Improving the Utility of Anonymized Datasets through Dynamic Evaluation of Generalization Hierarchies

Vanessa Ayala-Rivera*, Thomas Cerqueus†, Liam Murphy*, Christina Thorpe*
*Lero@UCD, School of Computer Science, University College Dublin, Ireland
†Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
e-mail: vanessa.ayala-rivera@ucdconnect.ie, thomas.cerqueus@insa-lyon.fr, liam.murphy@ucd.ie, christina.thorpe@ucd.ie

***Abstract*—The dissemination of textual personal information has become a key driver for innovation and value creation. However, due to the possible content of sensitive information, this data must be anonymized, which can reduce its usefulness for secondary uses. One of the most used techniques to anonymize data is generalization. However, its effectiveness can be hampered by the Value Generalization Hierarchies (VGHs) used to dictate the anonymization of data, as poorly-specified VGHs can reduce the usefulness of the resulting data. To tackle this problem, we propose a metric for evaluating the quality of textual VGHs used in anonymization. Our evaluation approach considers the semantic properties of VGHs and exploits information from the input datasets to predict with higher accuracy (compared to existing approaches) the potential effectiveness of VGHs for anonymizing data. As a consequence, the utility of the resulting datasets is improved without sacrificing the privacy goal. We also introduce a novel rating scale to classify the quality of the VGHs into categories to facilitate the interpretation of our quality metric for practitioners.**

***Keywords*-Privacy, Data Publishing, Data Quality, Generalization Hierarchies, Data Semantics, Anonymization**

I. Introduction

Currently, the volumes of data generated globally grow exponentially every year. Within this data, there is a large amount of textual personal information, such as socio-economic data and health care records. This fact has attracted the interest of companies in various sectors (e.g., insurance companies, retailers) to collect this data for creating new business models or delivering better services. This data has also become a fundamental resource for other organizations (e.g., research institutions), who can reuse this data beyond its original purpose. For this reason, data sharing has become a driver for innovation and value creation. However, this data may contain sensitive information about individuals (e.g., medical conditions, religious beliefs) that can bring harm to the involved parties if the disclosure of this information occurs. For example, individuals may suffer from discrimination or identity theft. Likewise, organizations may suffer from negative publicity, fines or other sanctions [18]. Hence, this data must be anonymized before being shared for analysis. Privacy-Preserving Data Publishing (PPDP) provides methods for publishing data without compromising the confidentiality of individuals, while trying to retain the utility of the data for a variety of tasks (e.g., to feed data mining models, perform query answering, create decision-support systems) [4]. Since all the potential usage scenarios for the data are commonly unknown at the time of publication, the produced anonymization solution should be useful enough to be adequately exploited by multiple data recipients. One widely-used technique of anonymization is *generalization*, which consists in replacing the original values of an attribute in a dataset with others that are less precise but semantically consistent [21]. The idea is that the original data loses its specificity, which reduces the probability of re-identifying the individuals contained in the published datasets. For example, one could generalize the terms *"oncology clinic"* and *"cosmetic surgery hospital"* to *"medical institution"* to protect the sensitive whereabouts of a person.

A common prerequisite of generalization algorithms is the use of *Value Generalization Hierarchies* (VGHs) to drive the anonymization. VGHs are tree-like structures that contain the original values of an attribute and their set of candidate generalizations, which constitute the anonymization solution space for the attribute. VGHs are usually created and evaluated by the data publishers (i.e., anyone involved in the dissemination of data in a safe and useful manner; hereinafter referred as *users*). Users often follow an iterative process for assessing the concepts and appropriate details to represent in the VGHs. This can yield multiple candidates of VGHs (for each attribute) that can be used for anonymization. Then, the users have to choose which VGH will be used to anonymize each attribute. All these tasks are often performed manually and rely on users' judgment.

A key problem of this practice is that the quality of the VGHs is evaluated in a subjective and informal way. Regardless this problem, the "correctness" of the VGHs that feed generalization algorithms is an aspect that is rarely questioned in the PPDP literature. This is because it is commonly assumed that users are fully capable of providing the adequate domain expertise to the VGHs based on their own knowledge and experience [12], [23]. To mitigate possible issues, knowledge engineers are often involved in the evaluation process. However, the process may become expensive due to the limited availability of

subject-matter experts and the time-consuming manual labor it requires. In any case, the decision about the quality of VGHs generally represents the subjective opinion of a single individual, and thus corresponds to only one interpretation of a domain. Clearly, the tasks of creating and evaluating VGHs for anonymization are challenging and the current practices are not effective [5], [22]. Likewise, the chosen VGHs play a key role in the utility remaining in the anonymized data and in the precision of the analysis performed, as both can decrease if poorly-specified VGHs are used in the data generalization.

Contributions. Considering these challenges, our research has centered on developing techniques to evaluate in an objective, quantifiable, and automatic way, the quality of textual data VGHs with the aim of improving their effectiveness for anonymizing data. In [5] we introduced the *Generalization Semantic Loss* metric (s-GSL), which captures the quality of a VGH with respect to its semantic consistency and taxonomy. However, s-GSL is based on a static strategy. That is, it assumes that, for a given set of domain values, there is an optimum "one-size-fits-all" VGH that would suit any input dataset containing such values. Hence, it does not exploit information from the input datasets. This makes s-GSL sensitive to data sparseness, which reduces its accuracy and limits its applicability. Also, s-GSL lacks a qualitative interpretation. In this paper we extend our previous work [5] by proposing an enhanced VGH quality metric (d-GSL) which, based on a dynamic evaluation scheme, exploits the frequency distribution of the input datasets. In this manner, d-GSL can predict with higher accuracy the effectiveness that VGHs will have in anonymization. This enables users to select the best VGH (among all candidates) for a particular dataset. Hence, the utility of the anonymized datasets will be better preserved without sacrificing the privacy goal. We also experimentally show how d-GSL enhances the process of VGH evaluation. Finally, we propose a rating scale to help users to classify the VGHs based on their quality (w.r.t. d-GSL) into categories. Each category includes an interval and a qualitative descriptor to offer practitioners an intuitive interpretation of the d-GSL metric.

The remainder of this paper is organized as follows: Section II provides the background and related work. Section III presents the proposed d-GSL metric and its rating scale. Section IV explains the methodology and the evaluation criteria selected for our experiments. Section V discusses our experimental results. Section VI provides a final discussion of our work. Finally, Section VII presents our conclusions and future work.

II. Background and Related Work

The collection and dissemination of large amounts of personal information have allowed organizations to benefit from the exploitation of this data. Therefore, a lot of effort has been invested into studying PPDP techniques. In a typical PPDP scenario [4], a trusted organization collects data about individuals for their own business reasons. However, this data may be shared in a sanitized form with third parties or in the public domain under various circumstances (e.g., commercial use, research, legal reasons). The essence of PPDP is to release anonymized datasets that have good utility for a variety of tasks.

Generalization is a commonly used technique in PPDP to achieve privacy. One advantage of generalization is that (unlike perturbation techniques that apply noise to data) it preserves the truthfulness of the data. This property is achieved by using VGHs. An example VGH is shown in Fig. 1. The leaves at level 0 (L0) represent the original distinct values of an attribute in the 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. VGHs are usually specified for the quasi-identifier (QID) attributes of a dataset, which are those that can be linked to external information and lead to the re-identification of people in published anonymized datasets.

Although 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) [17] with the aim of providing formal privacy guarantees. In our work, we focus on k-anonymity, a widely-adopted privacy model that consists in making each record indistinguishable from a group of at least *k-1* other records [21].

Example 1. To illustrate generalization-based k-anonymization, consider Table I showing a table with socio-economic records. Among the attributes, *name* is the identifier (ID), *occupation* is the QID, and *salary* is the sensitive attribute (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 II shows a 3-anonymous version of Table I. Note that the generalization created two groups of records that share the same QID value. Within each group, individuals are indistinguishable from each other.

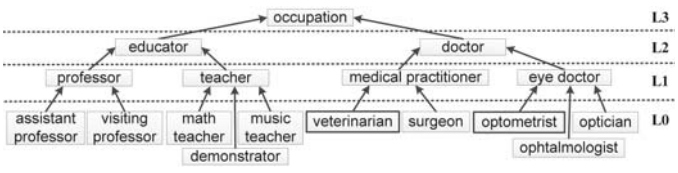

Figure 1. VGH for *occupation*

Table I
EXAMPLE SOCIO-ECONOMIC DATASET

	ID	QID	SA
Record#	**Name**	**Occupation**	**Salary**
1	Bob	Music Teacher	51 000
2	John	Surgeon	278 000
3	Clare	Veterinarian	86 000
4	Alice	Math Teacher	47 000
5	Owen	Assistant Profesor	75 000
6	Jack	Optician	107 000

Table II
A 3-ANONYMOUS VERSION OF TABLE I

	QID	SA
Record#	**Occupation**	**Salary**
1	Educator	51 000
4	Educator	47 000
5	Educator	75 000
2	Doctor	278 000
3	Doctor	86 000
6	Doctor	107 000

The creation and evaluation of VGHs for numerical attributes have been well studied in the literature [6], [22]. For numerical attributes, it is relatively easy to evaluate if the data quality has been preserved after anonymization (e.g., by retaining their statistical properties, or by minimizing the size of an interval). In contrast, little research has been done to study the quality of textual data VGHs. Previous studies have discussed the role that VGHs play in the utility of anonymized data [6], [11], [16], [22]. They indicate that a good VGH would improve the usefulness of the data, whereas a poor VGH would reduce the precision of the data. Although these works have helped us 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.

A closely related work 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 [20]. 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). However, 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 [8], [13]. However, their 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 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 reuses 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 [7]. Moreover, multiple ontologies can be integrated to complement each other and have a more complete source of knowledge [19].

III. PROPOSED APPROACH

In this section, we motivate the use of data semantics for evaluating textual data VGHs. We also discuss the details of our proposed solution and the advantages of using a dynamic strategy over a static one. Finally, we present a rating quality scale that serves to better interpret d-GSL.

A. Semantics Data Preservation in Textual Data

Data semantics is an implicit aspect of textual data. However, traditional metrics used in PPDP for measuring the amount of generalization in textual attributes do not usually consider it. Instead, metrics that are better suited for numerical attributes are typically used [16], [21]. For example, a common approach is to map each textual value to a numeric one, then the amount of data distortion occurred by generalization is quantified by the length of the interval in which the original values have been grouped [16]. Although these types of metrics capture a certain level of data distortion, they do not capture the loss in the meaning of textual values. For example, consider that *illness* is generalized to *disease* (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. Likewise, the loss when *cafeteria* is replaced by *restaurant* should be lower than replacing it by *tavern*, as the first two concepts are more similar. Considering this, we have incorporated the use of data semantics in our VGH evaluation approach with the aim of capturing more accurately the possible loss of information incurred in the VGH. This would help to identify those VGHs that better retain the semantics of the original data. As a result, the anonymized data will also better retain its semantic usefulness, which would enable users to extract more useful conclusions.

B. Static vs. Dynamic Selection of VGHs

In the traditional anonymization process, the design and selection of VGHs are performed the first time that a dataset needs to be published by an organization. If a new dataset that belongs to the same domain (which has already been modeled) requires anonymization, the selected VGH usually remains static. This is because, as discussed in Section I, the complexity and cost of the VGH creation process can be significant, and thus may be unsuitable for an organization to carry it out periodically. Thus, the static VGH (which was the best according to the evaluation process) is used to perform the anonymization of subsequent datasets; assuming that such VGH will be the best regardless of the input dataset. However, using a single static VGH for all anonymizations may not be the best strategy in PPDP. This is because the effectiveness of the VGHs (and the utility of the anonymized data) can be impacted by the characteristics of the input datasets. For this reason, our work proposes to use a dynamic strategy in the evaluation of the candidate VGHs. That is, VGHs should be evaluated depending on the input dataset, as the decision about the "best" VGH can change when the distribution of the datasets is considered.

C. Dynamic VGH Evaluation Scheme Overview

In our work, the quality of the VGHs is represented by a score, which dynamically adjusts to the distribution of the input datasets. This strategy would allow a more accurate evaluation of the VGHs and thus, a better prediction of their effectiveness to conduct the data anonymization. Fig. 2 depicts the contextual view of our solution in PPDP: (1) Organizations collect personal information and are required to share it under different circumstances. However, these datasets must be anonymized before being disseminated. (2) The users choose the QIDs to be generalized from the datasets. (3) For each QID, users create a set of candidate VGHs modeling the given domain and evaluate them based on their own knowledge and experience. In the traditional anonymization process, these tasks are performed manually, so they are time-consuming and error-prone. Considering this, users need to have an efficient and effective manner of assessing candidate VGHs to decide which VGH best fits each input dataset. (4) Our proposed approach is a dynamic VGH evaluation scheme which captures, in a score denoted as d-GSL, the degree of data semantics that each VGH loses in their specification. The lower the d-GSL score, the less information loss incurred in the VGH. Our solution integrates knowledge bases and semantic similarity metrics. The knowledge base is implemented in the form of ontologies, which act as a gold standard in which the domain expert knowledge is reflected. Ontologies often represent the consensus opinion of a panel of experts thus, we mitigate the risk of having partial interpretations and single judgments over the domains represented in the VGHs. The semantic content of the ontologies is exploited by semantic similarity metrics to measure the proximity between the original values of a dataset and its possible generalizations; resembling human behavior regarding the judgment of similarity between two terms. (5) The output of our approach is one d-GSL score per evaluated VGH, which allows the users to compare the quality of the candidate VGHs and select the best for each data scenario (e.g., those ranked #1). Our solution also defines a rating scale in which the d-GSL score can be mapped to a qualitative category to ease the interpretation of d-GSL for practitioners. (6) After evaluation, the best VGHs (in terms of d-GSL) can then be used to feed the algorithm that anonymizes the data with more guarantees that the chosen VGHs will help to better preserve the semantics of the original data, hence, retain the data usefulness.

D. Computing the Dynamic GSL Score

In order to assess the quality of a VGH in terms of its semantics preservation and the input dataset, we evaluate the degree of information loss that could result from the transformations of the original values in the input dataset. To perform this, we first measure the semantic distance between the original values at the leaf nodes $L = \{l_1, l_2, \ldots, l_n\}$ and their corresponding candidate generalizations at the ancestor nodes of each level *i* of the VGH $A^i = \{a_1^i, a_2^i, \ldots, a_k^i\}$. To consider also the dataset distribution, we consider the frequency of occurrence of the original values $F = \{f_1, f_2, \ldots, f_n\}$. The semantic loss caused by the generalization of a value to its ancestor in level i is given by (1):

$$D_j^i = f_j \cdot SemDist(l_j, a^i) \tag{1}$$

where l_j is the j-th leaf node of the VGH, f_j is the number of times that the l_j node value appears in the input dataset, and a^i is the ancestor of l_j in level i.

The score representing the quality of a VGH, V, is captured by the *Dynamic Generalization Semantic Loss* (d-GSL). For this measure, lower values are better, as it would mean a lower semantic loss. The d-GSL score is given by (2):

$$d\text{-}GSL(V) = \sum_{i=1}^{h} w_i \cdot \mathcal{G}(D_1^i, \ldots, D_n^i) \tag{2}$$

where i is the index of a level in the VGH, h denotes the height of the VGH, w_i is a predefined weight associated with the level i, and $\mathcal{G}$ is an aggregation function applied to all the D_j^i distance scores. The weights should be specified such that they sum up to 1. They can be used to consider taxonomical characteristics of the VGHs (e.g., height) into the evaluation score. For example, a generalization occurring in a coarse-grained VGH would cause a higher information loss than one in a fine-grained VGH. Moreover, weights can also be used to penalize the abstraction/specificity of the terms in the VGH. For example, to magnify the differences in semantics for the generalization at the lower levels of the

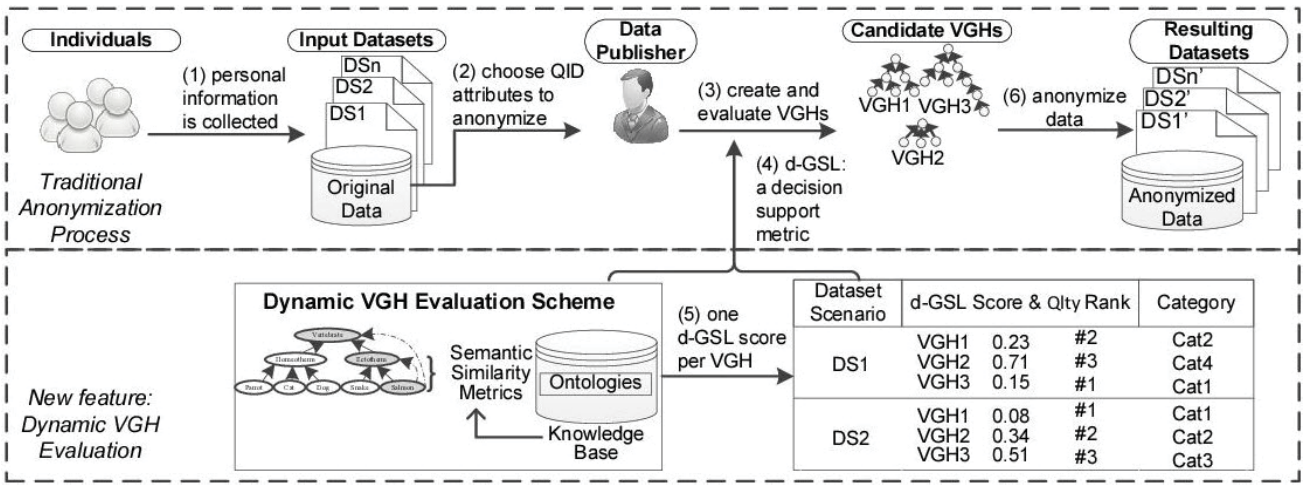

Figure 2. Contextual View of Dynamic VGH Evaluation Scheme in PPDP

VGH, where the more specific levels are found. The weights can be given by any of the equations in (3):

$$w_i = \begin{cases} \dfrac{1}{h}, & \text{if } w_i \text{ is uniform} \quad (3a) \\ \dfrac{(h+1-i)}{\sum_{j=1}^{h} j}, & \text{if } w_i \text{ is based on level } i \quad (3b) \end{cases}$$

The function $\mathcal{G}(D_1^i, \ldots, D_n^i)$ can be any aggregation mechanism that allows to combine the D_j^i scores into a single representative value for each level of a VGH, with the aim of assessing their quality. The definition of the function $\mathcal{G}$ is based on the intended analysis of the users. For example, to identify the individual generalizations that are causing the maximum losses (maximum value as in Eq. 4a), or to compare the overall quality of each level of the VGH (average value as in Eq. 4b). This is useful because under some schemes of anonymization, the generalization occurs at the domain level, that is, all the original values end at the same level of the VGH (e.g., full-domain generalization [4]).

$$\mathcal{G}(s_1, ..., s_n) = \begin{cases} \max\limits_{i \in [i,n]} s_i, & \text{if } \mathcal{G} \text{ is maximum} \quad (4a) \\ \dfrac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} f_i}, & \text{if } \mathcal{G} \text{ is average} \quad (4b) \end{cases}$$

Example 2. To illustrate the benefits of leveraging data semantics for VGH evaluation, let us consider a dataset where the selected QID to anonymize is *occupation*. The user has defined (based on her own knowledge) the VGH shown in Fig. 1 to perform the generalization of the QID. Ideally, more general terms are located at higher levels in the VGH and more specialized terms are lower in the VGH. However, there are imprecisions in the VGH which might not be easy to identify at first sight. For example, the ancestor terms specified for *"optometrist"* are semantically inconsistent. This is because although those terms refer to people involved in eye caring, *"optometrist"* is not an *"eye-doctor"* or *"doctor"*. These type of issues can be easily detected by inspecting those transformations using our semantic-based approach. For example, the semantic loss caused by the transformations of *"optometrist"* to *"eye-doctor/doctor"* (computed with Eq. 1) are 0.3846 and 0.3333, respectively; which indicate a high loss. In contrast, the ancestors specified for *"veterinarian"* are correct, as the transformations to *"medical practitioner/doctor"* incur lower losses, which are 0.0769 and 0.037, respectively.

Table III
EVALUATION OF VGH SHOWN IN FIG. 1

Quality Scores (using WordNet and Wu-Palmer metric)		
Dataset Scenarios	**s-GSL**	**d-GSL**
DS1: uniform	0.3403	0.3403
DS2: 90%, 10%	0.3403	0.4744
DS3: 50%, 40%, 10%	0.3403	0.3803

Example 3. To illustrate the benefits of evaluating VGHs using a dynamic scheme (such as d-GSL) over a static one (such as s-GSL), let us consider three datasets (DS) containing 100 records that need to be anonymized. These DS belong to the same domain but have different frequency distributions (as shown in the first column of Table III): DS1 has a uniform distribution of 10% of each term; DS2 has 90% of *"optometrist"* terms (poorly-specified branch) and 10% divided among the rest; DS3 has 50% of *"veterinarian"* terms (well-specified branch), 40% of *"optometrist"*, and 10% divided among the rest. Next, consider that we first evaluate the quality of the VGH using s-GSL. In this case, we obtain the same quality score independently of the input dataset (as seen in the s-GSL column). In contrast, when using the d-GSL metric (computed with Eq. 2), we obtain different scores per dataset, as d-GSL considers the impact

Table IV
RATING SCALE FOR THE D-GSL METRIC

Category	Ranges	Descriptor
Cat1	$0.0 \leq$ Quality Score < 0.2	Very Good
Cat2	$0.2 \leq$ Quality Score < 0.4	Good
Cat3	$0.4 \leq$ Quality Score < 0.6	Moderate
Cat4	$0.6 \leq$ Quality Score < 0.8	Poor
Cat5	$0.8 \leq$ Quality Score ≤ 1.0	Very Poor

of the data distribution to evaluate the quality of the VGH. For DS1, both quality scores yield the same result as the impact of having a single instance of each term is the same as having *n* instances. For DS2, d-GSL showed an increment in the loss incurred by the VGH, as in this scenario 90% of the data falls into a branch where the losses are high. Finally for DS3, d-GSL showed a small increment as the data is distributed between well-defined and poorly-specified branches. Having obtained different d-GSL scores for the same VGH shows the importance of considering information from the input datasets in the VGH evaluation process. Moreover, it shows that our dynamic scheme can capture more accurately the different dataset scenarios compared to the static scheme.

As shown by the above examples, our approach brings multiple benefits to the users in the evaluation of VGHs, such as: (1) Inconsistent specifications introduced in the VGH can be easily identified (e.g., misclassifications, redundancies); (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; (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.

E. VGH Quality Categories

We propose a rating scale consisting of five categories to classify the VGHs according to their d-GSL score. The aim is that these categories serve as a guide for the users to know what to expect about the utility of the data anonymized with the VGHs. These categories are shown in Table IV. They were inspired by the rule of thumb used for interpreting correlation coefficients (e.g., pearson, spearman), which is composed of a 5-point scale to offer a fair and intuitive range of qualitative descriptors. Furthermore, the ranges of the categories were derived from the VGH behaviors observed in our empirical evaluation. Each category has an ordinal scale (i.e., the descriptor) which qualitatively expresses the ratings of quality, so that practitioners can better interpret the d-GSL metric. Lower categories are better as they indicate that a VGH has a lower semantic loss, thus retaining more information in their specification. The ranges of the categories cover the interval of [0,1], which matches how the d-GSL is expressed. This is because we used the Wu and Palmer metric to measure semantic similarity; if an alternative metric is used, the range of the metric would only need to be normalized to the [0,1] interval.

IV. EXPERIMENTAL SETUP

Below, we present our experimental methodology and describe the testbed of VGHs, the datasets, and the evaluation criteria used in our experiments.

A. Experimental Methodology

We conducted a series of experiments that pursued three objectives: (1) to investigate how the effectiveness of a VGH is subjective to the dataset that will employ it for anonymization; (2) to demonstrate how and why using a dynamic VGH quality assessment scheme, which considers the distribution of the input datasets, is better than a static scheme; and (3) to demonstrate how a rating scale can be applied to classify the quality of VGHs.

For this purpose, we created a set of candidate VGHs modeling the same domain for a socio-economic attribute (see Section IV-B). We then evaluated the quality of those VGHs using the d-GSL and s-GSL metrics. s-GSL was chosen as the rival metric for comparison as it represents, to the best of our knowledge, the first quantitative mechanism to assess the quality of VGHs to perform data anonymization. Later, we conducted the anonymization of datasets (see Section IV-B) using the candidate VGHs and the Datafly algorithm (a popular k-anonymity based algorithm) [21]. We tested different levels of privacy, varying the k-values $\in$ [30..100]. In this manner, all the anonymity levels of the candidate VGHs were covered, which guaranteed a fair comparison. Finally, we calculated the usefulness of the anonymized datasets using task-independent data utility metrics (see Section IV-C).

B. Evaluated VGHs and Datasets

Our testbed consisted of 252 VGHs. Those VGHs were created by perturbing the semantical content of an *ideal* VGH which was constructed by extracting the minimal taxonomy from WordNet (ontology widely used due to its broad coverage of concepts [14]) for our evaluation data. To obtain a varied range of VGH quality scores for our tests, we applied transformations to the ideal VGH, such as: mixing the leaf nodes and replacing the terms of the ancestors with another one selected from a list of candidate terms extracted from WordNet (that are within a semantic similarity boundary). All VGHs were constructed over the same set of leaf concepts. As evaluation data, we used the *Insurance* dataset [2] which contains personal information (in tabular format) that can be of interest to an insurance company for carrying out a risk assessment on potential clients. From this dataset, we focused on the attribute of *occupation*, which has the highest diversity of values. To test the generality of our solution, we derived multiple datasets

per VGH. This strategy allowed us to considerably diversify the range of evaluated scenarios and to show the impact that a dataset can have over the anonymization performance of the VGHs. To achieve this, for each VGH we computed the semantic loss that each branch (from leaf to root node) incurred in the VGH. We then identified the branches with the minimum and maximum semantic losses. Based on these branches, we generated 10 datasets (each composed of 1,100 records) per VGH. The name assigned to each dataset reflected the frequency distributions of the worst branch, the best branch, and the rest of the terms in the dataset. For example, for the dataset 70w20b10r, the 70% of the data was the worst branch, while 20% was the best, and 10% was distributed among the rest of the terms.

C. Evaluation Criteria

VGH Quality. The quality of the VGHs is expressed in terms of our d-GSL metric and the s-GSL metric. As both metrics leverage on measures of semantic similarity, in our experiments we used two widely-used path-based metrics: Wu and Palmer (WUP) and Leacock and Chodorow (LCH) [14]. This strategy allowed us to prove the generality of the d-GSL with respect to the used semantic similarity metric. The WUP metric measures the depth of two given concepts in the taxonomy and the depth of their least common subsumer. The LCH metric measures the length of the shortest path between two concepts considering the depth of the taxonomy. Our implementation used the WS4J library [3], which relied on WordNet 3.0 to compute the semantic similarity between two words.

Data Utility. The level of usefulness that remained in the datasets after anonymization was measured using two task-independent data utility metrics (as not knowing in advance the analysis task is an essential premise of PPDP): Semantic Sum of Squared Errors (SSE) [8] and Semantic Information Loss (SemILoss) [13]. SSE measures the level of intra-group homogeneity in a group of anonymized records. SemILoss measures how semantically different the anonymized values are (on average) compared to the original ones. For both metrics, lower values are better, as it indicates a higher utility of the data.

VGH Quality and Data Utility Correlation. To analyze the degree of correlation between the scores representing the quality of VGHs and the utility of the anonymized datasets, we calculated the Spearman's rank order correlation (r_{Spm}). It measures the strength of a monotonic (but not necessarily linearly related) relationship between paired data. r_{Spm} can take values from -1 to +1; the closer the value is to ± 1, the stronger the relationship.

VGH Quality Categories. Based on our proposed rating scale, we created clusters based on both VGH quality scores, and compared them against the clusters created based on the data utility scores. In this manner, we could determine which of the two cluster partitions (those based on d-GSL or those based on s-GSL) was more in agreement with the clusters obtained from the data utility scores. The idea is that the higher the agreement between the partitions, the higher the accuracy of the quality metric. This was measured using two representative clustering evaluation metrics: the Wallace coefficient (Wallace) and the Normalized Mutual Information (NMI) metric [9]. Wallace 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 tool available at [1].

V. Experimental Results

Correlation Comparison. This analysis focused on assessing the capacity of the d-GSL and s-GSL metrics to capture (a priori) the effectiveness of the VGHs for anonymizing data. Figs. 3a and 3b show the correlation between the VGH quality metrics and the data utility per dataset. Fig. 3a shows the correlation calculated using WUP to compute the VGH quality scores, and SSE to quantify the data utility. The d-GSL metric obtained stronger correlations (between 0.91 and 0.98) than the ones obtained for s-GSL (between 0.64 and 0.81). This means that d-GSL was more precise in determining the effectiveness of VGHs. Similar results are observed in Fig. 3b. It shows the correlation calculated using LCH to compute the VGH quality scores, and SemILoss to quantify the data utility. Results show that d-GSL continued to exhibit a stronger correlation (between 0.79 and 0.90) than s-GSL (between 0.73 and 0.85).

To complement the validation, the highest and the lowest correlations obtained for each quality metric are shown in Figs. 4 and 5 (for the 50w40b10r and 05w90b05r dataset scenarios, respectively). It can be observed how the relationship between the quality of the VGHs and the utility of the anonymized data exhibits a more linear and positive monotonic trend for the d-GSL metric.

In conclusion, this analysis demonstrated that d-GSL is a better metric than s-GSL as it provides a more accurate representation of the VGHs' quality.

Data Distribution Sensitivity Analysis. This analysis centered on assessing the influence of the input datasets (with different frequency distributions) on the effectiveness of the VGHs. For the sake of brevity, we only present the comparison of ten sample VGHs (randomly picked from our testbed) that belong to the category 2 (according to their s-GSL score). Fig. 6a shows the d-GSL and s-GSL scores of the sample VGHs. Since s-GSL is computed using a static scheme, there is one s-GSL score per VGH. For this reason, we use the s-GSL score as the baseline for our comparisons. In contrast, d-GSL uses a dynamic scheme, thus each VGH

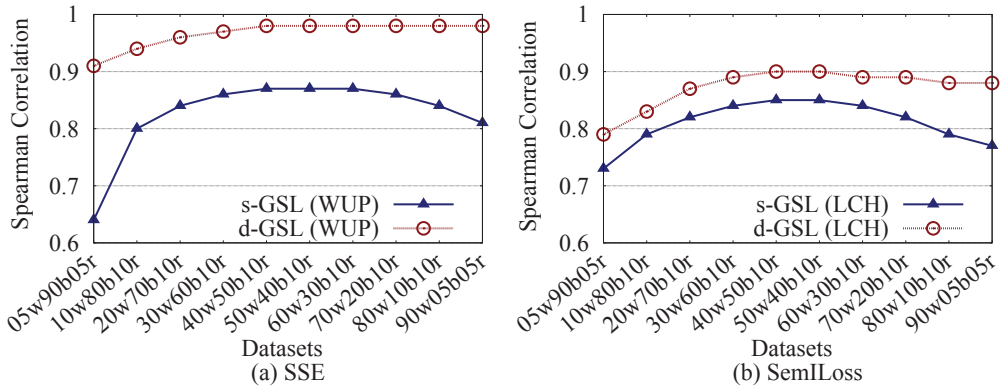

Figure 3. Correlation Comparison of Quality Scores vs (a) SEE and (b) SemILoss Data Utility

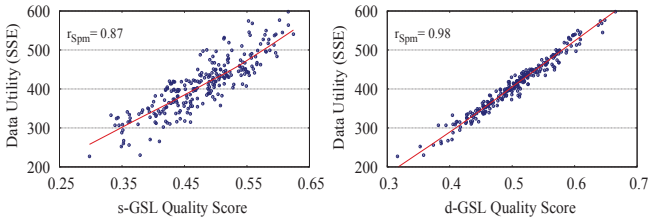

Figure 4. VGH Quality Scores vs Data Utility in the 50w40b10r dataset (Highest r_{Spm})

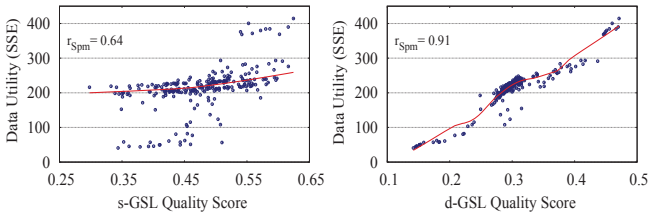

Figure 5. VGH Quality Scores vs Data Utility in the 05w90b05r dataset (Lowest r_{Spm})

has one d-GSL score per input dataset. It can be seen how the s-GSL scores remained within the range of a single quality category (i.e., [0.3 to 0.39] corresponding to cat2), whereas the d-GSL scores fluctuated among four categories (i.e., [0.14 to 0.67] corresponding to cat1 to cat4). Based on their quality scores (per dataset), the VGHs were ranked to identify the best VGH among all candidates. That is, the VGH that would yield the lower information loss in the anonymized data. These rankings are shown in Fig. 7a. For example, based on d-GSL and the 90w05b05r input dataset, the VGH3 would be the best. Whereas, if the input dataset were 05w90b05r, the best would be VGH2. In summary, Figs. 6a and 7a represent the expected anonymization performance of the VGHs based on their d-GSL and s-GSL quality scores.

Figs. 6b and 7b show the real performance of the sample VGHs. That is, the information loss that the datasets suffered after being anonymized with the VGHs. Fig. 6b shows the data utility magnitudes (in terms of SSE), whereas Fig. 7b shows the data utility rankings of the VGHs for each dataset scenario. To validate the accuracy of the quality metrics (d-GSL and s-GSL) for anticipating the utility of the anonymized data, we compared the expected rankings (shown in Fig. 7a) with the real rankings (shown in Fig. 7b). Overall, the d-GSL metric predicted more accurately the performance of the VGHs than s-GSL. For example in Fig. 7a, s-GSL ranked the VGH2 in 2nd place, however its performance degraded with some datasets (e.g., 70w20b10r and 90w05b05r) until falling to 9th place. In contrast, this scenario is well-captured by d-GSL, as the d-GSL scores exhibit the same trends in both quality (Fig. 7a) and utility plots (Fig. 7b). We also compared the expected magnitude trends (Fig. 6a) with the real magnitude trends (Fig. 6b). In this case, we only validated that the trends remained, as the correlation between VGH quality and the utility offered by the VGHs is not linear (as previously discussed in our correlation results).

In conclusion, this analysis demonstrated that there is no "best-fit-for-all" VGH for all datasets. Instead, the best VGH can change depending on the input datasets. Hence, the utility of an anonymized dataset can be improved when the VGH evaluation is performed using a dynamic scheme (i.e., d-GSL) instead of a static scheme (i.e., s-GSL).

VGH Quality Categories Comparison. This analysis assessed the empirical categories we proposed in Section III-E. Fig. 8 shows the average data utility of the VGHs grouped by quality categories based on their d-GSL score. For the sake of brevity, we only present the results for three datasets, as similar behaviors were obtained for the others. Within each dataset, the highest utility was obtained for the Cat1 as the VGHs belonging to this group reduced the information loss the most, in comparison to the rest of the categories. This demonstrates that a VGH that belongs to a lower category would be more effective to perform the anonymization of data than one that belongs to a higher category. It can also be noticed that the information losses are generally lower

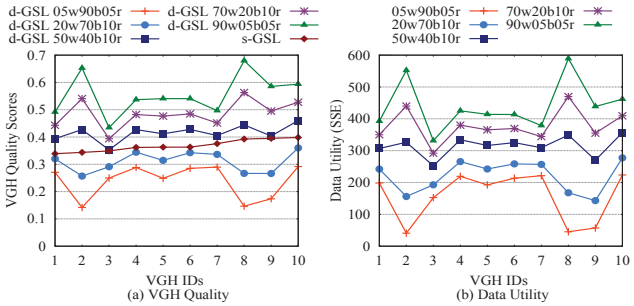

Figure 6. Magnitudes for (a) VGH Quality Scores and (b) Data Utility Scores

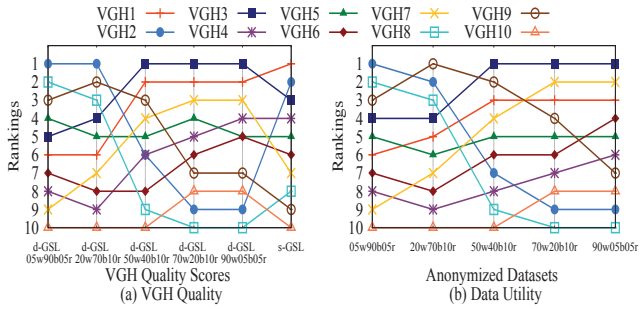

Figure 7. Rankings of (a) VGH Quality Scores and (b) Data Utility Scores

in the 05w90b05r dataset. This behavior is explained by the fact that this dataset contains the highest proportion of well-defined terms (i.e., those that belong to the best branch, as explained in Section IV-B).

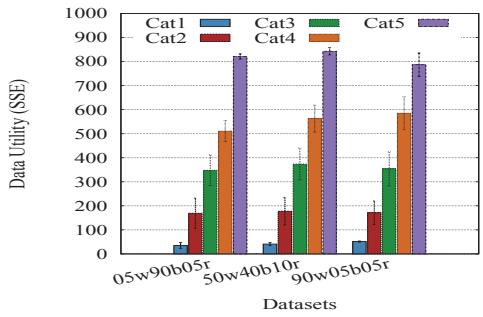

Figure 8. Utility for d-GSL based Categories

We also grouped the VGHs in categories based on their s-GSL scores; their behaviors were similar to the ones exhibited by the d-GSL categories (shown in Fig. 8). To validate which of the two category groups was more accurate, we compared the groups ("clusters") based on the d-GSL and s-GSL scores, against the groups created based on the data utility scores. That is, we evaluated if the VGHs are "classified" in the same way when they are grouped by their quality than when they are grouped by the data utility. Then, we measured the agreement between the groups. Figs. 9a and 9b depict the level of agreement between the data utility

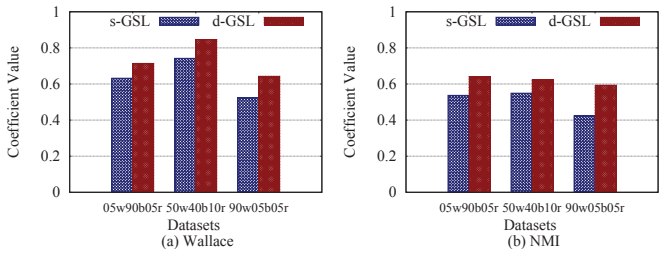

Figure 9. Category Agreement for s-GSL and d-GSL w.r.t. (a) Wallace and (b) NMI

clusters and the VGH quality clusters measured with Wallace and NMI. The VGH classification based on d-GSL always outperformed the one based on s-GSL. This means that s-GSL could have overestimated or underestimated the real data utility yield by the VGHs in each category.

In conclusion, this analysis proved that VGHs classified in lower categories are more likely to yield a higher data utility than those in higher categories. We also showed that the agreement level between the categories created based on quality and data utility was higher when the classification was performed with d-GSL (rather than s-GSL). That is, the expected VGHs' effectiveness indicated by d-GSL was more in accordance with their real effectiveness.

VI. Final Discussion

In our empirical evaluation, the effectiveness of d-GSL was tested with VGHs applied in the anonymization of tabular data (one of the most used formats in data sharing). However, the applicability of our solution can be broader, as VGHs are the most used approach in generalization to protect privacy in different types of data. For example, in semantic trajectory data [15], VGHs are used to hide sensitive places where a person has stopped (e.g., an oncology clinic); while in transactional data [10], they are used to hide sensitive items in purchases (e.g., pregnancy test).

Our approach uses an a priori strategy for evaluating the quality of VGHs. That is, the potential effectiveness of VGHs is estimated before anonymizing the data, which allows to save users' time and prevent applications from using inappropriate VGHs. Regarding the privacy implications of our approach, it does not affect the privacy goal set by users. In our experiments, the anonymized datasets improved their utility while still satisfying their corresponding k-anonymity levels. Thus, the anonymized datasets kept the same protection and vulnerabilities of the privacy model used. In our experiments, we used k-anonymity. However, our approach is not tied to a specific privacy model or its associated goal. This decision was made not only to make d-GSL independent of the privacy model but also because it is not known a priori how many generalizations will be needed to satisfy a privacy goal (e.g., k-anonymity level), thus our assumption is that it can be equally satisfied (same probability) at any level of the VGH.

VII. Conclusions and Future Work

In this paper, we proposed a dynamic VGH evaluation approach which exploits the frequency distribution of the input datasets. Our approach yields a score (d-GSL) which acts as predictor of the effectiveness of VGHs to perform the anonymization of data. The d-GSL score enables users to effectively compare multiple VGHs for a given domain and select the one that will better retain the usefulness of the original data. We also proposed a rating scale that will help users to classify the VGHs based on their quality (in terms of d-GSL) into categories. Each category includes an interval and a qualitative descriptor to offer practitioners an intuitive interpretation of the d-GSL. Our results demonstrated that the utility of anonymized datasets is improved (without sacrificing the privacy goal) when the selection of the best VGH is based on a dynamic scheme (i.e., d-GSL) instead of a static scheme (e.g., s-GSL). Our results also demonstrated that d-GSL is more accurate than s-GSL, as it was better correlated with the utility of the datasets anonymized with the evaluated VGHs. Furthermore, we showed that VGHs classified in lower categories are more likely to yield a higher data utility than those in higher categories.

In our future work, we intend to investigate which other aspects of PPDP might be suitable to extend our VGH evaluation solution. Likewise, we plan to broaden the validation of d-GSL applying task-specific utility metrics (e.g., data mining). Another interesting idea is to consider the potential vulnerability of the VGHs to different privacy attacks. We plan to use this additional knowledge to develop a global cost evaluation function which can assess VGHs from different perspectives. It would be also interesting to evaluate our solution further using more datasets and privacy models. Finally, we also intend to explore how to automatically generate well-defined VGHs or improve an "imperfect" VGH.

Acknowledgments

This work was supported, in part, by Science Foundation Ireland grant 10/CE/I1855. We also thank A. Omar Portillo-Dominguez for the helpful discussions and reviews.

References

[1] Comparing Partitions. http://www.comparingpartitions.info.

[2] Insurance Datasets. https://github.com/ucd-pel/Datasets/tree/master/Insurance.

[3] WS4J library. https://code.google.com/p/ws4j.

[4] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy. A Systematic Comparison and Evaluation of k - Anonymization Algorithms for Practitioners. *Trans. on Data Privacy*, 7(3):337–370, 2014.

[5] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy. Ontology-Based Quality Evaluation of Value Generalization Hierarchies for Data Anonymization. In *PSD*, 2014.

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

[7] M. D'Aquin and N. F. Noy. Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *Web semantics (Online)*, 11:96–111, 2012.

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

[9] N. Grozavu, G. Cabanes, and Y. Bennani. Diversity analysis in collaborative clustering. In *IJCNN*, pages 1754–1761, 2014.

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

[11] T. Li and N. Li. Towards optimal k-anonymization. *DKE*, 65:22–39, 2008.

[12] B. C. S. Loh and P. H. H. Then. Ontology-Enhanced Interactive Anonymization in Domain-Driven Data Mining Outsourcing. In *ISDPE*, pages 9–14, 2010.

[13] 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, 2012.

[14] L. Meng, R. Huang, and J. Gu. A Review of Semantic Similarity Measures in WordNet. *Int. J. of Hybrid Information Tech*, 6(1):1–12, 2013.

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

[16] M. E. Nergiz and C. Clifton. Thoughts on k-anonymization. *DKE*, 63(3):622–645, 2007.

[17] A. M. Pinto. A comparison of anonymization protection principles. In *IRI*, pages 207–214, 2012.

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

[19] 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.

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

[21] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzz.*, 10(05):571–588, 2002.

[22] V. K. Vatsavayi and S. K. Adusumalli. Cost effective dynamic concept hierarchy generation for preserving privacy. *J. Info. Know. Mgmt.*, 13(04):1450035, 2014.

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