Automated k-Anonymization and l-Diversity for Shared Data Privacy

Anne V.D.M. Kayem[1,2(✉)], C.T. Vester[1], and Christoph Meinel[2]

[1] Department of Computer Science, University of Cape Town, Rondebosch, Cape Town 7701, South Africa
akayem@cs.uct.ac.za
http://infosec.cs.uct.ac.za/
[2] Hasso-Plattner-Institute, Potsdam, Germany
http://hpi.de/meinel/lehrstuhl.html

Abstract. Analyzing data is a cost-intensive process, particularly for organizations lacking the necessary in-house human and computational capital. Data analytics outsourcing offers a cost-effective solution, but data sensitivity and query response time requirements, make data protection a necessary pre-processing step. For performance and privacy reasons, anonymization is preferred over encryption. Yet, manual anonymization is time-intensive and error-prone. Automated anonymization is a better alternative but requires satisfying the conflicting objectives of utility and privacy. In this paper, we present an automated anonymization scheme that extends the standard k-anonymization and l-diversity algorithms to satisfy the dual objectives of data utility and privacy. We use a multi-objective optimization scheme that employs a weighting mechanism, to minimise information loss and maximize privacy. Our results show that automating l-diversity results in an added average information loss of 7 % over automated k-anonymization, but in a diversity of between 9–14 % in comparison to 10–30 % in k-anonymised datasets. The lesson that emerges is that automated l-diversity offers better privacy than k-anonymization and with negligible information loss.

Keywords: Automated data anonymization · Multi-objective optimization · k-anonymity · l-diversity · Data outsourcing

1 Introduction

A common challenge faced by law enforcement agencies in developing world regions is that of analyzing large volumes of crime data [7,27]. Recent statistics from the United Nations (UN) and World Bank (WB) [28] estimate that violent crime cost Guatemala an estimated $2.4 billion or 7.3 % of her Gross Domestic Product (GDP) in 2007, and the Mexican government estimated the costs of violence in 2007 at $9.6 billion, primarily from lost investment, local business and jobs. The UN and WB also estimated that, in 2007, Jamaica and Haiti could have increased their GDP by 5.4 % merely by bringing down their crime levels to that of Costa Rica [28]. In South Africa for instance, it is estimated that

S. Hartmann and H. Ma (Eds.): DEXA 2016, Part I, LNCS 9827, pp. 105–120, 2016.
DOI: 10.1007/978-3-319-44403-1_7

more than a million of the approximately 2 million crimes reported annually, are never resolved [17,31]. Surveys indicate that corruption and police ineffectiveness fuel fears of disclosure and the general belief that most offenses go unresolved [17]. Challenges faced by the law enforcement authorities include limited "in-house" computational processing power which makes handling large volumes of crime data challenging and perhaps more importantly, the lack of data analytics expertise which is essential in identifying relevant data for crime resolution. Outsourcing the data to a third-party Data Analytics Service Provider (DASP) offers a cost effective management solution to the data analytics problem but the sensitivity of the data makes pre-processing to protect the data a necessary step before the data is transferred to the DASP.

Existing solutions based on encrypting the data before it is transferred to the DASP are time-intensive in terms of query response time which is undesirable when performance as well as data protection are a concern [4,9,11,12,18,30,33]. Data protection alternatives such as anonymization, are a better solution from the performance perspective. Manual anonymization is however, a time-consuming and error-prone procedure that can result in inadvertent disclosures of information. A further concern with manual anonymization is the challenge of preventing new releases of anonymized datasets from being adversarially combined with historical data to provoke linking and inferential attacks.

In this paper, we present an automated anonymization scheme that extends the standard k-anonymization and l-diversity algorithms to satisfy the dual objectives of data utility and privacy. The automated scheme employs a multi-objective optimization approach that uses a weighting mechanism to maximize information utility (minimize information loss) and diversity to maximise privacy by circumventing linking and inference attacks. This is handled via a two pronged approach where in the first step we maximize information utility under a modified k-anonymity algorithm in a manner that ensures security against linking attacks. In the second step, we extend the k-anonymity algorithm based on the concepts of l-diversity to provide protection against inference attacks. Our results indicate that l-diverse datasets incur an average information loss of 7 % over k-anonymised datasets, but offer better privacy (protection against linking and inference attacks) with a diversity of between 9–14 % in comparison to 10–30 % in k-anonymised datasets. The lesson that emerges is that in automated anonymization, augmenting k-anonymization with l-diversity offers better privacy and at a negligible cost to utility.

The outline of the paper is as follows. In Sect. 2 we provide an overview of the literature on privacy preserving data publishing. We proceed in Sect. 3 with a specification of our proposed multi-objective scheme to support k-anonymization and l-diversity in automated data anonymization. In Sect. 4, we present results from experiments conducted on a prototype implementation platform [27]. We offer conclusions and suggestions for future work in Sect. 5.

2 Related Work

Privacy preserving data publishing combines efficient protection with availability in data analytics [6,16,19,22,25,32,36]. There are two tenets to privacy

preserving data publishing. The first is to anonymize and then mine the data [2,3,6,16] and the second, to mine and then anonymize the released query results [1–3]. The second approach is better suited to users without the adequate in-house human-capital and computational resources. For this reason, we focus on privacy preserving data publishing schemes where the onus is to anonymize and then share.

Anonymization algorithms can be classified into two main groups namely, syntactic and probabilistic models [10]. Syntactic models have a well defined data output format, such that for small data sets privacy traits can often be confirmed by visually inspecting the data. Privacy violation adversarial models are constructed based on generally available information and generalizations drawn from the syntactic and semantic meaning of the underlying data. k-anonymity [1,6,19], l-diversity [25], and t-closeness [22] algorithms as well as their variants are classified under this category.

On the other hand, probabilistic privacy models employ data perturbations based primarily on noise additions to distort the data [10,34]. Perturbation approaches have been critiqued for being vulnerable to inferential attacks based on adversarial knowledge of the the true underlying distributions of the data [24]. Dwork et al. [15] proposed addressing this caveat with the notion of differential privacy. Differential privacy basically requires that the adversary learns no more from a published data set when one record (or individual) is present in, or removed from, the data set [34]. Attempts have also been made to combine attributes from both syntactic and probabilistic models to form hybrid anonymization approaches. Examples include probabilistic k-anonymity [2], and differential privacy with t-closeness [10]. However, automating these approaches for application on mixed data (categorical and numerical) in ways that minimize information loss and maximize privacy is a challenging problem [16,20].

Since crime data includes a mix of numerical and categorical data, we have opted to focus on syntactic anonymization models, specifically k-Anonymity and l-Diversity. For reasons, centered around high processing costs, we decided against considering the t-closeness scheme. Recall that one of the constraints we mentioned, is the limitation on computational processing power that the organizations face. Work on k-Anonymity was pioneered by Sweeney [29] as an approach to sharing data in plain text without revealing private or sensitive information about individuals. The principle behind k-anonymity is to use the notion of bucketization to create k sets of data (equivalence classes) such that for every tuple there exist at least $k-1$ tuples that have the same quasi-identifier[1] values. Sweeney's work [29] triggered a plethora of schemes such as [13,14,21,23] aimed at performance improvement and circumventing inferential attacks.

Various l-diversity schemes have been proposed to address this drawback by considering that sensitive attributes are the main reason behind disclosures of information used to provoke inferential attacks [8,23,26]. l-diversity requires in addition, that the most frequent sensitive attribute occurences in an equivalence

[1] Quasi-identifiers: Attributes which independently or combined can be used to uniquely identify an individual.

class (EC) should not appear more than $\frac{1}{l}$ times in the EC. So, at least l distinct sensitive values must exist in each EC. As in k-anonymity schemes, efficiently obtaining usable but privacy preserving data sets is provably NP-Hard [35] and so, optimization heuristics have been proposed to improve on the basic l-diversity scheme [13,14,26,35]. We note that l-diversity has the drawback of being dependent on the distribution of sensitive attributes in the data set and so, sensitive attribute values with high probability mass functions (that is some values have a very high frequency and others a very low frequency of occurrence) are prone to provoking high information loss in the anonymized data set. In addition l-diversity only considers the frequency of specific values within independent ECs and not in the dataset as a whole which can result in inadvertent inferential disclosure. t-closeness addresses this caveat but requires a high degree of computational resources. Other issues are centered on the semantics of generalizations and the effect these generalizations have on enabling information disclosures [13,22,25].

In the following section, we propose augmented k-anonymity and l-diversity schemes to support automated data anonymization. The idea is to use the notion of Pareto optimality [5] that has the nice quality of considering that no optimal solution exists for a given problem but rather that the solution space consists of a set of optimal points [5]. This quality, is useful in designing an automated anonymization scheme in that it allows the scheme select the best optimal with respect to data utility and privacy at some given instant and to consider historical data releases. As mentioned before, automated data anonymization is a cost-effective and privacy preserving pre-processing step for data that is outsourced to DASPs. Application examples emerge for law enforcement authorities in developing world countries and organizations lacking the "in-house" computational processing power as well as the data analytics expertise. We now describe our proposed solution in the next section.

3 Multi-Objective Data Anonymization (MOA)

In this section we describe our multi-objective optimization scheme that is geared at supporting automated data anonymization via the k-anonymization and l-diversity algorithms. We begin by providing some basic notation to support our subsequent discussions.

3.1 Information Loss Notation

Let A be the attribute space (columns in a data table) such that $a \in A$ represents a specific attribute (column in the data table) in A and d represents a tuple that contains all the attributes in A.

We denote $T(a)$ as the generalization tree for numerical attributes and $K(a)$ is the generalization tree for categorical attributes. Furthermore, $T(a)_{max}$ and $T(a)_{min}$ denote the upper and lower limits respectively for numerical attribute generalizations while $t_{d,i}(a)_{max}$ and $t_{d,i}(a)_{min}$ represent the upper and lower

limits of the generalization of an attribute a in tuple d during the i^{th} iteration of the anonymization algorithm.

Finally, $K(a)_{total}$ is the total number of leaf nodes generated for $K(a)$ and P is the number of nodes created by $K(a)$. $k(a)_p$ is a sub-tree of $K(a)$ rooted at a node $p \in P$ and $k(a)_{p,total}$ is the number of leaf nodes in $k(a)_p$.

3.2 Information Loss and Severity Weighting

Once the data has been processed and generalized, the next step is to find a suitable balance between information loss and privacy. Minimizing information loss is useful in ensuring data usability while maximizing privacy ensures adequate data protection from adversarial access. In line with our goal of multi-objective optimization, we employ a piece-wise function to handle information loss on both categorical and numerical data.

$$IL_{d,i}(a) = \begin{cases} \frac{k(a)_{p,total} - 1}{P-1} & \text{if categorical} \\ \\ \frac{t_{d,i}(a)_{max} - t_{d,i}(a)_{min}}{T(a)_{max} - T(a)_{min}} & \text{if numerical} \end{cases} \tag{1}$$

where the Information Loss Metric is given by:

$$LM_i(a) = \sum_{d \in D} \sum_{a \in A} IL_{d,i}(a) \tag{2}$$

To minimize information loss, we employ a weighting scheme for the loss metric which enables authorized end users to prioritize specific attributes during anonymisation. By this we mean that the data owner can decide to specify the Quasi-Identifiers (QIDs) that should contain more information without negatively impacting on data privacy. The weighting scheme acts as a sort of utility function that can be adjusted dynamically to allow the data owner decide what levels of privacy to sacrifice in favor of query result accuracy without negatively impacting on the overall privacy of the data. The weighted information loss metric ($IL_{weight,i}$) at the i^{th} iteration of the algorithm is computed as follows.

$$IL_{weight,i} = \sum_{d \in D} \sum_{a \in A} w_a \times IL_{d,i}(a) \tag{3}$$

where w_a is the weight assigned to attribute $a \in A$ by the data owner. Finally, to facilitate automated anonymization we use a sensitive attribute severity weighting $S(c)$ where $c \in SA$. SA is the list of sensitive attributes and $S(\cdot)$ maps the sensitive attribute category to its weight.

Example 1. In Table 1, SA denotes the list of offences (sensitive attribute) and $S(\cdot)$ maps the crime category to its weight, which in this case is simply the guideline sentence duration (in time - months, years...) for a given crime. So, $S(\text{Theft}) = 5$ indicates a sentence of 5 years. We note that following this scale, the risk of privacy loss for a tuple containing *"Robbery"* is higher than for a tuple with *"Disorderly Conduct"*.

Table 1. Crime severity weightings

Crime	Severity
Embezzlement	3
Disorderly conduct	3
Theft	5
Drunken driving	5
Robbery	7

We now describe our automated anonymization schemes, namely CG-Kanon and CG-Diverse that are extensions of the k-anonymization and l-diversity algorithms respectively.

3.3 CG-Kanon Scheme

Our proposed CG-Kanon scheme uses the severity weighting and bucketization, to hide tuples with highly sensitive values in larger ECs while tuples of lower sensitivity are classified in smaller ECs. For instance, a tuple concerning a "Robbery" should be classified in a 20-anonymity EC while "theft" could be placed in a lower level EC say, 5-anonymity. This idea of hiding more sensitive values in larger ECs does not affect the absolute level of k-anonymity for different sensitive attribute categories. It is instead a relative statement regarding the level of k-anonymity required for different sensitive attributes in the anonymized dataset. The severity weighting is converted to a severity penalty which is used by the CG-Kanon scheme. To do this, we compute an absolute required minimum level of k-anonymity ($k_{\min}$) for the dataset and use $k_{\min}$ to guarantee a global minimum level of k-anonymity that all ECs must adhere to in the dataset. We compute $k_{\min}$ as follows:

$$k_{min} = \max\left(k_{cons}, \min\left(S_D(\cdot)\right)\right) \tag{4}$$

where k_{cons} is a fixed minimum level of k and $S_D(\cdot)$ is the set of all severities for the dataset D. The definition of k_{min} shows that the global minimum level of k-anonymity is fixed at k_{cons} or at the lowest level of attribute sensitivity in the dataset when $\min\left(S_D(\cdot)\right) > k_{cons}$. If $k_{cons} = 5$ and $\min\left(S_D(\cdot)\right) = 3$ then $k_{min} = 5$. However if $\min\left(S_D(\cdot)\right) = 7$ then $k_{min} = 7$ instead. The CG-Kanon scheme uses $k_{\min}$ as the k-anonymity baseline when deciding on appropriate ECs for tuples based on sensitivity.

Once $k_{\min}$ has been computed, we compute the severity penalty for each classification since the CG-Kanon scheme requires this information to optimize the information loss and privacy cost-benefit trade-off. The severity penalty determines the level of loss of privacy for a single tuple $d \in D(\cdot)$ and is computed as follows.

$$SP_{d,i} = \frac{S_d(c)}{|e_{d,i}|} \tag{5}$$

where $D(\cdot)$ is the dataset, $S_d(c)$ is the severity weight of sensitive attribute $c \in d$, and E is the set of ECs such that $|e_{d,i}|$ is the size of the EC that a tuple d is classified in during the i^{th} iteration of the CG-Kanon scheme.

Example 2. From the severity penalty computation, highly sensitive attributes in small ECs result in high penalties and vice versa. So, if a "murder" report with a severity weighting of 25 were located in a 5-anonymity EC, a penalty of $\frac{25}{5} = 5$ is generated. An incident of "theft" with a severity weighting of 5 generates a severity penalty of 1, indicating that this information is comparatively less sensitive. The CG-Kanon scheme uses the severity penalty as a criterion besides, information utility, to determine tuple placement in ECs to minimize the overall sensitive information exposure risk.

Finally, the CG-Kanon scheme must compute the aggregate severity penalty, $SP_{tot,i}$, for the entire dataset, to determine whether the obtained anonymized dataset satisfies at least the threshold goals of privacy and utility. $SP_{tot,i}$ is computed as follows:

$$SP_{tot,i} = \sum_{d \in D} SP_{d,i} \tag{6}$$

and expresses the total severity penalty for the dataset as the summation of the severity penalties of the individual tuples. The $SP_{tot,i}$ is then feed into a fitness function to decide whether each tuple in D satisfies both objectives. We express the fitness function as follows:

$$FF_i^{CG-Kanon} = \frac{1}{\max\left(SP_{tot,i}, LM_{CG,i}\right)} \tag{7}$$

So, the result for $FF_i^{CG-Kanon}$ at iteration i is the inverse of the maximum of $SP_{tot,i}$ and $LM_{CG,i}$. Recall that a high $SP_{tot,i}$ indicates a strong risk of privacy exposure, while a high $LM_{CG,i}$ indicates a high level of information loss. Therefore, it is desirable that the fitness function generates results that iteratively converge towards a high value for $FF_i^{CG-Kanon}$, expressed by low values of $SP_{tot,i}$ and $LM_{CG,i}$ respectively.

The main drawback here is that, depending on tuple distribution, the diversity of the sensitive attributes in large ECs can be quite low and this negatively impacts on privacy. As well, a large proportion of tuples are suppressed to satisfy the minimum level of k-anonymity which results in high information loss. We addressed this by limiting the size of ECs to a pre-defined threshold size and as we discuss in Sect. 4, found that this reduces the number of suppressions to satisfy $k_{\min}$-anonymity. We still have the caveat of inferential attacks and so augment our CG-Kanon scheme with the CG-Diverse scheme (l-diversity algorithm inspired) to help circumvent these attacks.

3.4 CG-Diverse Scheme

Instead of using $SP_{tot,i}$ to classify tuples into ECs, the CG-Diverse scheme computes the average severity, AS_D, for D as well as the EC average severity weighting AS_e. The AS_D is computed for D and is used to start the anonymization

process to ensure that the target level of l-diversity in D is such that $l = AS_D$. We compute AS_D as follows:

$$AS_D = \frac{\sum_{d \in D} S_d(c)}{|D|} \tag{8}$$

A high AS_D implies a higher level of diversity in the entire dataset. As a stopping criterion for deciding when an acceptable level of k_{min} and AS_D has been satisfied by all the ECs, we bound the l-diversity range with the severity weighting scale and use AS_D to compute the fitness of the dataset with respect to privacy and utility. We employ the following modified fitness function, expressed as follows:

$$FF_i^{CG-Diverse} = \frac{1}{\max\left(AS_{D,i}, LM_{CG,i}\right)} \tag{9}$$

However, as mentioned before suppressing the ECs that fail to meet the required levels of AS_D and k_{min} would result in a high level of information loss. Therefore, we alleviate this problem by identifying ECs with a lower average severity (but adequate relative diversity) to avoid high suppression rates. This is achieved by assessing the privacy of individual ECs that do not meet the global AS_D-diversity requirement. To this end the EC average severity weighting AS_e is computed as follows:

$$AS_e = \frac{\sum_{d \in D} S_d(c)}{|e|} \tag{10}$$

The AS_e of an EC is compared to the relative diversity l_e, and if $AS_e > l_e$ the tuples in the EC are generalized to the highest possible level to avoid suppression. Alternatively, when the diversity is higher than AS_e no changes are made. We note that this procedure is computationally inexpensive since it simply requires comparing AS_e with the actual observed diversity of the EC.

Example 3. Table 2 shows the average severity measures calculated for a given sample dataset. The $AS_e = 5$ is calculated as follows: $\frac{5+3+7+5+5}{5}$ using the crime severity weightings given in Table 1. By considering Table 1, and Eqs. (8) and (10), the l-diversity range can be restricted to between 3–25, depending on the underlying dataset. Yet requiring ECs to satisfy the global level of AS_D-diversity might be too restrictive. We alleviate this issue by moving tuples between ECs to minimise the information loss due to suppression. For instance, in Table 2 we observe that in the 5-anonymity EC, "Robbery" has a severity of 7 which implies an inference risk. CG-diverse handles such cases by using the AS_e to move the tuple to the more appropriate 7-anonymity EC as highlighted in Table 2.

We are now ready to discuss our experimental platform, results and analysis.

4 Results and Analysis

We demonstrate the feasibility of our proposed automated data anonymization scheme with results from experiments conducted on a prototype crime data collection application [27]. A host server with an Ubuntu server 12.04 operating

Table 2. Average severity versus diversity

Age	Crime	Diversity (Equivalence Class)	AS_D (Dataset)	AS_e (Equivalence Class)
18 - 22	Theft	4	11	5.0
18 - 22	Embezzlement	4	11	5.0
18 - 22	Robbery	4	11	5.0
18 - 22	Drunken Driving	4	11	5.0
18 - 22	Theft	4	11	5.0
18 - 87	Rape	8	11	7.0
18 - 87	Vandalism	8	11	7.0
18 - 87	Robbery	8	11	7.0
18 - 87	Assault	8	11	7.0
18 - 87	Murder	8	11	7.0

system running on a 64 bit machine with 8 GB RAM and a processor speed of 3.2 GHz (Intel Xeon E3-1230 Quad Core) was used. The algorithms were implemented in Java 1.7.0_65 while Python 2.7.3 was used to run the web server. A PostgreSQL 9.1 database management system and a Postfix email server were used to store the dataset, both plain and anonymized. Our dataset consisted of 10000 records because this is a reasonable bound for daily average crime report rates per police station [17]. The attributes considered included "Age", "Suburb", "Crime" and "Reporter". Sensitive attributes such as "Names" and "Date of Birth" were removed during pre-processing. Quasi-identifiers which more closely match the k-anonymity requirement for CG-Kanon were generated before the anonymization process. This was done by generalizing attributes to the highest node in the generalization hierarchy (tree) for ECs that do not meet the k-anonymity requirement. We qualitatively assessed the anonymized data produced by the CG-Kanon and the CG-Diverse algorithms, by considering aspects such as information loss, classification accuracy and the impact of the weighting scheme on linking and inference attacks. Throughout the discussion of the results we refer to an anonymization based on the weightings of the quasi-identifiers (QIDs) used during the anonymization. This will be denoted as $A_{w_{Age}} : S_{w_{Suburb}} : R_{w_{Reporter}}$. For example where equal weights were assigned to the QIDs this will be denoted as an $A1 : S1 : R1$ anonymization, similarly where we use $A10 : S5 : R1$ weights of 10, 5, and 1 were used for the *Age*, *Suburb*, *Reporter* attributes respectively (Fig. 1). $k_{constant}$ was set to 5 for all results on CG-Kanon anonymization. Our minimum crime severity level for the data was set to 3 and in this case, $k_{min} = 5$. For CG-Diverse, we set our lowest diversity level to 3 for all anonymization runs as a standard minimum privacy level. Since on average, the lower severity crimes were located in such ECs, this was acceptable. All algorithms were allowed to run for 30 min after which the algorithm was stopped. Pre-experiment sampling revealed that running for shorter periods, say 15 min resulted in high severity penalties and information loss for larger ECs, with only between 3–6 % of tuples meeting the minimum anonymity level. Running for much longer resulted in better success rates, but

at the price of time. Once stopped the anonymized data was checked for compliance with the desired level of privacy. Tuples not satisfying the privacy criteria on termination were processed further according to the respective CG-Kanon and CG-Diverse algorithms (Fig. 1). Figure 2 shows the CG-Kanon algorithm classifying data using ECs only with no severity weighting support. We note that the crimes are clustered around smaller sized ECs which is good for protection against inference attacks, but bad for information loss. When the severity penalty is applied, we note as shown in Fig. 3 that more severe crimes are classified in larger ECs but this has the caveat of introducing inferential disclosure. For instance, from Fig. 3 one can see directly that more severe crime has a higher frequency with *"Murder"* being as high as 31 %. We address this with the CG-Diverse scheme. As shown in Figs. 4 and 5, based on the $A1 : S1 : R1$ weighting and an average severity level of 11, the global diversity and average severity of each EC is evaluated before suppressing the QIDs. When compared to Figs. 2 and 3, we note that the average diversity in CG-Kanon varies between 10 % and 30 % while that of CG-Diverse is much lower at 9 % to 14 % and consequently lowers inferential risk.The desired lower frequency (i.e. higher diversity) for more

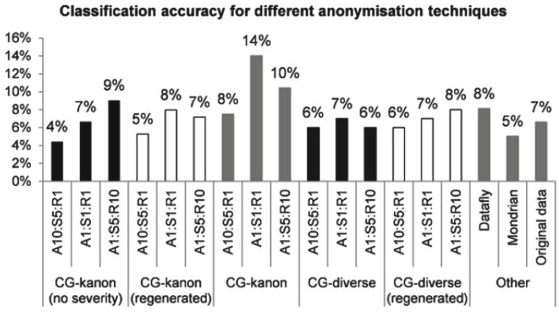

Fig. 1. Classification accuracy of CG-Kanon and CG-Diverse

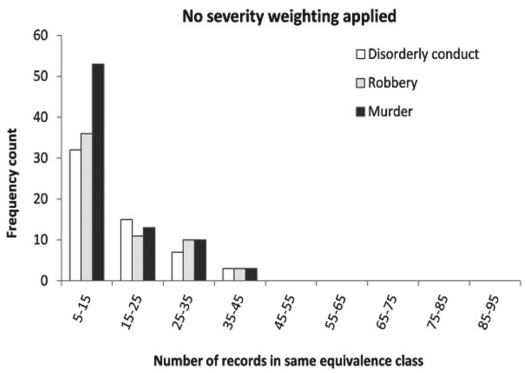

Fig. 2. Severity impact on dataset (no severity weighting)

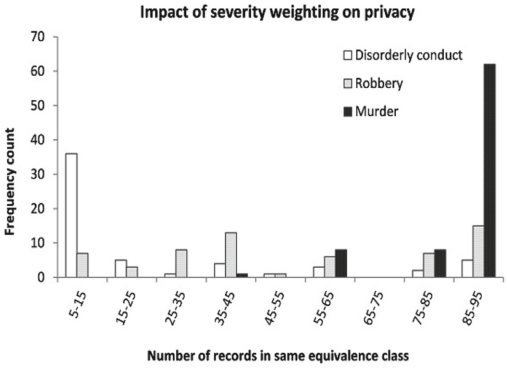

Fig. 3. Impact of severity weighting on privacy

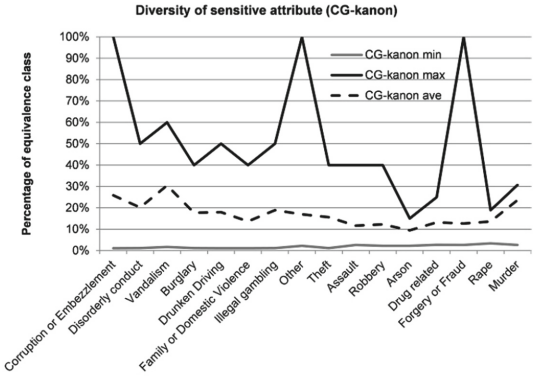

Fig. 4. Sensitive attributes frequency for CG-Kanon using A1:S1:R1

severe crimes is evident in CG-diverse whereas in CG-Kanon there is no such correlation. More severe crimes (Rape and Murder) in this case actually have lower average diversity and consequently less risk of inferential exposure. In addition we see the deviation from the mean frequency for more severe crimes is lower as severity increases. So not only does the average diversity increase as crime severity increases but the variance decreases as well. This gives us more certainty that more severe crimes will be less vulnerable to inference attacks. Finally, we note that l-diversity guarantees at least k-anonymity where $k = l$. The lowest diversity of 3 may appear weak from the privacy perspective when compared to the global diversity of 11 but it is unlikely, practically speaking, that severe crime (sensitive data) will be included in such lower diversity ECs. For instance, if we revisit our earlier results for CG-Kanon where the most serious crime ("Murder") was in an EC of size 90 and still only achieved a 3-diversity. Figures 6 and 7 show the aggregated information losses for different weighting schemes after termination of the algorithm. We selected three weighting schemes to mon-

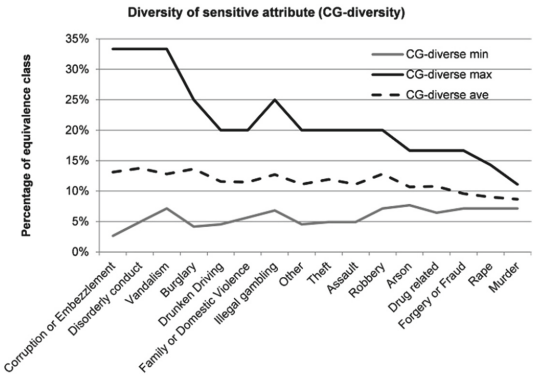

Fig. 5. Sensitive attributes frequency for CG-Diverse using A1:S1:R1

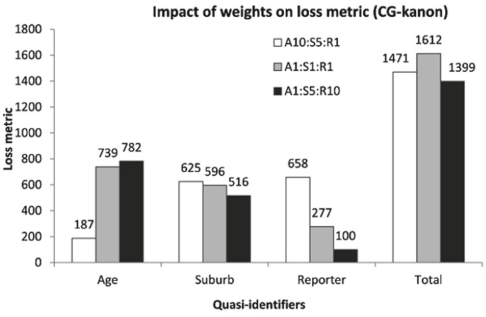

Fig. 6. Information loss for CG-Kanon

itor how the algorithms perform when attributes with varying granularity are weighted differently. For instance the $A10 : S5 : R1$ scheme overweights the *Age* attribute which is highly granular and under weighs the *Reporter* attribute, while $A1 : S5 : R10$ test the opposite scenario and $A1 : S1 : R1$ is equivalent to having no weighting scheme. The marginal increase in information loss for CG-diverse relative to CG-Kanon seems quite acceptable given the improved privacy provided by CG-Diverse. For our results the information loss across the three weighting schemes was on average 7 % higher for CG-diverse. However, this reduced data utility is acceptable given our desire for better anonymized data privacy. One further insight relates to the number of parameters that are used for the fitness function in selecting QIDs. We see from Figs. 8 and 9 that information loss for CG-diverse is a much lower proportion of its starting value than for CG-Kanon. This is attributed to the fact that CG-Kanon searches for solutions that minimize both the information loss and the severity penalty, in addition to satisfying k-anonymity. While CG-diverse only minimizes information loss and endeavours to meet the diversity requirement. The additional parameter (severity penalty) for CG-Kanon increases the search space and reduces the efficiency

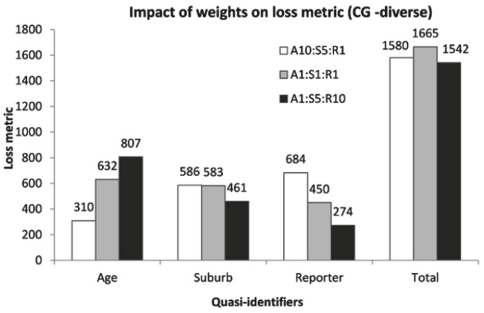

Fig. 7. Information loss for CG-Diverse

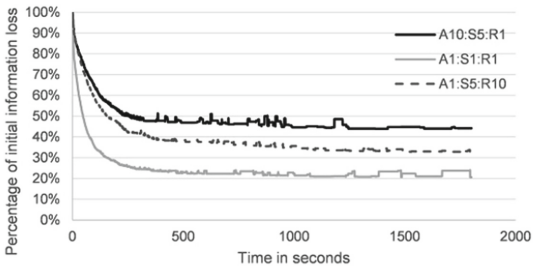

Fig. 8. Information loss reduction versus time (CG-Kanon)

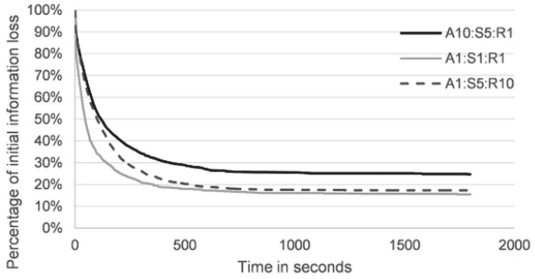

Fig. 9. Information loss reduction versus time (CG-Diverse)

of the algorithm. For instance, at termination the reduction in the initial information loss for $A10 : S5 : R1$ in CG-diverse (Fig. 9) was 74 % compared to 55 % for CG-Kanon (Fig. 8).

5 Conclusions

We presented two algorithms namely, CG-Kanon and CG-diverse that augment the standard k-anonymity and l-diverse algorithms to facilitate automatic classification and anonymization of data. In particular, we considered crime data

because it contains a large volume of sensitive data and is vulnerable to linking and inferential attacks. To match privacy with utility, we used a random sampling approach without replacement so, historical released reports were excluded from being selected in subsequent releases. The sampling approach also offers the advantage of reduced computational complexity and therefore runtime for our algorithms which is a plus for use in computationally constrained environments. To reduce information loss, we also used a fitness function to improve classification accuracy, and privacy. Our results demonstrate that CG-diverse incurs an average information loss of 7 % over CG-Kanon, but with a diversity of between 9–14 % in comparison to 10–30 % CG-Kanon. So, we can conclude that, since CG-Diverse offers anonymity levels that are at least equal to CG-Kanon's, the percentage of information loss incurred does not significantly affect query response accuracy and in addition, provides stronger privacy guarantees than CG-Kanon.

Possible avenues for future work include evaluating CG-Kanon and CG-Diverse on de facto anonymization benchmarks such as the Adult's census dataset from the UC Irvine machine learning repository. Additionally, evaluations of robustness to other known attacks against k-anonymization and l-diversity will be useful for practical purposes. Finally, we should also consider parametrizing the t-closeness model for better performance under constrained conditions as an interesting candidate for overcoming the drawbacks of CG-Kanon and CG-Diverse.

Acknowledgements. The authors gratefully acknowledge funding for this research provided by the National Research Foundation (NRF) of South Africa, and the Hasso-Plattner-Institute (HPI). In addition, the authors are grateful for the anonymous reviews.

References

1. Aggarwal, C.C.: On k-anonymity and the curse of dimensionality. In: Proceedings of the 31st International Conference on Very Large Data Bases, VLDB 2005, pp. 901–909. VLDB Endowment (2005)
2. Aggarwal, C.C.: On unifying privacy and uncertain data models. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE 2008, pp. 386–395. IEEE Computer Society, Washington, DC (2008)
3. Aggarwal, C.C., Yu, P.S.: Privacy-Preserving Data Mining: Models and Algorithms, 1st edn. Springer, New York (2008)
4. Arasu, A., Eguro, K., Kaushik, R., Ramamurthy, R.: Querying encrypted data. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD 2014, pp. 1259–1261. ACM, New York (2014)
5. Aytug, H., Koehler, G.J.: New stopping criterion for genetic algorithms. Eur. J. Oper. Res. **126**(3), 662–674 (2000)
6. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: 21st International Conference on Data Engineering (ICDE 2005), pp. 217–228, April 2005

7. Burke, M., Kayem, A.: K-anonymity for privacy preserving crime data publishing in resource constrained environments. In: 28th International Conference on Advanced Information Networking and Applications Workshops, AINA 2014 Workshops, Victoria, BC, Canada, 13–16 May 2014, pp. 833–840 (2014)
8. Ciriani, V., Vimercati, S.D.C., Foresti, S., Samarati, P.: k-anonymous data mining: a survey. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining: Models and Algorithms, pp. 105–136. Springer, Boston (2008)
9. Ciriani, V., De Capitani Di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Combining fragmentation and encryption to protect privacy in data storage. ACM Trans. Inf. Syst. Secur. **13**(3), 22:1–22:33 (2010)
10. Clifton, C., Tassa, T.: On syntactic anonymity and differential privacy. Trans. Data Priv. **6**(2), 161–183 (2013)
11. De Capitani Di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Encryption policies for regulating access to outsourced data. ACM Trans. Database Syst. **35**(2), 12:1–12:46 (2010)
12. De Capitani Di Vimercati S., Foresti, S., Paraboschi, S., Pelosi, G., Samarati, P.: Shuffle index: efficient and private access to outsourced data. ACM Trans. Storage **11**(4), 19:1–19:55 (2015)
13. Dewri, R., Ray, I., Ray, I., Whitley, D.: Exploring privacy versus data quality trade-offs in anonymization techniques using multi-objective optimization. J. Comput. Secur. **19**(5), 935–974 (2011)
14. Dewri, R., Whitley, D., Ray, I., Ray, I.: A multi-objective approach to data sharing with privacy constraints and preference based objectives. In: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, GECCO 2009, pp. 1499–1506. ACM, New York (2009)
15. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
16. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: Fast data anonymization with low information loss. In: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB 2007, pp. 758–769. VLDB Endowment (2007)
17. Gould, C., Burger, J., Newham, G.: The saps crime statistics: what they tell us and what they don't. SA Crime Quaterly (2012). https://www.issafrica.org/uploads/1crimestats.pdf
18. Hang, I., Kerschbaum, F., Damiani, E.: ENKI: access control for encrypted query processing. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD 2015, pp. 183–196. ACM, New York (2015)
19. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 279–288. ACM, New York (2002)
20. Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, pp. 193–204. ACM, New York (2011)
21. Last, M., Tassa, T., Zhmudyak, A., Shmueli, E.: Improving accuracy of classification models induced from anonymized datasets. Inf. Sci. **256**, 138–161 (2014). Business Intelligence in Risk Management
22. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, pp. 106–115, April 2007
23. Lin, J.L., Wei, M.C.: Genetic algorithm-based clustering approach for k-anonymization. Expert Syst. Appl. **36**(6), 9784–9792 (2009)

24. Liu, K., Giannella, C., Kargupta, H.: A survey of attack techniques on privacy-preserving data perturbation methods. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining: Models and Algorithms, pp. 359–381. Springer, Boston (2008)
25. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data **1**(1), 1–52 (2007)
26. Nergiz, M.E., Tamersoy, A., Saygin, Y.: Instant anonymization. ACM Trans. Database Syst. **36**(1), 2:1–2:33 (2011)
27. Sakpere, A.B., Kayem, A., Ndlovu, T.: A usable and secure crime reporting system for technology resource constrained context. In: 29th IEEE International Conference on Advanced Information Networking and Applications Workshops, AINA 2015 Workshops, Gwangju, South Korea, 24–27 March 2015, pp. 424–429 (2015)
28. Seckan, B.: Violent crime in the developing world: research roundup. Journalist's Resource: Research on today's New topics (2012). http://journalistsresource.org/studies/international/development/crime-violence-developing-world-research-roundup
29. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **10**(5), 557–570 (2002)
30. Wang, F., Kohler, M., Schaad, A.: Initial encryption of large searchable data sets using hadoop. In: Proceedings of the 20th ACM Symposium on Access Control Models and Technologies, SACMAT 2015, pp. 165–168. ACM, New York (2015)
31. Website: South Africa's police: something very rotten. In: The Economist: Middle East and Africa (2012). http://www.economist.com/node/21557385
32. Wicker, S.B.: The loss of location privacy in the cellular age. Commun. ACM **55**(8), 60–68 (2012)
33. Wong, W.K., Kao, B., Cheung, D.W.L., Li, R., Yiu, S.M.: Secure query processing with data interoperability in a cloud database environment. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD 2014, pp. 1395–1406. ACM, New York (2014)
34. Xiao, Q., Reiter, M.K., Zhang, Y.: Mitigating storage side channels using statistical privacy mechanisms. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS 2015, pp. 1582–1594. ACM, New York (2015)
35. Xiao, X., Yi, K., Tao, Y.: The hardness and approximation algorithms for l-diversity. In: Proceedings of the 13th International Conference on Extending Database Technology, EDBT 2010, pp. 135–146. ACM, New York (2010)
36. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.C.: Utility-based anonymization using local recoding. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, pp. 785–790. ACM, New York (2006)