evaluation of Homicide, which is the dataset whose attributes are the most diverse (i.e., highest cardinalities), took an average of 38 sec (with a standard deviation of 3.3 sec). Moreover, the unitary time cost of evaluating a single VGH in all datasets was 310 ms (with a standard deviation of 45 ms). Similarly, the CPU usage never exceeded 22%



(a) Execution Time (b) CPU Usage (c) Memory Consumption

Figure 16: Efficiency of Evaluation w.r.t. (a) Execution time, (b) CPU and (c) Memory

(meaning that there was a considerable amount of idle resources to support larger workloads). In terms of memory, our evaluation only used approximately 45% of the available memory. Finally, only a minimal amount of time (i.e., less than 3% in all cases) was spent on GC. This was another positive indicator that the memory settings were appropriate for the computation of GSL. The factors influencing the computational costs of GSL are the semantic similarity metrics used and the taxonomical properties of the VGHs (e.g., depth, breadth, number of leaves).

Summary. In conclusion, the results of this experiment showed that the usage of GSL is lightweight in terms of its consumption of computational resources, making it practical for real-world usage.

6.4 Final Discussion

The presented experimental results have proven how the usage of GSL can substantially enhance the performance of the anonymization process by accurately estimating the quality of the candidate VGHs. In the following paragraphs, we provide guidelines for data publishers to understand the conditions under which GSL can be useful and discuss the wider application of the approach.

a) In this paper, we proposed a quantitative mechanism to assess the quality of VGHs. The GSL score can be used by data publishers to identify (or design) well-specified VGHs so that their effectiveness is estimated before conducting the anonymization process. As a result, not only the effort and expertise required to properly evaluate VGHs can be reduced, but also higher quality VGHs can be used in anonymization which improves significantly the utility of the resulting anonymized data. It should also be noted that while the utility of the anonymized data is enhanced (by better preserving the semantics of the original data), the privacy remains the same, as the privacy goal set by the data publishers is not impacted. In our experimental evaluation, the k-anonymity levels were satisfied while the semantic utility of the data was improved. This means that the anonymized data kept the same protection degree and vulnerabilities of the chosen privacy model. For instance, when using k-anonymity, even though the data satisfies the given k-value, the re-identification risk may not be necessarily $1/k$. This is because k-anonymity assumes that each record in the table represents a distinct individual; otherwise, the achieved k may actually represent fewer than k record owners (hence involving a higher risk of re-identifying the data). This well-known k-anonymity vulnerability has been addressed in the literature by other anonymization models. For example, the (X, Y)-anonymity [73] stipulates that each value on X (i.e., QID) is associated with at least k distinct values on Y (i.e., a key in the table that uniquely identifies record owners). Therefore, it is important for data publishers to select the privacy model that better fits their needs. Moreover, even though our experimental evaluation used k-anonymity, our approach is not tied to a particular privacy model or its associated goal. Instead, the weights (discussed in Section 3.4) can be used to tailor GSL to the desired model.

b) As shown in our experiments, GSL captures well the quality of the VGHs. Nevertheless, we identified two scenarios in which the accuracy of the GSL score decreased: (1) when the data was over-generalized and (2) when non-optimal anonymization algorithms were used. The first scenario occurs when an anonymization solution reaches the root node of the VGH (i.e., maximum generalization). Such situation can

emerge when a QID, which has a low cardinality and/or which belongs to a homogeneous domain, is modeled in a VGH. When those factors are combined, the created candidate VGHs can be relatively flat and share similar structural properties. Thus, it is likely that those VGHs reach the maximum generalization more rapidly. If that occurs, the VGHs produce identical anonymization solutions (regardless of their GSL quality score). Thereby, it would be more challenging to differentiate the effectiveness of the VGHs from their anonymization results; as, under those circumstances, the quality of the anonymized solutions among the candidate VGHs becomes indistinguishable. These conditions can provoke a decrease in the correlation degree between the VGH quality and the data utility, which would lead to a reduction in the accuracy of the GSL metric. In the second scenario, we observed that GSL works better when the anonymization is performed using algorithms that produce globally optimal solutions. This is because for poorly-specified VGHs, the maximum loss of information may occur at any level of the VGH (as discussed in Motivation 3 of Section 3.3). Thus, iterative greedy strategies (e.g., Datafly) may favor anonymizations in those QID attributes that have well-specified VGHs (as they cause a lower information loss). This would lead to conduct the anonymization using a single attribute only, which can be an ineffective traversal strategy.

c) In our experiments, the accuracy of GSL was tested with VGHs applied in the anonymization of tabular data (one of the most used formats in data sharing). Nonetheless, the applicability of our solution can be broader, as VGHs are one of the most used resources in generalization to protect privacy in different types of data. For example, in semantic trajectory data [51], VGHs are used to hide sensitive places where a person has stopped (e.g., an oncology clinic); while in transactional data [30, 70] VGHs are used to hide sensitive items purchased by a person (e.g., pregnancy test), or to hide web search queries performed by a person (e.g., adult websites). Furthermore, the type of attributes explored in this work was categorical due to their increasing importance (in recent years) as a valuable source of information for data mining. Since GSL is currently based on the preservation of data semantics as a means to properly process textual data, it is only applicable to categorical values. However, an interesting area of potential future work that might be explored is the use of numerical measures that could be incorporated to GSL (e.g., numerical distances or the size of the generalization intervals) in order to support, in an integrated manner, numerical and categorical attributes.

d) In our experimental evaluation, the quality of the VGHs (represented by the GSL scores) and their effectiveness in anonymization (represented by the data utility metrics) have been correlated using the Spearman's rank order correlation coefficient. This analysis strategy was chosen because the relationship we aimed to assess (data utility and VGH quality) was monotonic, but not necessarily linear. Thus, we considered that the Spearman's coefficient was a good fit for our use case, allowing us also to compare across different types of metrics. A limitation of using a ranking strategy (such as the Spearman's coefficient) is that the actual values/magnitudes of the scores are not being compared. Therefore, even when there are minor differences between the GSL scores of two or more VGHs, these types of cases are not captured by the rankings. Alternatively, other analysis strategies that consider the actual values of the scores (e.g., Pearson correlation) could be used.

e) There are certain aspects of the VGH that can influence their performance in anonymization. For example, although taller VGHs tend to perform better, this is not always the case. From a semantic quality point of view, preserving the semantic consistency (of the modeled domain) in the VGHs matters the most, while other aspects, such as the structural properties of the VGH (e.g., its depth) and the defined privacy goal have a secondary influence. To illustrate this point, let us consider a scenario where there are two candidate VGHs to anonymize a specific attribute: one VGH is tall (e.g., 16 levels) but suffers a very high semantic lost at level 1; the other VGH is relatively flat (e.g., 4 levels) but suffers only a minimum semantic lost at level 1. Under this scenario, the utility of the anonymized data produced by the flat VGH would drastically outperform the tall one in all the anonymization cases that end at level 1. Moreover, if a VGH is well-defined (from a semantic point of view), it usually implies that the more granular the VGH is, the lower the information loss will be as levels are climbed up in the VGH (in comparison with the loss exhibited by a coarse VGH). However, one needs to take into account the inherent tradeoff that exists (when disseminating anonymized data) between the amount of information revealed and the usefulness of the data. Hence, the appropriate depth (i.e., level of granularity) for the VGH would be determined by the users according to their requirements. For example, accommodating diverse types of profiles for data recipients (which can involve different levels of trustworthiness) such as: interdepartamental, outsourced partners, or public in general.

f) In terms of the potential time-savings that APES can achieve (through the usage of GSL), they are directly related to the number of configurations that need to be tested. They consist of the number of candidate VGHs/VGH sets, the complexity of the domains modeled in the VGHs, the tested privacy constraints, and the number of VGHs that belong to the VGH set. Therefore, the biggest time savings are obtained when data publishers have various complex candidate VGHs per QID and also several QIDs are in-scope for anonymization (so that there will be candidate VGH sets). Under these conditions, APES is able to mitigate most of the significant effort required to identify the best VGHs for anonymization purposes. It is also worth mentioning that there will be gains even if the actual number of candidates VGHs is moderate. This is because APES significantly reduces the effort and expertise required to identify the best VGHs. Furthermore, as data publishers typically have to release the anonymized data on a recurrent basis, the capability of automatically identifying a priori the appropriate VGHs can substantially reduce the complexity of such iterative process.

g) In our experimental evaluation, we considered datasets widely-used in the literature. As our results have shown, the obtained positive results (in terms of accuracy, time-savings, and costs) are evident for all four datasets, and so it is expected that GSL can yield similar results when using other datasets. Likewise, it is expected that GSL should be applicable to other usage scenarios that rely on VGHs (or similar types of taxonomies). The same principle applies to the semantic similarity measures. Our results have shown that GSL worked well with both tested measures (i.e., Wu and Palmer, and Leacock and Chodorow). Therefore, it is expected that GSL can perform similarly when using other semantic similarity measures.

h) In our experimental evaluation, we used WordNet as the knowledge base. However, GSL is not tied to any specific ontology. As long as the appropriate semantic similarity libraries (e.g., WS4J for WordNet) are available (or developed) to interface

with the desired ontologies, our prototype can be easily adapted to use any ontology. This strategy can be useful to make GSL better suited to domains which highly rely on specific terminology (e.g., biomedicine). Furthermore, for those usage scenarios where relying on a single ontology is not enough, there are several techniques in the literature which are well-accepted solutions to address the limitations in coverage of an ontology. In particular, we consider useful those techniques which can integrate multiple ontologies in order to complement each other and generate a more complete source of knowledge. Similarly, for those cases in which only a few terms do not exist in the chosen ontology, mapping those non-existing terms to similar existing ones might suffice for estimating the quality of VGHs.

i) In terms of the costs of using GSL, our results have shown that calculating the metric is lightweight in terms of the amount of computational resources required. Moreover, the obtained results can be a useful input information for a capacity planning process. This would allow data publishers to estimate the system requirements required by an experimental environment to support a particular usage scenario or dataset.

j) Finally, based on all the previously discussed points, it is concluded that a metric that can evaluate a priori the quality of the VGHs before anonymizing the data, as GSL does, can offer significant benefits to data publishers and the anonymization process in general.

7 Conclusions And Future Work

VGHs play an important role in the utility of the anonymized data. This is because VGHs drive the generalization process in numerous privacy-preserving algorithms. Moreover, while the evaluation of VGHs for numerical attributes is well studied in the literature, techniques to assess the quality of categorical VGHs are scarce. To address this issue, in our previous work we presented GSL, a metric that captures the quality of VGHs for categorical data with respect to their semantic consistency and taxonomic organization. The aim of this paper was to comprehensively evaluate GSL in terms of its benefits (i.e., its accuracy and the achieved time-savings) and its costs (i.e., its computational resources) in order to offer data publishers a valuable reference regarding the performance of GSL. For this purpose, a prototype was built and a set of experiments were performed. The experimental results demonstrated the feasibility of using GSL to perform an a priori evaluation and selection of VGHs for anonymization. They showed how the utility of anonymized datasets was improved (without sacrificing the privacy goal) when the selection of the best VGH was based on GSL. This is because there was a strong positive correlation between the quality of the VGHs (represented by the GSL scores) and the utility of the anonymized data (evaluated in terms of general-purpose utility metrics and clustering quality). The results also demonstrated how GSL can enhance the efficiency of the anonymization process and the effectiveness of a VGH-based anonymization algorithm, by avoiding costly trial-and-error anonymization cycles. Finally, the results showed that GSL is lightweight in terms of its consumption of computational resources, making it practical for real-world usage.

There are different directions for future research in our work. One interesting direction is to define a set of categories (or ranges) for the GSL score that verbally describes the quality of a VGH (e.g., good/moderate/poor). This would serve as a guide for data publishers to help them better interpret the GSL score. For that purpose, we plan to evaluate our solution

further using more datasets and privacy models. Likewise, we also plan to evaluate the accuracy of GSL in other application-context scenarios, such as classification and aggregate query answering. Moreover, we intend to investigate which other aspects of PPDP might be suitable to extend our VGH evaluation solution. Finally, we plan to explore how to automatically generate well-defined VGHs (or improve an "imperfect" VGH) by leveraging the GSL metric.

8 Acknowledgments

This work was supported, in part, by Science Foundation Ireland grant 10/CE/I1855 to Lero - the Irish Software Research Centre (www.lero.ie). This work was also supported, in part, by Science Foundation Ireland grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Irish Software Research Centre (www.lero.ie). The authors thank the anonymous reviewers for their helpful comments and suggestions. We also thank Dr. A. Omar Portillo-Dominguez for the valuable discussions and reviews.

References

[1] Chicago Homicides. `https://data.cityofchicago.org`. Last accessed: 2017-03-04.

[2] Insurance Dataset. `https://github.com/ucd-pel/Datasets/tree/master/Insurance`. Last accessed: 2017-03-04.

[3] Medical Subject Headings. `http://www.nlm.nih.gov/mesh/`. Last accessed: 2017-03-04.

[4] UTD Anonymization ToolBox. `http://cs.utdallas.edu/dspl/cgi-bin/toolbox/`. Last accessed: 2017-03-04.

[5] Vatant, B., Wick, M.: Geonames ontology. `http://www.geonames.org/ontology/`. Last accessed: 2017-03-04.

[6] WS4J library. `https://code.google.com/p/ws4j/`. Last accessed: 2017-03-04.

[7] D. Abril, G. Navarro-Arribas, and V. Torra. Towards semantic microaggregation of categorical data for confidential documents. In *Int. Conf. on Modeling Decisions for Artificial Intelligence*, pages 266–276. Springer, 2010.

[8] S. K. Adusumalli and V. V. Kumari. An efficient and dynamic concept hierarchy generation for data anonymization. In *Int. Conf. on Distributed Computing and Internet Technology*, pages 488–499. Springer, 2013.

[9] C. C. Aggarwal and S. Y. Philip. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-Preserving Data Mining*, pages 11–52. Springer, 2008.

[10] S. G. Anco Hundepool, Aad van de Wetering, Ramya Ramaswamy, Luisa Franconi, Alessandra Capobianchi, Peter-Paul de Wolf, Josep Domingo, Vicenc Torra, Ruth Brand. μ-ARGUS User Manual version 3.2, 2003.

[11] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy. Ontology-based quality evaluation of value generalization hierarchies for data anonymization. In *Privacy in Statistical Databases*, 2014.

[12] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Transactions on Data Privacy*, 7(3):337–370, 2014.

[13] A. Ballatore, M. Bertolotto, and D. C. Wilson. The semantic similarity ensemble. *Journal of Spatial Information Science*, (7):27–44, 2014.

[14] M. Batet, A. Valls, and K. Gibert. Improving classical clustering with ontologies. In *World conference of the IASC, Japan*, pages 137–146, 2008.

[15] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *Int. Conf. on Data Engineering*, pages 217–228, 2005.

[16] J. Brank, M. Grobelnik, and D. Mladenić. A survey of ontology evaluation techniques. In *Conference on Data Mining and Data Warehouses*, 2005.

[17] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[18] T. Caliski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27, 1974.

[19] A. Campan, N. Cooper, and T. M. Truta. On-the-fly generalization hierarchies for numerical attributes revisited. In *Secure Data Management*, pages 18–32, 2011.

[20] J. Carriço, C. Silva-Costa, J. Melo-Cristino, F. Pinto, H. De Lencastre, J. Almeida, and M. Ramirez. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes. *Journal of Clinical Microbiology*, 44(7):2524–2532, 2006.

[21] C. Dai, G. Ghinita, E. Bertino, J. Byun, and N. Li. TIAMAT: a tool for interactive analysis of microdata anonymization techniques. *VLDB Endowment*, 2(2):1618–1621, 2009.

[22] M. D'Aquin and N. F. Noy. Where to publish and find ontologies? a survey of ontology libraries. *Web semantics (Online)*, 11:96–111, Mar. 2012.

[23] Y. Ding, D. Fensel, and M. Klein. The role of ontologies in eCommerce. In *Handbook on Ontologies*. Springer, 2004.

[24] J. Domingo-Ferrer and J. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *Trans. on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[25] J. Domingo-Ferrer, D. Sánchez, and G. Rufian-Torrell. Anonymization of nominal data based on semantic marginality. *Information Sciences*, 242:35–48, Sept. 2013.

[26] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[27] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. *Introduction to Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 2011.

[28] T. Guo, D. Schwartz, F. Burstein, and H. Linger. *Codifying Collaborative Knowledge: Using Wikipedia as a Basis for Automated Ontology Learning*, pages 289–310. Palgrave Macmillan UK, 2015.

[29] J. Hartmann, P. Spyns, A. Giboin, and D. Maynard. Methods for ontology evaluation. *Knowledge Web Deliverable D1.2.3*, 2005.

[30] Y. He and J. F. Naughton. Anonymization of set-valued data via top-down, local generalization. *VLDB Endowment*, 2(1):934–945, 2009.

[31] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[32] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *Int. Conf. on Knowledge Discovery and Data Mining*, pages 279–288, 2002.

[33] B. Kenig and T. Tassa. A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery*, 25(1):134–168, 2012.

[34] S. Kisilevich, Y. Elovici, B. Shapira, and L. Rokach. kACTUS 2: Privacy preserving in classification tasks using k-Anonymity. *Protecting Persons While Protecting the People*, 5661:63–81, 2009.

[35] F. Kohlmayer, F. Prasser, and K. A. Kuhn. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. *Journal of Biomedical Informatics*, 58:37–48, 2015.

[36] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense

identification. pages 265 – 283. The MIT Press, Cambridge, MA, 1998.

[37] S. Lee, S.-Y. Huh, and R. D. McNiel. Automatic generation of concept hierarchies using WordNet. *Expert Systems with Applications*, 35(3):1132–1144, 2008.

[38] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Int. Conf. on Management of Data*, pages 49–60, 2005.

[39] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Int. Conf. On Data Engineering*, page 25, 2006.

[40] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Int. Conf. On Data Engineering*, pages 106–115, 2007.

[41] T. Li and N. Li. Towards optimal k-anonymization. *Data & Knowledge Engineering*, 65:22–39, 2008.

[42] Lichman M. UCI Machine Learning Repository, 2013.

[43] D. Lindberg, B. Humphreys, and A. McCray. The unified medical language system. *Methods of Inf. in Medicine*, 32(4):281–291, 1993.

[44] Y. Liu, T. Wang, and J. Feng. A semantic information loss metric for privacy preserving publication. In *Database Systems for Advance Applications*, pages 138–152, 2010.

[45] B. Loh and P. Then. Ontology-enhanced interactive anonymization in domain-driven data mining outsourcing. In *International Symposium on Data, Privacy and E-Commerce*, pages 9–14, 2010.

[46] A. Machanavajjhala and D. Kifer. l-diversity: Privacy beyond k-anonymity. *Trans. on Knowledge Discovery from Data*, 1(1):3, 2007.

[47] S. Martínez, D. Sánchez, and A. Valls. Semantic adaptive microaggregation of categorical microdata. *Computers & Security*, 31(5):653–672, July 2012.

[48] S. Martínez, D. Sánchez, and A. Valls. A semantic framework to protect the privacy of electronic health records with non-numerical attributes. *Journal of Biomedical Informatics*, 46(2):294–303, Apr. 2013.

[49] S. Martínez, D. Sánchez, A. Valls, and M. Batet. Privacy protection of textual attributes through a semantic-based masking method. *Information Fusion*, 13(4):304–314, Oct. 2012.

[50] D. McCuinness. Ontologies come of age. In *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, pages 171–194. MIT Press, 2003.

[51] A. Monreale and R. Trasarti. C-safety: a framework for the anonymization of semantic trajectories. *Transactions on Data Privacy*, 4:73–101, 2011.

[52] M. Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71, 2012.

[53] R. Navigli. Word sense disambiguation. *ACM Computing Surveys*, 41(2):1–69, Feb. 2009.

[54] M. E. Nergiz and C. Clifton. Thoughts on k-anonymization. *Data & Knowledge Engineering*, 63(3):622–645, Dec. 2007.

[55] A. O. Portillo-Dominguez, P. Perry, D. Magoni, M. Wang, and J. Murphy. Trini: an adaptive load balancing strategy based on garbage collection for clustered java systems. *Software: Practice and Experience*, 2016.

[56] G. Poulis and A. Gkoulalas-Divanis. SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms. *Int. Conf. on Extending Database Technology*, pages 620–623, 2014.

[57] F. Prasser, F. Kohlmayer, and K. A. Kuhn. A benchmark of globally-optimal anonymization methods for biomedical data. In *Symposium on Computer-Based Medical Systems*, pages 66–71, 2014.

[58] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[59] T. R. Ringenberg and J. M. Taylor. Semantic anonymization of medical records. In *Int. Conf. on Systems, Man and Cybernetics*, pages 1450–1455, 2014.

[60] S. Romanosky, D. Hoffman, and A. Acquisti. Empirical analysis of data breach litigation. *Journal of Empirical Legal Studies*, 11(1):74–104, 2014.

[61] P. Samarati. Protecting respondents identities in microdata release. *Trans. on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[62] D. Sánchez, A. Solé-Ribalta, M. Batet, and F. Serratosa. Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain. *Journal of Biomedical Informatics*, 45(1):141–55, Feb. 2012.

[63] A. Solé-Ribalta, D. Sánchez, M. Batet, and F. Serratosa. Towards the estimation of feature-based semantic similarity using multiple ontologies. *Knowledge-Based Systems*, 55:101–113, 2014.

[64] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[65] S. Staab and R. Studer. *Handbook on ontologies*. Springer, Berlin, 2009.

[66] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[67] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Int. Conf. on World Wide Web*, pages 697–706. ACM, 2007.

[68] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst*, 10(05):571–588, Oct. 2002.

[69] T. Tassa, A. Mazza, and A. Gionis. k-concealment: An alternative model of k-type anonymity. *Transactions on Data Privacy*, 5:189–222, 2012.

[70] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. *VLDB Endowment*, 1(1):115–125, 2008.

[71] V. Torra. *Information fusion in data mining*, volume 123. Springer, 2013.

[72] V. K. Vatsavayi and S. K. Adusumalli. Cost effective dynamic concept hierarchy generation for preserving privacy. *Journal of Information & Knowledge Management*, 13(04):1450035, 2014.

[73] K. Wang and B. Fung. Anonymizing sequential releases. In *Int. Conf. on Knowledge Discovery and Data Mining*, pages 414–423. ACM, 2006.

[74] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

[75] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Association for Computational Linguistics*, pages 133–138, 1994.

[76] J. Xu, W. Wang, J. Pei, and X. Wang. Utility-based anonymization using local recoding. In *Int. Conf. on Knowledge Discovery and Data Mining*, pages 785–790, 2006.

[77] Y. Xu, T. Ma, M. Tang, and W. Tian. A survey of privacy preserving data publishing using generalization and suppression. *Applied Mathematics & Information Sciences*, 8(3):1103, 2014.

[78] H. Zakerzadeh and S. L. Osborn. Delay-sensitive approaches for anonymizing numerical streaming data. *Int. Journal of Information Security*, 12(5):423–437, 2013.

[79] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen. A hybrid approach for scalable subtree anonymization over big data using MapReduce on cloud. *Journal of Computer and System Sciences*, 80(5):1008–1020, 2014.

[80] L. Zhou. Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3):241–252, Mar. 2007.

Appendix A Semantic Similarity Metrics

The **Wu and Palmer (WUP)** metric [75] is given by (9):

$$Sim_{WuP}(c_1, c_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \tag{9}$$

where c_1 and c_2 are the two concepts to be compared, N_1 and N_2 denote the number of *is-a* links on the path from c_1 and c_2 respectively, to their least common subsumer (LCS), and N_3 denotes the number of *is-a* links on the path from the LCS to the root of the taxonomy. The score range is (0,1] (1 for identical concepts).

The **Leacock and Chodorow (LCH)** metric [36] takes into account the number of nodes N_p on the shortest path between the two concepts (c_1 and c_2), and the maximum depth D of the taxonomy in which they occur. This metric is given by (10):

$$Sim_{LCH}(c_1, c_2) = -\log \frac{N_p}{2D} \tag{10}$$

Appendix B Data Utility Metrics

B.1 Task-Independent Measures

Semantic Information Loss (SemILoss). The overall SemILoss score (based on the equation presented in [49]) for an anonymized table T^* is given by (11):

$$SemILoss(T^*) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} sdist(x_{ij}, x_{ij}^*)}{n \cdot m} \tag{11}$$

where n is the number of records in the dataset, m is the number of QID attributes, x_{ij} is the original value of the jth attribute in the ith record, and x_{ij}^* is the anonymized value.

Semantic Sum of Squared Errors (Semantic SSE). The overall Semantic SSE score [25] for an anonymized table T^* is given by (12):

$$SemanticSSE(T^*) = \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} sdist(x_{ij}, x_{ij}^*)}{m} \right)^2 \tag{12}$$

where n is the number of records in the dataset, m is the number of QID attributes, x_{ij} is the original value of the jth attribute in the ith record, and x_{ij}^* is the anonymized value.

B.2 Task-Specific Measures

Clustering Comparison Methodology. To carry out the comparison, we firstly anonymized the original data, T, using each of the candidate VGHs, VGH_i, and obtained the anonymized datasets, T_i^*. Then, for each dataset, we generated a matrix with the semantic distance among the objects in the dataset. This matrix was used as the input for a hierarchical clustering method c which was applied to T and obtained clusters $c(T)$. We applied the same clustering method to T_i^* and obtained $c(T_i^*)$. Finally, we compared the similarity between the optimal cluster partitions obtained from $c(T)$ against the ones obtained from $c(T_i^*)$. The idea is that the more similar $c(T_i^*)$ is to $c(T)$, the lower the information loss. In this manner,

we could determine which VGHs produced the solutions that best retained the semantic utility of the original data. In the following paragraphs, we describe the metrics used to compare the quality of the sets of clusters between the original dataset (partition A) and the datasets anonymized with each candidate VGH (partition B).

Pairwise Agreement Metrics. The coefficients in these metrics are calculated based on a mismatch matrix [31]. This is built upon the different scenarios in which a pair of data points can fall: a, the number of point pairs that are in the same cluster in both A and B; b, the number of point pairs that are in the same cluster in A but not in B; c, the number of point pairs that are in the same cluster in B but not in A; or d, the number of point pairs that are in different clusters in A and B. Intuitively, $a + d$ are considered as the number of agreements between partitions A and B; and $b + c$ are the number of disagreements between partitions A and B. The pairwise agreement metric used in this work was the Rand index [58]. For these metrics, higher values are better: 0 indicates that the two data clusters do not agree on any pair of points and 1 indicates that the data clusters are exactly the same. The Rand index represents the ratio of agreement between both matches and mismatches. It is given by (13):

$$\mathcal{R}_{AB} = \frac{a+d}{a+b+c+d} \tag{13}$$

Entropy-Based Metrics. These metrics are built upon concepts from information theory to measure how much information is shared between partitions of clusters. In this work, we used the Normalized Mutual Information (NMI) metric [66], which is expressed by (14a):

$$NMI_{AB} = \frac{2 \cdot I(A,B)}{H(A) + H(B)} \tag{14a}$$

The agreement between two partitions, measured by the mutual information I is given by (14b):

$$I(A,B) = \sum_{i=1}^{R} \sum_{j=1}^{Q} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{n_i n_j / N^2} \tag{14b}$$

where R and Q denote the number of clusters in partitions A and B respectively, and n_{ij} denotes the number of shared patterns between clusters $C_i \in A$ and clusters $C_j \in B$.

The entropy H of the partition A is computed taking the frequency counts as approximations for probabilities. It is expressed by (14c):

$$H(A) = -\sum_{i=1}^{R} \frac{n_i}{N} \log \frac{n_i}{N} \tag{14c}$$

where n_i represents the number of patterns in cluster $C_i \in A$. The entropy H of the partition B is calculated in the same manner as in Eq. 14c.

The values of NMI range between 0 and 1. Larger values of NMI indicate a higher similarity between the partitions.