User-defined Privacy Preferences for k-Anonymization in Electronic Crime Reporting Systems for Developing Nations

Aderonke Busayo Sakpere
Department of Computer Science, University of Cape Town, Cape Town, South Africa

Keywords: Data Anonymity, Streaming Data, Crime Reporting, User-defined Privacy.

Abstract: Existing approaches that protect data from honest-but-curious data mining service providers include k-anonymity technique, which is considered a better alternative to previously proposed techniques. However k-anonymity technique adopts a generic paradigm approach to privacy enforcement in its model. Owing to the fact that real-life users have different privacy requirements, there is need to address this generic paradigm approach in K-anonymity in order to improve its efficiency. Our proposed approach integrates the concept of a three tier-privacy level (low, medium and high) into k-anonymity to achieve anonymization. This helps us to identify individual users' best choice and how users' privacy preference can be incorporated into the K-anonymity model, as opposed to the generic approach currently adopted. Our preliminary survey presents facts that help to understand factors that influence the choice of users' privacy preference during crime reporting. Results also show that the following factors affect people's privacy choice: Age Group, Personality, Community Need and Cultural Background (Adaptive).

1 INTRODUCTION

Increasing rates of crime occurrences in developing nations have raised government concerns for safety. A recent report shows that about 3.3 million crimes occur yearly in South Africa (SAIRR, 2013). Thus, a lot of crime data/report are archived by the security agencies, which have potential to generate knowledge-driven decision support in tackling crime issues if efficiently analysed. Manual data analysis is no longer a viable approach to crime prediction and prevention because the large volume of data make manual data analytics a time consuming process. In areas where on-site data analytics expertise is limited, outsourcing the data to a third-party data analytics service provider is a good solution. Due to lack of analytical expertise within the law enforcement agencies in developing countries, these crime reports are often not analysed or mined in order to predict and prevent future crime occurrences. It therefore makes sense to involve third parties that have the expertise.

Seeing that crime data is privacy sensitive, it makes sense to ensure crime data is protected from an "honest-but-curious" data analytic/miner. The use of techniques common in cryptography, access control and authentication for protecting crime data are not sufficient to prevent third parties from identifying subjects in the dataset. This is because they do not guard against inferences (Sweeney, 2002b). Other data protection techniques such as swapping, perturbation and additive noise have been studied for protecting outsourced data from unauthorized access but compromises data integrity (Sakpere and Kayem, 2014). A recent privacy preserving techniques is differential privacy. However, some theoretical research on this technique has shown infeasible results (Dwork, 2006). Another disadvantage of differential privacy is that it makes use of additive noise. According to Guo and Zhang, 2013, too much noise makes analysis of anonymized data more difficult to analyse. K-anonymity techniques are a better alternative to protecting outsourced data because it does not compromise the integrity (truthfulness) of anonymized data (Bayardo and Agrawal, 2002). Furthermore, results from various theoretical and practical research and application of k-anonymity has shown its suitability for privacy preservation ((Guo and Zhang, 2013; Samarati, 2001; Sweeney, 2002b; Sweeney, 2002a).

K-anonymity preserves privacy by ensuring that each record corresponds to at least k-1 other records with respect to their Quasi-Identifier, where k is a pre-assigned integer variable and $k > 1$ (Samarati, 2001; Sweeney, 2002b; Sweeney, 2001). Quasi-Identifier is mainly a combination of one or more

Sakpere A..
User-defined Privacy Preferences for k-Anonymization in Electronic Crime Reporting Systems for Developing Nations.
Copyright © 2015 SCITEPRESS (Science and Technology Publications, Lda.)

non-explicit identifiers. These non-explicit attributes when combined together can potentially identify individuals. Examples of Quasi-Identifiers are Age and Sex. K-anonymity achieves this by using generalization and suppression (Samarati, 2001; Sweeney, 2002b). Generalization replaces a specific value with a more general but semantically consistent value (Sweeney, 2002a). Suppression involves withholding a value completely (Sweeney, 2002a). Therefore, if an attacker wants to identify a man in a released table and the only information he has is his birth date and gender. K-anonymity ensures there are k men in the table with the same birth date and gender. K is a pre-assigned integer that is greater than one. Tables 1 and 2 illustrate how k-anonymity works.

Table 1: Crime Victims' data in an explicit form.

Name	Age	Gender
Rose	25	Female
Mary	29	Female
Scoth	32	Male
Smith	36	Male

Table 2: Crime Victims' data in an anonymous form.

Name	Age	Gender
****	20-29	Female
****	20-29	Female
****	30-39	Male
****	30-39	Male

Table 1 shows data that needs to be anonymised. Table 2 is an anonymised version of table 1 using k-anonymity, where $k = 2$ and QI = (Age, Sex). From table 2, each sequence of values in QI has at least two occurrences. Hence, the probability of re-identification occurrence is $1/k$.

2 OUTLINE OF OBJECTIVES

K-anonymity model has been identified as a more promising alternative in data privacy preservation, as opposed to previously proposed techniques that have shortcomings in areas such as additive noise and inference attack to mention a few. However, k-anonymity still requires refinement in certain aspects of its model, such as incorporating user-defined factors into the model for privacy preservation, as it ordinarily uses a generic paradigm for this purpose. Our key research questions are:

1. What are the major factors that determine the privacy level preference of a person who has been a victim of crime or a potential victim?
2. How can user defined privacy preference be integrated into k-anonymity, to improve its efficiency in privacy preservation?

3 RESEARCH PROBLEM

K-anonymity uses the same privacy level (i.e. k-value) for all individuals in the data set. The use of the same privacy for all users is unrealistic in real-life because individuals tend to have varying privacy protection requirements (Xiao and Tao, 2006; Gedik and Liu, 2008). Furthermore, the use of the same privacy preference for all users mean individual's privacy need is misrepresented. As a result, some users may be over-protected, while some others may be under-protected. This implies that over-protection could lead to high loss of information and under-protection could lead to inadequate protection (Xiao and Tao, 2006). Information loss is used to quantify the amount of information that is lost due to k-anonymization (Kabir and Bertino, 2011). The consequence of a high information loss is that it debases the utility of the released anonymized database for data mining or analysis (Byun and Li, 2006).

To illustrate this problem, let's assume a user named Mary prefer her details to be known when her details is released to a third party for data mining purposes. On the other hand, Smith might prefer that his details are well protected before they are released to third party for analysis or mining purposes. In addition, there are individuals who are indifferent about their privacy. As a result, it makes sense to integrate individuals privacy preference into k-anonymity.

4 STATE OF THE ART

The need for data protection, especially when needed for research and data mining purposes has led to the development of several privacy enforcing algorithms that are based on techniques such as swapping, substitution, perturbation and additive noise (Sweeney, 2001). A major problem that the use of these techniques face, is the difficulty in relating them to the legal and societal norms of privacy (Jiang and Clifton, 2006). Furthermore, these techniques can produce "untruthful data" (Bayardo and Agrawal, 2002). As a result, Sweeney (Sweeney, 2002b) came up with k-anonymity to solve these deficiencies. Recently, a new technique named differential privacy has emerged to ensure privacy. It achieves this by the use of a randomized mathematical function. However, the use of rigorous mathematical computation involved

in differential privacy makes it computationally intensive (Dwork, 2006).

One of the pioneer work in personalized privacy is by Aggrawal & Yu (Aggarwal and Philip, 2008). They achieved personalized privacy through the use of k-anonymity by allowing a user to select an integer, i, (where $1 \leqslant i \leqslant n$) to indicate his/her privacy preference. This implies that in an anonymised table, T, the user must be included in a QI-group with at least size i. A drawback of this is that it might be difficult for users to set a realistic k-value in real-life especially in Crime Reporting System where users might be under duress or shock as a result of the crime. Also, setting a realistic k-value implies that users must understand the principle of k-anonymity.

An equally novel approach in achieving personalized anonymisation using the concept of k-anonymity is the work of Xiao and Tao, 2006. In their work, an individual specifies the degree of privacy protection for his/her sensitive values. Their solution assumes that each sensitive attribute has a classification tree and each record owner specifies a guarding node in the tree. Guarding nodes depend on user's personal privacy preferences and indicates how users want their sensitive values to be represented. A major drawback of their approach is that a guarding node requires that a hierarchy-tree be defined on sensitive attribute. However, hierarchical trees are difficult to define on numerical and transactional data. Another drawback is that in real-life, it is unclear how individual record owners would set their guarding node (Aggarwal and Philip, 2008).

Gedik and Liu, 2008 achieved personalized k-anonymity by allowing users to specify their preferred k-anonymity value. A setback of this is that users may need to understand the concept of k-anonymity in order to be able to choose an appropriate k-value which may not be practical in real-life.

Another research targeted towards including users privacy preference in k-anonymity is the work of Kabir and Bertino, 2011. In their approach, they only considered the privacy level of individuals who do not care about the disclosure of their details. Their work did not encompass the personal privacy preference of individuals who care about their privacy.

We therefore note that the issue of incorporating users preference to cope with anonymization of data in a manner that is usable in real-life is yet to be studied. This study is necessary in order to generate reliable anonymized reported crime data for third party service providers.

5 METHODOLOGY

The user study approach was used in order to determine factor(s) that influence people's privacy during crime report in real-life. We conducted this preliminary survey in the University of Cape Town, South Africa. Twenty-four participants were recruited to source user experiences with reporting crimes. The participants consisted of twenty users who had been affected by crime before and only four users whom had never been personally affected by crime. Questionnaires and face-to-face interviews were used to gather user's privacy preference during crime reporting. The questionnaire was designed to confirm the validity of the claims in the research of Xiao and Tao, 2006, and Gedik and Liu, 2008 that users have different privacy levels. In addition we also designed the questionnaire to confirm the claims of Chuang and Kuo, 2011 that users find it easier to determine their privacy level using a three-tier privacy level. Additionally the questionnaire aimed to gather other factors such as gender, age and crime level that influences peoples' privacy during crime reporting.

A three-tier privacy level preference consists of low, neutral (medium) and high. It is conceived that the willingness of an individual to share information is inversely related to his/her privacy level preference. Our choice of three-tier is based on the research of Chuang and Kuo, 2011 that believe users can only recognize their privacy requirements between three levels. A high privacy level indicates an extreme privacy consciousness, whereas a low privacy level depicts a lower privacy consciousness. Therefore, neutral privacy level is an intermediate.

5.1 Survey Analysis

Figure 1 and Table 3 illustrate both visual and quantitative contents of data collected. Such summary is necessary to obtain preliminary information about the relationship among the variables collected. All the collected survey data comprised of 24 subjects and eight categorical variables: *Sex*, *Age group*, *Present education level/Occupation*, *Highest education qualification* (HEQ), *Victim of crime*, *Crime experienced*, *Preferred privacy level* (PPL), and *Reason for choice of privacy* (RCP). All the subjects interviewed are postgraduate students whose *Occupation* and *HEQ* are exactly related. For example, a student enrolled for PhD has a Masters degree as his/her *HEQ*. Without loss of generality, *HEQ* will consequently be deleted from the analysis data. Table 3 and Figure 1 provide summaries of the different categories of each of the variables left.

5.2 Survey Model

The response of interest in the survey is the attribute Preferred Privacy Level (PPL) variable. To model the PPL, we use multivariate logistics regression model (Skrondal, 2003) because we are interested in exploring the relationships among the variables (Skrondal, 2003) as stated in our research question. That is, we want to know variables that influence peoples' privacy choice.

Consider a set of k explanatory variables; $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$, such that, $X_j = \{x_{1j}, x_{2j}, \ldots, x_{nj}\}$. Also, assume that each set of observations, $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}\}$, corresponds to measurements collected from a specific subject i. X represents our dependent variables such as age, gender. These measurements could be any combination of continuous, discrete or categorical variables. Further assume that for each $\mathbf{x}_i$, a corresponding binary variable response of interest y_i was subsequently observed. Hence, the responses constitute a vector $Y = \{y_1, y_2, \ldots, y_n\}$. Y represents our variable of concern i.e. PPL. Let the binary classes of Y be representable by 0 and 1. Logistic model approach posits that the relationship between Y and $\mathbf{X}$ can be modeled as;

$$\log\left(\frac{p^{(1)}}{1-p^{(1)}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad (1)$$

where, β_0 is the coefficient of intercept and $p^{(1)}$ is the observed proportion of subjects in response class 1 i.e. the probability of belonging to response category 1. Obviously, the probability of belonging to response category 0 will be $1 - p^{(1)}$. Equivalently, Equation 1 could be represented in terms of odds ratio as;

$$\frac{p^{(1)}}{1-p^{(1)}} = \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k\right) \quad (2)$$

Thus, a primary aim of most logistic regression analysis is to estimate the coefficients $\underline{\beta}$. β's are the coefficients of our exploratory variables.

5.3 Survey Result

Our results obtained showed that the odds ratio (OR) for `Male` under `Neutral-PPL` implies that, on average, there is about 27.27% (= 1.2727 - 1) higher chance that a male will choose a neutral preference level over other preference levels. OR measures association between our explanatory variables (such as age group, gender) and our response of interest (PPL).

Our result further shows the estimated odds ratio for `Male` under `Low-PPL` is 1. This implies that, on average, the (independent) choices of a male and a female will be similar when considering choosing between low and other privacy levels.

With respect to age-group, our result illustrates that, on average, relative to an individual in the `26-30` age group, an individual in the `31-35` age group has a significantly higher chance of choosing a high privacy preference level over other privacy preference levels. The estimated odd of this claim is 12.5:1 with a standard error of about 3.4723. This deduction is supported by the plot on Figure 2. The large standard er-

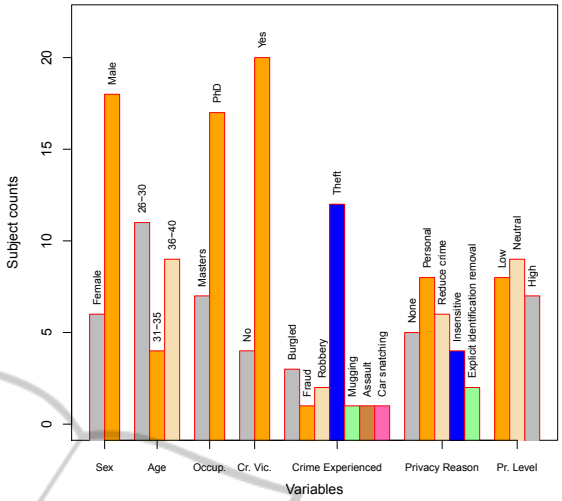

Figure 1: Histogram illustrating the distribution of subjects over the different categories of variables surveyed in the primary study of privacy level preference.

Table 3: Description of the different categories of the surveyed variables used in the ensuing analysis. The table shows number of subjects observed for each variable category.

Variable	*Categories*	*Subjects*
Sex	`Male`	18
	`Female`	6
Age group	`26 - 30`	11
	`31 - 35`	4
	`36 - 40`	9
Occupation	`PhD`	17
	`Masters`	7
Privacy choice reason	`None`	5
	`Adaptive`	1
	`Personal`	8
	`Insensitive`	4
	`Reduce crime`	6
	`Explicit ID delete`	2
Crime Victim?	`Yes`	20
	`No`	4
Crime experienced	`Burgle`	3
	`Fraud`	1
	`Robbery`	2
	`Theft`	12
	`Mugging`	1
	`Assault`	1
	`Car snatching`	1
Preferred privacy level	`Low`	8
	`Neutral`	9
	`High`	7

ror observed on the table is highly inevitable given the sample size of the analysis data. This proves that age is a key factor in choosing privacy preference. This implies that age is a key element in choosing privacy.

In order to test if there is a relationship between the type of crime experienced, Privacy Choice and Preferred Privacy Level (PPL), we used subset regressions. All subsets regression begins by fitting separate models of the response against each of the explanatory variables. Subset Regression is a procedure to check for relationship between individual variable (e.g. Crime Experienced) and PPL. The intercept is also considered as an explanatory variable in this scenario. Based on some model selection criteria the best among these one-variable models is selected. This work uses the *Alkaike Information Criteria* (AIC) as its model selection criterion: the lower the AIC the better. Our result shows that, the different categories of `Privacy choice reasons` seems effective, except of `IPR: Insensitive` which was not selected as an element of an optimal categorization in any of the subset sizes. This implies that the following factors affect people's privacy choice: personality (PPR), Reduction of crime (RPR), Cultural Background (Adaptive).

Applying similar ideas, as those from the previous paragraph, to the results for `Crime experienced` allows substantial pruning of the categories of the variable to just `Robbery`, `Theft` and, `Car snatching`. From our study, only these three crimes affect People's Privacy Choice out of the seven different types of crime investigated.

In summary, although it might be necessary to gain more power through increased sample size in the main study, the hypothesis of non-optimal categorization of the `Privacy choice reason` and `Crime experienced` variables cannot be rejected. It is therefore claimed that, it is sufficient to say that people's privacy choices is affected by the following reasons: `Age`, `Personality`, `Community Need`, `Explicit Identity Removal` and `Cultural Background:Adaptive`. On the other hand, the following categories of `Crime experienced` affects people's privacy choice: `Robbery`, `Theft` and, `Car snatching`.

6 EXPECTED OUTCOME

Our preliminary survey has shown the feasibility of our approach in our domain area. Results from our survey shows that the feasibility of integrating a three-tier privacy into k-anonymity. As future work, we will be carrying out a real-life implementation of integration of our personalised privacy into k-anonymity algorithm using three-tier privacy scheme. We also hope to increase our survey subjects.

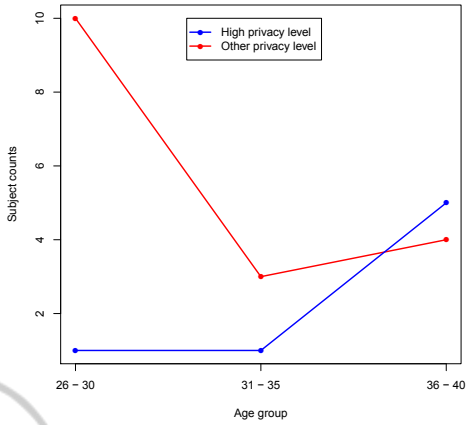

Figure 2: Illustration of the degree of association between high privacy preference level and potential confounder `Age group`.

7 CONCLUSION

Analyzing data is crucial for knowledge-driven decision support in most organisations, including the crime domain. However, lack of data analytics expertise within most law enforcement agencies in developing nations has necessitated the need to have third-parties (honest-but-curious) data analytics provider intervene in order to aid fast crime report analysis. Data Anonymization techniques have proven to be reliable solutions to help in overcoming the limitations that prevents the engagement of third-parties expertise in anlysing data. Nonetheless, anonymisation techniques proposed in the past such as Cryptography and Swapping have limitations such as additive noise and inference attack. K-anonymisation technique has proven to be a promising alternative in overcoming these shortcomings. However, the generic paradigm approach to privacy enforcement in K-anonymity model needs to be refined, in order to improve efficiency in this regard.

Therefore, this research re-emphasizes the need to integrate users privacy preference into k-anonymity model, thereby improving its efficiency. The users privacy preference ensures that users privacy specification or need is well represented, such that the users data is neither over-protected nor under-protected. Our proposed approach integrates the concept of a three tier-privacy level (low, medium and high) into k-anonymity to achieve anonymization. This helps us to identify individual users best choice and how

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

users privacy preference can be incorporated into the K-anonymity model.

To establish the feasibility of our approach, we carried out a preliminary survey in our domain area. Results from our survey show that during crime reporting, there is about 27.27% higher chance that a male person will choose a neutral preference level over other preference levels. Our work is relevant and critical in that it improves upon the K-anonymisation model, thereby improving its efficiency. The next phase of our research would consider verifying the scalability of our approach by introducing different attacks such as unsorted matching attack.

REFERENCES

Aggarwal, C. C. and Philip, S. Y. (2008). TA general survey of privacy-preserving data mining models and algorithms. *Springer US*.

Bayardo, R. J. and Agrawal, R. (2002). Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on (pp. 217-228)*.

Byun J. W., Bertino, K. A. B. E. and Li, N. (2006). Efficient k-anonymity using clustering technique.

Chuang, I., L. S. H. K. and Kuo, Y. (2011). An effective privacy protection scheme for cloud computing. In *13th International Conference on Advanced Communication Technology (ICACT, 2011), IEEE, 260-265*.

Dwork, C. (2006). Differential privacy. In *Automata, languages and programming (pp. 1-12)*.

Gedik, B. and Liu, L. (2008). Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *Mobile Computing, IEEE Transactions on, 7(1), 1-18*.

Guo, K. and Zhang, Q. (2013). Fast clustering-based anonymization approaches with time constraints for data streams. *Knowledge-Based Systems, Elsevier*.

Jiang, W. and Clifton, C. (2006). A secure distributed framework for achieving k-anonymity. *The VLDB JournalThe International Journal on Very Large Data Bases, 15(4), 316-333*.

Kabir, M. E., W. H. and Bertino, E. (2011). Efficient systematic clustering method for k-anonymization. *Acta Informatica, 48(1), 51-66*.

SAIRR (2013). Over half of crime go unreported, press release for immediate release.

Sakpere, A. B. and Kayem, A. V. D. M. (2014). *A state of the art review of data stream anonymisation schemes*. Information Security in Diverse Computing Enviroments, 24. IGI Global, PA, USA., USA.

Samarati, P. (2001). Protecting respondents identities in microdata release. *Knowledge and Data Engineering, IEEE Transactions on, 13(6), 1010-1027*.

Skrondal, A. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika 68(2): 267287*.

Sweeney, L. (2001). Computational disclosure control: A primer on data privacy protection. *Thesis (PhD), Massachusetts Institute of Technology, Cambridge, MA, 2001*.

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10 (05), 557-570*.

Sweeney, L. (2002b). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10 (05), 557-570*.

Xiao, X. and Tao, Y. (2006). Personalized privacy preservation. *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS