Automatic Construction of Generalization Hierarchies for Publishing Anonymized Data

Vanessa Ayala-Rivera, Liam Murphy, and Christina Thorpe

Lero@UCD, School of Computer Science, University College Dublin, Dublin, Ireland
vanessa.ayala-rivera@ucdconnect.ie,{liam.murphy,christina.thorpe}@ucd.ie

Abstract. Concept hierarchies are widely used in multiple fields to carry out data analysis. In data privacy, they are known as Value Generalization Hierarchies (VGHs), and are used by generalization algorithms to dictate the data anonymization. Thus, their proper specification is critical to obtain anonymized data of good quality. The creation and evaluation of VGHs require expert knowledge and a significant amount of manual effort, making these tasks highly error-prone and time-consuming. In this paper we present AIKA, a knowledge-based framework to automatically construct and evaluate VGHs for the anonymization of categorical data. AIKA integrates ontologies to objectively create and evaluate VGHs. It also implements a multi-dimensional reward function to tailor the VGH evaluation to different use cases. Our experiments show that AIKA improved the creation of VGHs by generating VGHs of good quality in less time than when manually done. Results also showed how the reward function properly captures the desired VGH properties.

1 Introduction

Microdata (i.e., records about individuals) is a valuable resource for organizations. By exploiting it, companies acquire knowledge to improve or create new business models. For this reason, many organizations are actively collecting and publishing data. However, data must be anonymized before being shared for analysis as it may contain sensitive personal information (e.g., medical conditions) that can bring harm to the involved parties if it is disclosed (e.g., negative publicity, fines, identity theft). Privacy-Preserving Data Publishing (PPDP) offers methods for publishing data without compromising individuals' confidentiality, while trying to retain the data utility for a variety of tasks [5].

k-Anonymity is a fundamental principle to protect privacy in the release of microdata [5,21]. It requires that each record appears at least with k occurrences with respect to the quasi-identifiers (QIDs), i.e., attributes that can be linked to external information and reidentify individuals in anonymized datasets. *Generalization* is the most widely used technique to achieve k-anonymity [21]. It consists in replacing the original QIDs' values with less precise (but semantically consistent) ones, reducing the risk of reidentification (e.g., *"surgeon"* with *"doctor"*). Generalization is usually conducted using concept hierarchies, known as *Value Generalization Hierarchies* (VGHs), which indicate the transformations that an attribute can undergo. Fig. 1 shows an example of a VGH. The leaves (L0) correspond to the real values of an attribute in the dataset, and the ancestors (L1 to L3) correspond to the candidate values used for generalization.

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-47650-6_21

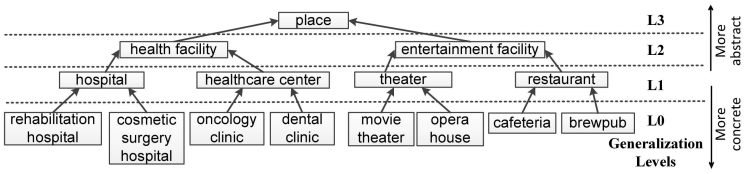


Fig. 1: A VGH for the attribute *place*

VGH design is a burdensome process for data publishers (i.e., people involved in the dissemination of data in a safe and useful manner; hereinafter referred as *users*) as one VGH needs to be created per QID, based on the input dataset. If the input values change, VGHs must be modified accordingly, which requires additional manual effort. While it is feasible to create VGHs of small size, the effort considerably increases when larger VGHs are required (e.g., open-ended surveys), or in scenarios where data constantly changes (e.g., streaming data). To tackle this issue, various approaches to generate VGHs automatically have been proposed [8]. However, most of them are designed for numerical attributes, while methods applicable to categorical data remain scarce. Numerical approaches often consist in creating intervals that fit the distribution of the input data. Thus, they are not suitable for categorical data, as its inherent semantics is ignored (a key factor to preserve its meaning). The construction of categorical VGHs presents even more challenges [12]: Disambiguation of the concepts' senses, defining meaningful labels to represent clustered lower level concepts, etc.

Traditionally, categorical VGHs are designed by users based on their own knowledge and experience, as it is commonly assumed that they are fully capable of bringing adequate domain expertise to the construction of VGHs [8]. A key problem of this practice is that the quality of VGHs is evaluated in a subjective and informal way. This issue can lead to misclassifications or inconsistencies which significantly impact the quality of the anonymized data. To mitigate this issue, knowledge engineers often participate in the evaluation process. However, the process may become expensive due to the limited availability of experts and the laborious work involved. Consequently, the design of VGHs is normally a highly error-prone and time-consuming process.

Considering these challenges, our paper has the following contributions:

1. A knowledge-based framework (AIKA) to automatically construct and evaluate categorical VGHs for anonymization, which considers users' preferences.
2. A comprehensive practical evaluation of AIKA, consisting of a prototype and a set of experiments to assess the benefits of AIKA for the creation and evaluation of VGHs for anonymization, as well as the costs of using AIKA.
3. A case-study comparing the quality and efficiency of the VGHs generated by AIKA against VGHs manually created.

2 Related Work

Several methods for creating "good" VGHs (i.e., those that yield a good utility in the data after anonymization) have been proposed in literature. However, most of them focus on numerical attributes. For instance, the authors of [8]

presented an approach for creating numerical hierarchies on-the-fly based on agglomerative hierarchical clustering. In general, these approaches are unsuitable for categorical data, as semantics is ignored. Most of the existing work focusing on categorical data belongs to the field of knowledge engineering. There, various techniques exist to create concept hierarchies whose aim is usually to facilitate the understanding of documents and processes, or to enhance semantic interoperability [11, 22]. However, their direct applicability 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. Alternatively, some authors [10, 15] have proposed the use of ontologies (instead of VGHs) to anonymize data. However, this can bring significant restrictions to anonymization. For instance, ontologies cannot be easily tailored to diverse publishing scenarios. Also, the fine granularity of ontologies can overexpose information to an adversary. For these reasons, our work only uses ontologies as a source of knowledge for the creation and evaluation of VGHs; leveraging the fact that various large and consensus ontologies have been made available [9].

3 AIKA Framework

In this section, we provide the context of our solution and describe the methods proposed for the automatic construction and evaluation of VGHs for PPDP.

3.1 Overview

To address the need for assisting the users in the design of VGHs, we followed a typical design science research approach [17] to develop our solution. It consists of a knowledge-based framework (AIKA) for the automatic construction and evaluation of VGHs to be used in data anonymization. Our goal is to offer a mechanism that not only reduces the human effort and expertise required to design and evaluate VGHs, but also improves the quality of the generated VGHs. Fig. 2 depicts the contextual view of AIKA in PPDP: (1) A trusted entity collects personal data and is required to publish it. Thus, datasets must be anonymized before being disseminated. (2) The user selects the QIDs to be generalized from the datasets. (3) For each QID, the user manually creates candidate VGHs modeling their corresponding domains. (4) Once the user is confident about the created VGHs, they are used to anonymize the data. (5) The user then evaluates the utility and disclosure risk of the anonymized data. (6) If they are acceptable, the data is released. Otherwise, a new anonymization cycle starts (Step 3). AIKA fits into Step 3, where the VGHs are designed. (3a) AIKA consists of two components: a *constructor* and an *evaluator*. The constructor (see Section 3.2) automatically generates various candidate VGHs for a particular domain by exploiting information from a knowledge base and the original dataset. Note that the constructor does not generate a single "optimal" VGH but a set of VGHs

that can fulfill the needs of different use cases. The candidate VGHs are passed to the evaluator (see Section 3.3), where the VGHs are objectively assessed with quantifiable metrics from multiple perspectives. (3b) The user can inspect the VGHs and adjust (or re-evaluate) them as needed. (4) After evaluation, the best VGHs can be used to drive the data anonymization with more guarantees that those VGHs will help to retain the desired level of data usefulness and disclosure risk (hence eliminating the need of costly trial-and-error anonymization cycles).

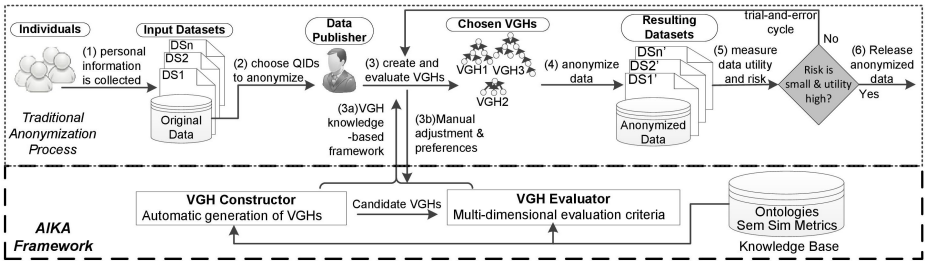


Fig. 2: Contextual View of AIKA Framework in PPDP

3.2 VGH Constructor

The constructor consists of a method that automatically generates and tailors VGHs, based on the input datasets and a knowledge source that models the domain expert knowledge and human judgment. Next, we describe the elements involved in the VGH construction process (depicted in Fig. 3).

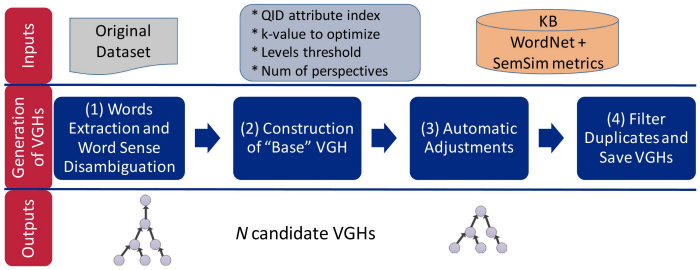


Fig. 3: AIKA - VGH Constructor

Knowledge Base (KB). AIKA exploits a KB to perform various tasks such as: semantic relationships exploitation, word-sense disambiguation, and measurement of similarity between words. The KB is encapsulated in ontologies, which act as a gold standard in which the domain expert knowledge is reflected. Since ontologies often represent the consensus opinion of a panel of experts, the risk of having partial interpretations and single judgments over the domains represented in the VGHs is mitigated. The semantic content of the ontologies is exploited by semantic similarity metrics to measure the proximity between the original values of a dataset and their possible generalizations. In this work we use WordNet as

KB, and Wu and Palmer as semantic similarity metric (two of the most widely used resources in knowledge-based systems) [16]. Note that relying on a single ontology does not represent a limitation for AIKA, as several works support the integration of ontologies [20]. Also, the ontology used by AIKA is configurable.

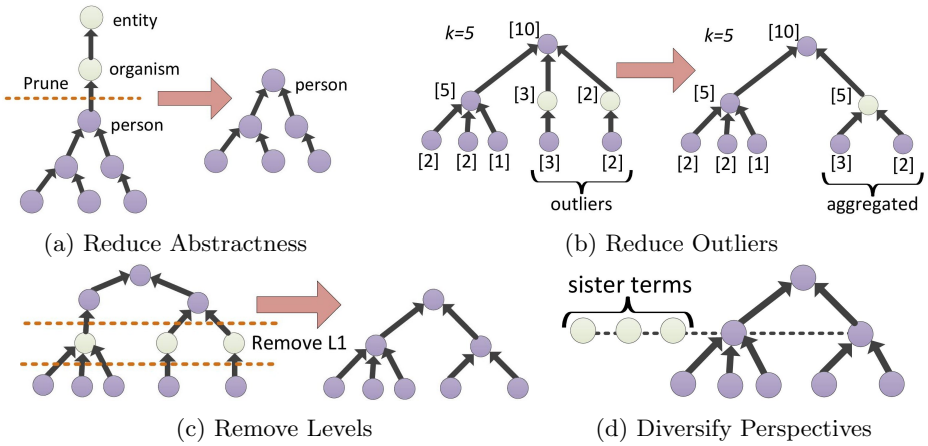
(1) Words Extraction and Word Sense Disambiguation (WSD). First, the constructor identifies the leaf nodes of the VGH (by extracting the distinct values of the QID from the input dataset), and calculates their frequencies of occurrences. Next, WSD is performed, which involves defining the right sense for the words. In AIKA we use the adapted Lesk algorithm [7] which is a gloss-based method that relies on the definition of a word (using WordNet as gloss dictionary). This technique is suitable for microdata anonymization because there is no background context that can be used (e.g., documents or corpus). To mitigate the possibility of any remaining noise (i.e., incorrect senses), AIKA allows users to provide (or adjust) the senses of the individual participant words.

(2) Construction of "base" VGH. To start the generation of VGHs, AIKA extracts the minimal hierarchy that subsumes all the leaf values from the ontology. That is, for each leaf, it extracts the hypernym tree from WordNet. Then, all branches are merged into the "base" VGH. This VGH forms the basis for all other candidate VGHs, which will be later derived from it. Using the subsumption hierarchy is appropriate in our scenario as it reflects the principle of specialization/generalization used by data generalization techniques.

(3) Automatic Adjustments. This step consists in applying a series of automatic transformations to the "base" VGH with the objective of deriving multiple candidate VGHs that can be used to fulfill the requirements of different use cases. This is because the released anonymized data is intended to be used by multiple parties for different purposes. In general, such transformations vary the taxonomic structure and degree of data semantics of the "base" VGH, hence the characteristics of the derived candidate VGHs are diversified. Below, we describe the different types of adjustments performed by the VGH constructor:

a. Reduce abstractness (Fig. 4a) prunes the hierarchy at the lowest level where all the branches are connected. This adjustment naturally meets the monotonicity property [13] extensively used in anonymization: if the generalization T^* at level i preserves privacy, then every generalization of T^* at level $i+1$ also preserves privacy. That is, all successors of an anonymous state are also anonymous.

b. Reduce outliers (shown in Fig. 4b) avoids over-generalizing the data by reducing the possible outliers in the VGH (e.g., due to data sparseness). The aim is to tailor the VGHs for a given syntactic privacy model (e.g., k-value for k-anonymity) so that the privacy condition is satisfied at the lowest possible level (where the information loss is lower). When it is possible (i.e., the frequency sum of the outliers is $\geq k$ and the semantic consistency of the VGH is respected), the outliers can be aggregated into groups so that k is satisfied. The new node (common ancestor of a group of outliers) can be one of three possibilities: one of the parents of the outliers; one of the outliers, promoted as parent; or the root node replicated (implying the full suppression of the values). All these alternatives are viable depending on the data anonymization scenario.



(a) Reduce Abstractness (b) Reduce Outliers

(c) Remove Levels (d) Diversify Perspectives

Fig. 4: Automatic Adjustments in VGH Constructor

c. Reduce levels (shown in Fig. 4c) removes full generalization levels in the VGH based on a desired threshold of taxonomic levels to be preserved. The aim is to make the VGHs taller (fine-grained) or flatter (coarse-grained), depending on how the user wants to refine the anonymizations. This is useful to manipulate the level of safety of the anonymized data, as fine-grained usually means a higher utility but also a higher risk of disclosure. Hence, diverse profiles of data recipients (with different trustworthiness levels) can be supported such as: releasing data to an outsourced partner, or public in general.

d. Diversify perspectives (shown in Fig. 4d) applies facets to the candidate generalizations by organizing the concepts in alternative ways. The aim is to offer different perspectives about a domain. The "base" VGH is mainly created using the subsumption relationship (*is-a*) in an ontology. However other semantic relationships can be used, such as: meronym (e.g., *part-of, substance-of*) and sibling (e.g., *sister terms*). For example, *animals* can be organized in *vertebrate* and *invertebrate* but also in *ectotherm* and *homeotherm*, depending on the user's needs. The nodes to be replaced by a facet are the ancestors. The feasibility of a concept to be considered as a facet is given by a semantic similarity boundary, which determines its relevance with respect to the ancestor to be replaced.

(4) Filter Duplicates and Save VGHs. Once the automatic adjustments have been applied, the constructor cache has VGHs of diverse characteristics (potentially including repeated ones). Thus, this step consists in filtering out the duplicate VGHs so that only unique VGHs are preserved. To identify if a VGH is "equal" to another one, we use a filtering strategy based on adjacency matrix representation. Unlike techniques based on graph traversing, this approach allowed to accurately capture the equality of two VGHs for the PPDP context. That is, two VGHs qualify as equal if they have the same nodes connected in the taxonomy; even if the branches are not arranged in the same order. Finally, the unique candidate VGHs are saved (in XML format) into disk so they can be inspected by the user to be adjusted, evaluated, or used in anonymization.

3.3 VGH Evaluator

The evaluator (shown in Fig. 5) consists of a method for the multi-dimensional evaluation and ranking of VGHs. It is based on the combination of a set of metrics that capture, in an objective and quantifiable way, the quality of VGHs from different perspectives that are relevant in anonymization. The input is a list of candidate VGHs and a set of weights that represent the user's preferences with respect to the evaluation metrics. The usage of weights allows the evaluation phase to be tailored to assess a particular use case. A KB (described in Section 3.2) is also used to evaluate the semantic similarity between attribute values.

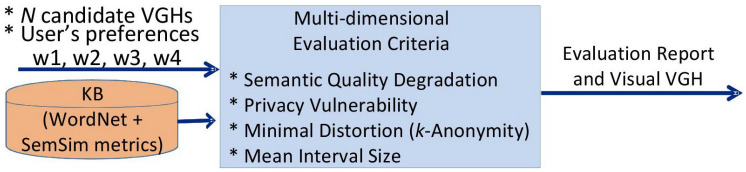


Fig. 5: AIKA - VGH Evaluator

Multi-dimensional Evaluation Criteria. In the following paragraph, we describe the evaluated VGH aspects and how they are measured: (1) *Semantic quality degradation* assesses the proper specification of the VGH concepts in terms of semantics. For this purpose, we apply the generalization semantic loss metric [6], which captures the quality of a VGH in terms of the semantic consistency and taxonomic organization. (2) *Privacy vulnerability* assesses the VGH susceptibility to inferences by attackers. For this purpose, we apply the semantic variance metric [19], which measures the semantic dispersion of the modeled concepts. The idea is that fine-grained taxonomies can give more information to an adversary and make the data vulnerable to attacks, thus, the more spread the concepts are, the more the privacy vulnerability. Considering this aspect helps to exclude extremely detailed VGHs. (3) *Minimal distortion* captures the ratio between the minimum level of the VGH at which the syntactic privacy condition (e.g., k-value) is satisfied, and the total height of the VGH. This metric was inspired by the minimal distortion principle discussed in [21]. (4) *Mean interval size* captures the average size of the generalization intervals (i.e., average number of children of the ancestor nodes); the more coarse the groups are, the more indistinguishable the original values are and thus, the more the information loss.

As first step, the evaluator calculates the scores of all aspects, per VGH. Then, the VGHs are ranked (per aspect) in order to obtain a reward equivalent to their ranking position in descending order. That is, VGHs with a low score will obtain a higher reward. This is because lower values are better for all evaluation metrics. Next, the user preferences (weights) are applied to the given reward, and an overall score for a VGH is calculated by the function $E(V)$ given by (1):

$$\mathcal{E}(V) = w_1 \cdot semq + w_2 \cdot priv + w_3 \cdot distrn + w_4 \cdot isize \quad (1)$$

where $semq$, $priv$, $distrn$, $isize$ are the aspects evaluated in the VGH according to the multi-dimensional criteria; and w_1, w_2, w_3, w_4 are the weights assigned by

the user to indicate the importance of each aspect. The best VGH is the one that maximizes $E(V)$ given the chosen weights. This is given by (2):

$$f(\mathcal{E}(V)) = \max(\mathcal{E}(V))|w_1, w_2, w_3, w_4) \quad (2)$$

4 Experimental Evaluation

The experiments aimed three objectives: (1) to assess the benefits of using AIKA (i.e., its capability to create good quality VGHs and estimate their effectiveness in anonymization); (2) to assess the costs of using AIKA (in terms of computational resources); and (3) to compare AIKA's benefits and costs against those of manually generated VGHs. As evaluation data, we used four publicly available datasets: Adult [14] consists of census information; German Credit [14] contains credit applicants information; Chicago Homicide [1] has information about homicides filed by the Chicago police; Insurance [2] contains personal information useful for risk assessment. For each dataset, we chose the categorical attributes with the most heterogeneous values as QIDs (Table 1) to diversify the tested domains; then, we generated VGHs for them using AIKA. To assess the performance of the VGHs in anonymization, we used the commonly-used anonymization algorithm Datafly [21] (from the UTD Anonymization Toolbox [3]). We also tested a broad range of privacy levels, varying the k-values $\in$ [2..100]. All experiments were done in a computer with an Intel Core i7-4702HQ CPU at 2.20Ghz, 8GB of RAM, Windows 8.1 64-bit, and HotSpot JVM 1.7 with a 1GB heap. Finally, AIKA's prototype was developed in Java, internally using the WS4J library [4].

Table 1: QIDs considered for VGH creation and anonymization

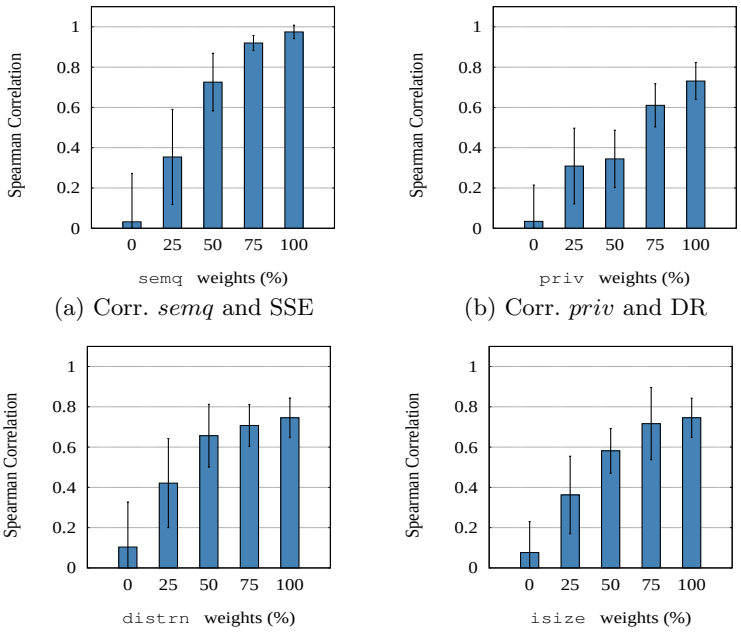
Dataset	Attribute	Card.	Col Index	Dataset	Attribute	Card.	Col Index
HomicideVictims	Location	96	46	Adult	Occupation	14	7
	PHome	11	48	GermanCredit	Purpose	12	3
	POutdoor	33	56	Insurance	Occupation	60	3
	CausalFactor	47	59		Workplace	29	4
	VicRelation1	95	71		Hobby	40	5
	OffRelation1	95	72		PlaceOfHobby	32	6
	WClub	57	106				
	WKnife	25	109				

AIKA's benefits. This analysis focused on assessing the quality of the VGHs, by measuring their effectiveness for anonymizing datasets (our use case). For this purpose, we firstly evaluated the VGHs using AIKA's multi-dimensional criteria (Eq. 1). We tested the full spectrum of weights (i.e., [0..100%]) in increments of 25% per aspect. This strategy involved 35 sets of weights, one for each possible weight permutation and the four aspects (e.g., w_1=75,w_2=25,w_3=w_4=0). This allowed us to rank the VGHs from best to worst per weighted aspect. Next, we conducted the anonymization of the datasets using the VGHs and calculated the usefulness of the resulting datasets using four utility/risk metrics (each one associated with a desired aspect of the VGH). To measure the data utility, we used three commonly-used task-independent metrics: Semantic Sum of Squared Errors (SSE) [10], Generalized Information Loss (GenILoss) [5], and Average Equivalence Class Size (C_{AVG}) [5] which are related to the *semq*, *distrn*, and

isize aspects, respectively. To measure the data disclosure risk (DR), we used record similarity [15], associated with the *priv* aspect. Due to space constraint, we only present the most relevant results (as this experiment involved the generation/evaluation of approximately 1.4K VGHs and 138K anonymized solutions).

To assess how well the properties of the VGHs were captured by AIKA's evaluator, we calculated the degree of correlation between the VGH quality scores and the quality of the anonymized datasets. For this purpose, we used the Spearman's rank order correlation (r_{Spm}), which measures the strength of a monotonic (but not necessarily linear) relationship between paired data. r_{Spm} can take values from -1 to +1. The closer the value is to ± 1, the stronger the relationship. The results showed that AIKA worked well (Fig. 6), as a strong level of correlation (i.e., $r_{Spm} \geq 0.60$) was achieved by all metric/aspect combinations when a high weight was used (e.g., 75% and 100%). Fig. 6a shows the results of the *semq* aspect. There, it can also be noticed how the correlation level gradually decreases following a trend similar to the decrease in the *semq* weight. This is consequence of considering other aspects and exemplifies the trades-off that are experienced in anonymization (i.e., one sacrifices utility to enforce privacy). This behavior is also reflected in the standard deviations of the low weights, which tend to be higher than those of higher weights. Figs. 6b, 6c, and 6d depict the results of the other aspects. It can be noticed how the aspects behaved similarly, as they achieved comparable levels (and trends) of correlations.

To complement this analysis, an example of the correlation plots is shown in Fig. 7. It can be noticed how the VGH quality rankings closely resemble the



(a) Corr. *semq* and SSE (b) Corr. *priv* and DR

(c) Corr. *distrn* and GenILoss (d) Corr. *isize* and C_{AVG}

Fig. 6: Correlations between VGH evaluator criteria and data quality metrics

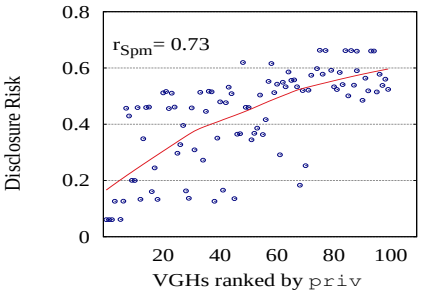


Fig. 7: *priv* (w_2=100%) vs DR

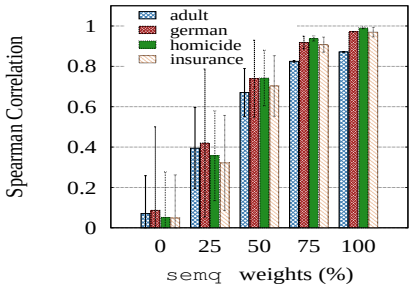


Fig. 8: Corr. *semq* and SSE per dataset

disclosure risk of the anonymized solutions. We also carried out a breakdown of the correlation results per dataset. No serious variations in the results were observed, showing that AIKA's rankings were accurate irrespectively of the dataset. This behavior is exemplified by Fig. 8, which present the correlation breakdown for the *semq* aspect. Similar results were obtained for the other aspects.

AIKA's costs. We also studied the costs of using AIKA, in terms of computational resources: memory consumption, CPU usage, and execution time. Garbage collection (GC) was also monitored as it is a key performance concern in Java [18]. Results showed that AIKA is lightweight in terms of CPU and memory: its average CPU usage did not exceed 26% (peak reached by the constructor), while its average memory consumption did not exceed 847MB (peak reached by the evaluator). Both utilizations were considered tolerable as the computer was far from exhausting its resources. AIKA also proved to be efficient in terms of execution time: the average execution time of the constructor was 2.7 secs (per QID), while for the evaluator it was 21.6 secs. Finally, the GC was only significant for the constructor, where it represented 11% of the execution time. In contrast, it involved less than 3% for the evaluator (meaning that its memory settings were appropriate). This information is shown in Table 2.

Table 2: Resources' utilizations of AIKA components

AIKA component	Avg CPU (%)	Std CPU (%)	Avg MEM (MB)	Std MEM (MB)	Avg Exec. Time (sec)	Std Exec. Time (sec)	Avg MaGC Time (sec)	Std MaGC Time (sec)
Constructor	25.25	1.27	247.80	54.35	2.73	0.66	0.30	0.03
Evaluator	18.87	0.69	846.09	316.82	21.64	2.80	0.71	0.15

AIKA's VGHs (A-VGHs) vs. Manual VGHs (M-VGHs). Sixteen researchers from our department participated in this experiment. Due to their limited availability, we focused on one dataset (i.e., Insurance). This dataset was chosen as its attributes belong to relatively common domains. This allowed us to define an improvement baseline (as the gains in more complex domains would be higher). We provided the participants with a set of leaf terms for each domain. They then defined the ancestor nodes and organized all terms, ending at the root node (also provided). To specify the VGHs, participants used their own knowledge, plus other auxiliary sources (e.g., dictionaries) except WordNet (AIKA's current knowledge base). Finally, the experiment was not time-bounded.

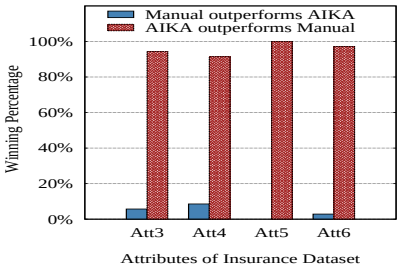


Fig. 9: Winning Percentages

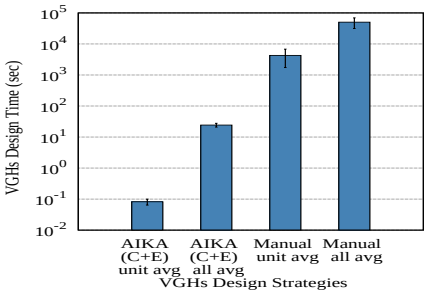


Fig. 10: Efficiency Comparison

To compare the quality of the two VGH sets, we firstly evaluated them using AIKA (with the 35 sets of weights previously discussed) and analyzed their corresponding quality rankings. This analysis showed that the A-VGHs drastically outperformed the M-VGHs, as in more than 95% of the 140 cases, an A-VGH was ranked #1. This is depicted in Fig. 9, which shows the number of wins (i.e., ranks #1) achieved by each VGH type. We also compared their differences in rankings and reward scores. This showed that when an A-VGH was not the best (i.e., did not win), the ranking difference was minimal (only 1 place). On the contrary, M-VGHs always lost by several places (an average of 14). The same behavior was observed in terms of reward scores. Also, in the few cases where M-VGHs won, those VGHs were created by the participants who invested the longest time designing the VGHs (meaning that they were expensive wins).

Next, we assessed the time-savings gained by AIKA. First, the time required by AIKA to create/evaluate (C+E) one VGH was compared against the time reported by the participants. This comparison showed that AIKA offers significant time-savings, as its unitary cost was 99.99% smaller. We also compared the time required to create all VGHs of each type. This also proved AIKA's usefulness, as the time-savings were also significant (an average decrease of 99.95%). These results are depicted in Fig. 10. It is also worth noting that: (i) The manual effort only considers the intrinsic evaluation performed during the construction of the VGHs. If any extrinsic evaluation would be performed, the time-savings would be higher; (ii) AIKA created/evaluated more VGHs (an average of 100) than the participants (16), meaning that the domains were more exhaustively explored.

5 Conclusions And Future Work

This paper presents AIKA, a knowledge-based framework to automatically construct and evaluate VGHs for the anonymization of categorical data. Our experiments proved that AIKA can accurately create and determine which VGH is the most appropriate for a given scenario. AIKA also proved to be lightweight in terms of computational resources. Finally, results showed that AIKA's VGHs are not only better than manual ones, but also AIKA was significantly faster. As future work, we plan to evaluate AIKA with other ontologies, extend it to support phrases, and make it more configurable to release AIKA as a publicly-available tool. Finally, although AIKA has been tested in anonymization, its applicability

can be broader. Thus, we plan to apply it to other areas where concepts are hierarchically ordered and data semantics is the main property to be preserved.

Acknowledgments. This work was supported with the financial support of the Science Foundation Ireland grants 10/CE/I1855 and 13/RC/2094.

References

1. Chicago Homicides. `https://data.cityofchicago.org`
2. Insurance. `https://github.com/ucd-pel/Datasets/tree/master/Insurance`
3. UTD ToolBox. `http://cs.utdallas.edu/dspl/cgi-bin/toolbox/`
4. WS4J library. `https://code.google.com/p/ws4j/`
5. Ayala-Rivera, V., McDonagh, P., Cerqueus, T., Murphy, L.: A Systematic Comparison and Evaluation of k -Anonymization Algorithms for Practitioners. Trans. Data Priv. 7(3), 337–370 (2014)
6. Ayala-Rivera, V., McDonagh, P., Cerqueus, T., Murphy, L.: Ontology-Based Quality Evaluation of Value Generalization Hierarchies for Data Anonymization. In: PSD (2014)
7. Banerjee, S., Pedersen, T.: An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: CICLing. pp. 136–145 (2002)
8. Campan, A., Cooper, N., Truta, T. M.: On-the-fly generalization hierarchies for numerical attributes revisited. In: Secur. Data Manag. pp. 18–32 (2011)
9. D'Aquin, M., Natalya NF.: Where to Publish and Find Ontologies? A Survey of Ontology Libraries. Web semantics (Online) 11, 96–111 (2012)
10. Domingo-Ferrer, J., Sánchez, D., Rufian-Torrell, G.: Anonymization of nominal data based on semantic marginality. Information Sciences 242, 35–48 (2013)
11. Kröll, M., Fukazawa, Y., Ota, J., Strohmaier, M.: Concept Hierarchies of Health-Related Human Goals. In: KSEM. pp. 124–135 (2011)
12. Lee, S., Huh, S.-Y., McNiel, R. D.: Automatic generation of concept hierarchies using WordNet. Expert Syst. Appl. 35(3), 1132–1144 (2008)
13. LeFevre, K., DeWitt, DJ., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: Int. Conf. Manag. Data. pp. 49–60 (2005)
14. Lichman M.: UCI Machine Learning Repository (2013)
15. Martínez, S., Sánchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. Inf. Fusion 13, 304–314 (2012)
16. Meng, L., Huang, R., Gu, J.: A Review of Semantic Similarity Measures in WordNet. Int. J. of Hybrid Information Tech 6(1), 1–12 (2013)
17. Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V., Bragge, J.: The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. In: DESRIST. vol. 24, pp. 83–106 (2006)
18. Portillo-Dominguez, A.O., Wang, M., Magoni, D., Perry, P., Murphy, J.: Load balancing of java applications by forecasting garbage collections. In: ISPDC (2014)
19. Sánchez, D., Batet, M., Martínez, S., Domingo-Ferrer, J.: Semantic variance: An intuitive measure for ontology accuracy evaluation. EAAI 39, 89–99 (2015)
20. Solé-Ribalta, A., Sánchez, D., Batet, M., Serratosa, F.: Towards the estimation of feature-based semantic similarity using multiple ontologies. Knowledge-Based Systems 55, 101–113 (2014)
21. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl.-Based Syst 10(05), 571–588 (2002)
22. Wang, Y., Liu, W., Bell, D.: A Concept Hierarchy Based Ontology Mapping Approach. In: KSEM. pp. 101–113 (2010)