I.J. Information Technology and Computer Science, **2015, 10, 1-7**
Published Online September 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2015.10.01

L–Diversity-Based Semantic Anonymaztion for Data Publishing

Emad Elabd, Hatem Abdulkader, Ahmed Mubark
Faculty of computers and information, Menoufia University, Egypt
E-mail: {emadqap@gmail.com, hatem6803@yahoo.com, ahmed_mubark9@yahoo.com}

Abstract—Nowadays, publishing data publically is an important for many purposes especially for scientific research. Publishing this data in its raw form make it vulnerable to privacy attacks. Therefore, there is a need to apply suitable privacy preserving techniques on the published data. *K*-anonymity and *L*-diversity are well known techniques for data privacy preserving. These techniques cannot face the similarity attack on the data privacy because they did consider the semantic relation between the sensitive attributes of the data. In this paper, a semantic anonymization approach is proposed. This approach is based on the Domain based of semantic rules and the data owner rules to overcome the similarity attacks. The approach is enhanced privacy preserving techniques to prevent similarity attack and have been implemented and tested. The results shows that the semantic anonymization increase the privacy level and decreases the data utility.

Index Terms—Data publishing, Semantic anonymization, Privacy preserving, Semantic rules, L-Diversity.

I. INTRODUCTION

Due to the emerging of new technologies and tools for data outsourcing and publishing such as Cloud Computing, privacy preserving of the published data is become one of most important research topics. Many companies and organization collect personal data, stored and published on the cloud for research purposes. These data often contain sensitive information about individuals such as medical records that shows the types of diseases that each person had. When releasing microdata, it is necessary to prevent the sensitive information of the individuals from disclosed, sharing or publishing. Various techniques are designed to reduce risks of disclosing the data and preserve the privacy of the released microdata. Data anonymization is one of the privacy preserving techniques widely used of making the original data worthless to any person except the owners. Generally, released data are stored in format of microdata tables. Microdata table divided into identifiers, quasi-identifiers (*QI*) and sensitive attributes (*SA*). Although, explicit identifiers such as names, phone numbers and social security numbers about individuals are removed or encrypted before the microdata are released, the privacy disclosure problem still exists where there are two types of information, identity and attribute, are disclosure [1], [2]. Several models are used to treat the leakage of information about individual and improve data anonymization include *K*-anonymity[5] and *L*-diversity[16].

A. Motivation

To explain the similarity attach in the *L*-diversity approach, consider the following scenario. Table 1 shows the data of original medical records without identities. The table contains Gender, Zip Code and Age attributes as Quasi-identifier and Salary is a numeric sensitive attribute, Disease is categorical sensitive attribute. For the research purposed, this table should be publish. If the table is published in this raw form, it will be vulnerable for all types of privacy attack. Therefore, an anonymization technique should be applied before publishing. We will apply the *L*-diversity technique on the table before publishing it to guarantee the preserving privacy of its data. Table 2 derived from table1 after applying data anonymization with 3-diversity. Tuples divided into three equivalent classes. Each class consists of three different values sensitive attributes.

Table 1. Original microdata.

QID			SA	
Gender	**ZIP Code**	**Age**	**Disease**	**Salary**
Male	400071	35	bronchitis	10k
Male	400182	37	pneumonia	11k
Male	400095	39	stomach cancer	12k
Female	440672	54	gastritis	12k
Female	440123	58	Flu	15k
Male	440893	54	bronchitis	16k
Male	400022	41	gastric ulcer	16k
Male	400135	46	gastritis	17k
Female	400182	44	stomach cancer	18k

Table 2. 3-Diverse version of dataset.

QID			SA	
Gender	**ZIP Code**	**Age**	**Disease**	**Salary**
*	400*	>30	bronchitis	10k
*	400*	>30	pneumonia	11k
*	400*	>30	stomach cancer	12k
*	440*	5*	gastritis	12k
*	440*	5*	Flu	15k
*	440*	5*	bronchitis	16k
*	400*	>40	gastric ulcer	16k
*	400*	>40	gastritis	17k
*	400*	>40	stomach cancer	18k

One the problem of the resulted 3-diverse table in terms of privacy is the similarity attack. Suppose that an adversary by some way knows that Bob has record in the first equivalence class, then he can know that Bob's salary ranges from 10K-12K. As a result, the adversary can semantically concludes that Bob has a low salary. In addition, knowing that Bob's record belongs to third equivalence class enables the attacker to conclude that Bob has some stomach-related problems, because all the three diseases in the class are stomach-related. This leakage of sensitive information occurs because *L*-diversity does not consider semantic relation among sensitive values. Therefore, the data privacy is susceptible to be disclosed.

In this paper, an approach is proposed to overcome similarity attack for *L*-diversity anonymization technique. The proposed approach is based on the determination of the semantic rules that leads to the similarity attack and apply an anonymization process based on these rules.

The rest of the paper is organized as follows: the brief background about *K*-anonymity and ℓ -diversity are discussed in section 2. In section 3, the related works to anonymization methods and their related techniques are discussed. Section 4 presented the proposed model. The experimental results is shown in section 5. Finally in section 6, conclusion is provided.

II. Background of K-anonymity and L-diversity

To provide protection against attribute linkage, Lantaya Sweeney [5] proposed *k*-anonymity as model for privacy preserving of QI. In this model, each record in the released data table is indistinguishable with at least *k*-1 other records within the dataset with respect to a set of quasi-identifier attributes. Generalization and suppression methods in which attribute value is replaced by a less specific, more general value that is faithful to the original one are used to achieve *k*-anonymity requirement for data anonymization. In this case, adversary cannot uniquely identify individuals, so the individuals' privacy could been preserved. *K*-anonymity solved the problem of identity disclosure but it does not provide protect against attribute disclosure. It is unable to protect the privacy form homogeneity attacks and the background knowledge attacks. The privacy preserving *L*-diversity model [16] are proposed to overcome the drawback of *k*-anonymity by grouping the sensitive attribute values with the same quasi-identifier into equivalent class (EC).

Definition 1 (The *L*-diversity Principle):

An equivalence class (EC) is said to have *L*-diversity if there are at least *L* "well -represented" values for the sensitive attribute (SA). A table is said to have *L*-diversity if every equivalence class (EC) of the table has *L*-diversity [16]. *L*-diversity ensured privacy to sensitive attribute value of a particular person unless the adversary has enough background knowledge to eliminate $L-1$ sensitive attribute values in the person's EC. The *L*-diversity model provides privacy even when the data publisher does not know the background knowledge the adversary. There is a set of limitation of the *L*-diversity. One of the limitation is the Skewness attack since *L*-diversity does not consider the overall distribution of sensitive values. The second limitation is the similarity attack since it does not consider semantics of sensitive values [18]. The third limitation, it does not prevent the probabilistic inference attacks. Fourth limitation, it may be difficult and unnecessary to achieve.

III. Related Work

Privacy have become an important issue and considerable progress has been made with data anonymization. Publisher should take consider never publish microdata to researcher groups in its raw form. Most recent studies on privacy have focused on devising anonymization algorithms. One of the important approaches, proposed by Samarati and Sweeney [3–5], is *k*-anonymization. They presented a framework for generalization and suppression based k-anonymity, where the notion of generalization hierarchies was formally proposed. Given a predefined domain hierarchy, the problem of k-anonymity is how find the minimal domain generalization so that, for each tuple in the released microdata table, there exist at least $k-1$ other tuples that have the same quasi-identifiers.

Zhao at el.[6] introduced several privacy preserving methods in data publishing such as randomization, sampling, suppression, data swapping and perturbation. Suppression replace individual attributes with a* and generalization replacing individual attributes with a broader category.

Aggarwal et al. [7] show that suppressing the sensitive values chosen by individual records own is insufficient for privacy protection because the attacker can use association rules learnt from the data to recover the suppressed values. They proposed a heuristic algorithm to suppress a minimal set of values such that the hidden values are not recoverable by weakening the association rules. Padam Gulwani. [8] studied the database privacy problems caused by data mining technology and proposed algorithm for hiding sensitive data in association rules mining. Data swapping and data suppression two methods suggested to protect data privacy [9–11]. Although, they could not quantify how well the data is protected, the privacy preserving process takes various stages in its development according to the level of complexity in the existing technique. Aggarwal et al.[12], [13] discussed the curse of dimensionality related to k-anonymity and proposed a general model to solve the problem of finding optimal generalization and suppressions to achieve *k*-anonymity. It can accommodate a variety of cost metrics.

To improve the quality of the anonymized data, In [14] introduced incognito approach in which generalization hierarchies were explored in a vertical way to efficiently compute minimal and optimal generalizations. It computes a minimal solution to *k*-anonymity in the generalization hierarchy for each quasi-identifier. These solutions are combined to form the candidate

generalizations for the domain hierarchies of quasi-identifier pairs.

X. Xiao and Y. Tao. [15] proposed a new generalization framework based on the concept of personalized anonymity to perform the minimum generalization for satisfying everybody's requirements and retains the largest amount of information from the microdata.

Although, many works had proposed efficient algorithms for *k*-anonymity, there are a set of drawbacks of k-anonymity as a measure of privacy are appeared[16]. For instance, k-Anonymity does not provide privacy if the sensitive values in an equivalence class lack diversity. In addition, it does not provide privacy if the attacker has background knowledge

Machanavajjhala at el. [16] proposed an alternative property of ***L***-diversity to ensure privacy protection in the microdata disclosure, and demonstrated that algorithms developed for *k*-anonymity. The concept of ***L***-diversity introduced to prevent attackers with background knowledge. The ***L***-diversity would ensure that there are at least ***L*** distinct values for the sensitive attribute in each equivalence class. However, it does not prevent probabilistic inference attacks. This motivated the development of two stronger notions of ***L***-diversities called entropy ***L***-diversity and recursive (c,***L***)-diversity [16]. In the entropy ***L***-diversity, the different sensitive values must be distributed evenly enough for each equivalence class with different sensitive values. Recursive (c,*L*)-diversity also assures that "the most frequent value of sensitive attribute in each equivalence class is not too frequent, and the less frequent doesn't appear too rare". Less restrictive instantiation of diversity, called recursive (*c*,***L***)-diversity, compared most frequent sensitive values and least frequent sensitive values in equivalence class. The (α,k)-Anonymity which is like the (c,*L*)-diversity is proposed by Wong at el. [17] to investigate the privacy on data published in cloud.

Li et al. [18] proposed the t-closeness model which is an enhancement on the concept of ***L***-diversity. One characteristic of the ***L*** -diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data. *T*-Closeness considered the distribution of the sensitive attributes to be close to the distribution of that sensitive attribute values in the whole table. It restricts the distance between the distribution of a sensitive attribute in each equivalence class within the same quasi-identifiers and the distribution of the same sensitive attribute in the whole table. However, t-closeness doesn't make much improvement and harm the data utility when the value of *t* is small because enforcing *t*-closeness destroys the correlations between quasi-identifier attributes and sensitive attributes [19], [20].

Yeye et al. [21] proposed K^m-anonymity approach that is based on the top down local generalization process to record the number of transaction records.

Michal and Kern.[22] presented reasons why a publisher should never publish microdata to researcher groups in its raw form, studied *K*-Anonymity, ***L***-Diversity and *t*-Closeness as the principles to gain a certain anonymity level.

IV. The Proposed Approach

The proposed model is based on the ***L***-diversity approach and the semantic extraction techniques. The semantic extraction is based on a domain-based semantic rules repository and semantic rules that can be assigned by the owner of the data that will be published. The key idea of the proposed approach is that each semantic rule leads to a piece of data that can be extracted and effects on the data privacy, will be used in the anonymization of the data. The semantic rules that produce data which have an impact in the privacy are called "effective semantic rules". Therefore, applying the domain-based repository and data owner semantic rules will lead to a set of extracted information. The anonymizer determined from these information which pieces of information are important to be anonymized and then select the effective semantic rules that produce these pieces of information. These selected effective semantic rules will be used in the anonymization of the ***L***-diversity table. The approach can be summarized in the following steps:

1- Apply *L*-diversity on the original table.
2- Data extraction and effective semantic rules determination process.
3- Anonymization based on the effective semantic rules.

A. Data extraction and effective semantic rules determination

The semantic extraction process is based on the premise of relation among attribute values. This process is applied on the anonymized table that is resulted from the *L*-diversity process. It starts by clustering the anonymized table into equivalence classes (ECs). Each EC has domain generalization unique QI value and distinct sensitive attribute values. Then, the approach starts to check the semantic relations between sensitive values in each EC based on the semantic rules repository and data owner semantic rules. If there is an extracted piece of information that impacts on the privacy, then store this semantic rule to be used in the anonymization phase. The resulted anonymized table cannot be affected by the similarity attacks. Fig. 1 Shows the pseudo code algorithm for performing Data extraction and effective semantic rules determination process. The algorithm starts to scan the table by checking each equivalence class of the sensitive attributes. Then apply the semantic rules from the domain-based and data owner to check the type of the extracted data. If a piece of the extracted data impacts on the privacy then the algorithm store the rule which produce this data as an effective semantic rule and assign an anonymization action for this rule.

For instance, consider table 3 that has an anonymized data with Entropy 3-diversity. It consists of two sensitive attributes (SA){salary, disease} and three Quasi-identifier (QI) {zip code, Age, Nationality}. If the semantic a semantic rule saying that "salaries with values 0 to 20 K

is very low salary". If this rule is applied on the second the equivalence class which has Quasi-id {476**, [22, 30[, *}, then an extracted piece of information say that all members of this class have low salary. Therefore, the previous rule should be stored as an effective semantic rule for this class. In addition an anonymization action should be assigned to this rule. For instance increasing the values by a specific value to change their slide of salary to medium or high slide. Besides, the semantic rules can be applied on different categories of data rather than numeric. For instance, the semantic rule that say" the bronchitis, lung cancer, cough, flu and bronchitis are chest-related diseases" can be applied on the "disease" attribute on equivalence class which has Quasi-id {476**, [30, 40[, *}.

Algorithm: Semantic extraction algorithm
Input: L-diversity Table (T)
Output: Set of effective semantic rules
While T isn't end do
For each equivalence class V from T do
Apply semantic rules on the sensitive attribute;
If there is extracted data impacts on privacy resulted by a semantic rule then
A. Store the semantic rule as an effective semantic rule.
B. Set an anonymization action to this rule.
End while

Fig.1. Semantic extraction

Table 3. Micro data 3-diversity

zip code	Age	Nationality	disease	salary
148**	[22-30[	*	cancer	30k
148**	[22-30[	*	Heart disease	33k
148**	[22-30[	*	cancer	35k
148**	[22-30[	*	Heart disease	35k
148**	[22-30[	*	cancer	40k
148**	[22-30[	*	viral infection	38k
476**	[22-30[	*	gastric ulcer	8k
476**	[22-30[	*	gastritis	11k
476**	[22-30[	*	stomach cancer	6k
476**	[22-30[	*	stomach cancer	13k
130**	[30-40[	*	Heart disease	10k
130**	[30-40[	*	viral infection	9k
130**	[30-40[	*	cancer	15k
130**	[30-40[	*	viral infection	7k
130**	[30-40[	*	Heart disease	11k
476**	[30-40[	*	bronchitis	15k
476**	[30-40[	*	lung cancer	12k
476**	[30-40[	*	lung cancer	14k
476**	[30-40[	*	cough	18k
476**	[30-40[	*	flu	17k
476**	[30-40[	*	bronchitis	9k

B. Semantic Anonymization

Anonymization phase is based on the effective semantic rules which are resulted from the previous phase. After getting the effective semantic rules, the anonymizer assigns a suitable anonymization action to each EC stored in extracted knowledge table. The generalization is used for privacy preserving in this phase. In the generalization, the anonymizer try to find a general representation of a set of data in the EC that is susceptible to privacy attack in conjunction with the other EC in the table. This process is performed by using a merge method. If there is no EC that can be merged with the EC that is susceptible to privacy attack then the suppression technique is used. Suppression process starts to suppress values from the QIs of the susceptible EC. Then the anonymizer checks for the best match from the other EC according to QI. If there is no math in terms of QIs, Incognito algorithm [20] is used to perform generalization to the susceptible QIs EC and best EC matching. Then the merge process is performed between the two EC. In some cases, the generalization affects in the utility and destroy the data. Therefore, the anonymization can be applied directly on the sensitive data of the susceptible EC.

Table 4 shows the EC and the effective semantic rule for it and the proposed action for the anonymization. After applying the anonymization, the data is shown in Table 5. It is noticed that the EC which has QI {476**, [30, 40[, *} is merged with EC which has QI {476**, [22, 30[, *} into one EC where become has QI {476**, [20, 40[, *} and contain all sensitive values that were in the two EC. Therefore, prevent attacker from discloser all diseases in EC that are semantically related with chest disease. In addition, in the second EC the sensitive values is changed by addition 10 to the original sensitive value. Therefore, it prevents the attacker to know that salary is low for a certain person based in some background information.

V. Experiments Results

The effect of the similarity attack on data after semantic anonymization is tested in this section. We used two terms of data metrics, namely information loss and data utility. The Java expert system shell (JESS) rule engine and scripting (CLIPS) are used to write semantic rule. Rete algorithm to process rules is used in Java. The proposed model is implemented using Java. The specification of the platform is Intel core i5 2.10 GHz processor and RAM 3 GB on windows 8. The dataset used in the experiment is the adult dataset from the UC Irvine machine-learning repository [21]. According to the hierarchy described in [23], we can consider the first 7 attributes as the quasi-identifier and one as a sensitive attribute. Table 6 describes the attributes for dataset used in the experiments and the number of distinct values for each attribute. According to behavior of the people with different occupation, the values of the sensitive attribute occupation (14) divided into three equal-size group based on the semantic relation between the values as shown in table 7.

Table 4. Table with the effective semantic rules and the anonymization action.

zipcode	Age	National	disease	salary	Rules	action	apply
148**	[22-30[	*	cancer	30k	-	-	-
148**	[22-30[	*	Heart disease	33k	-	-	-
148**	[22-30[	*	cancer	35k	-	-	-
148**	[22-30[	*	Heart disease	35k	-	-	-
148**	[22-30[	*	cancer	40k	-	-	-
148**	[22-30[	*	viral infection	38k	-	-	-
476**	[22-30[	*	gastric ulcer	8k	<20→ Low	Generalization, Adding +10	FALSE
476**	[22-30[	*	gastritis	11k	<20→ Low	Generalization, Adding +10	FALSE
476**	[22-30[	*	stomach cancer	6k	<20→ Low	Generalization, Adding +10	FALSE
476**	[22-30[	*	stomach cancer	13k	<20→ Low	Generalization, Adding +10	FALSE
130**	[30-40[	*	Heart disease	10k	<20→ Low	Adding +10	TRUE
130**	[30-40[	*	viral infection	9k	<20→ Low	Adding +10	TRUE
130**	[30-40[	*	cancer	15k	<20→ Low	Adding +10	TRUE
130**	[30-40[	*	viral infection	7k	<20→ Low	Adding +10	TRUE
130**	[30-40[	*	Heart disease	11k	<20→ Low	Adding +10	TRUE
476**	[30-40[	*	bronchitis	15k	<20→ Low	Generalization, Adding +10	FALSE
476**	[30-40[	*	lung cancer	12k	<20→ Low	Generalization, Adding +10	FALSE
476**	[30-40[	*	lung cancer	14k	<20→ Low	Generalization, Adding +10	FALSE
476**	[30-40[	*	cough	18k	<20→ Low	Generalization, Adding +10	FALSE
476**	[30-40[	*	flu	17k	<20→ Low	Generalization, Adding +10	FALSE
476**	[30-40[	*	bronchitis	9k	<20→ Low	Generalization, Adding +10	FALSE
476**	[30-40[	*	bronchitis	15k	chest	Generalization, Adding +10	TRUE
476**	[30-40[	*	lung cancer	12k	chest	Generalization, Adding +10	TRUE
476**	[30-40[	*	lung cancer	14k	chest	Generalization, Adding +10	TRUE
476**	[30-40[	*	cough	18k	chest	Generalization, Adding +10	TRUE
476**	[30-40[	*	flu	17k	chest	Generalization, Adding +10	TRUE
476**	[30-40[	*	bronchitis	9k	chest	Generalization, Adding +10	TRUE

Table 5. 3- Diversity after anonymization.

zipcode	Age	Nationality	disease	salary
148**	[22-30[	*	cancer	30k
148**	[22-30[	*	Heart disease	33k
148**	[22-30[	*	cancer	35k
148**	[22-30[	*	Heart disease	35k
148**	[22-30[	*	cancer	40k
148**	[22-30[	*	viral infection	38k
130**	[30-40[	*	Heart disease	20k
130**	[30-40[	*	viral infection	19k
130**	[30-40[	*	cancer	25k
130**	[30-40[	*	viral infection	17k
130**	[30-40[	*	Heart disease	21k
476**	[20-40[	*	gastric ulcer	18k
476**	[20-40[	*	gastritis	21k
476**	[20-40[	*	stomach cancer	16k
476**	[20-40[	*	stomach cancer	23k
476**	[20-40[	*	bronchitis	25k
476**	[20-40[	*	lung cancer	22k
476**	[20-40[	*	lung cancer	24k
476**	[20-40[	*	cough	28k
476**	[20-40[	*	flu	17k
476**	[20-40[	*	bronchitis	9k

Table 6. Description of dataset used in the experiment.

	Attribute	Distinct values	Generalization type
1	Gender	2	Suppression(1)
2	Age	74	Ranges-5,10,20
3	Race	5	Suppression(1)
4	Marital Status	7	Hierarchy(2)
5	Education	16	Hierarchy(3)
6	Native Country	41	Hierarchy(2)
7	Work Class	7	Hierarchy(2)
8	Occupation	14	Sensitive Attribute

Table 7. Sensitive attribute group.

Group	Values
1	Tech-support, Craft-repair, Prof-specialty, Machine-op-inspct
2	Sales, Exec-managerial, Handlers-cleaners
3	Other-service, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces

In the previous data set, if equivalence classes are susceptible to the similarity attack, when it had all values on one group. L-diversity algorithms is used to generated (Entropy, recursive(c, L))-diversity with different values for L. The proposed approach applied the semantic rules and data owner semantic rules on the resulted L-diversity

to extract data that are nominated to be used in similarity attack. After all outlying classes from dataset are deleted, when the Entropy L=3 generated diversity table the proposed approach show 10 ECs can be used by the similarity attack. Likewise, the recursive(c=4,3)-diversity generates diversity table that contains 20 ECs vulnerable to the similarity attack. The proposed anonymization process to prevent extracted data from similarity attack is achieved. The results of the proposed approach shows a significant enhencement in terms of privacy.In the hand other there is a loss of infromation and decreasing in the utility of data due to the generalization and suppression in anonymization process.

A. Data Quality

We used two metric to measure the utility of anonymous data produced by Entropy L-diversity, recursive(c,L) where c=4 and the semantic anonymization. The first metric is the average size of equivalence classes generated by anonymization algorithm. The average size of the equivalence classes for the three anonymization approaches are shown in Figure 2 with different values of the variable L. The results shows that the semantic anonymization has the highest information loss when L is large which decrease the data utility. Therefore, the data publisher should balance between the level of the required privacy and the utility of the data (information loss). The second metric is the Discernibility Metric (DM)[24] which measures the number of tuples that are indistinguishable from each other. This process is based on assigning penalty to each tuple based on the number of tuples Indistinguishable from that tuple in anonymized table. The DM cost defines the information loss for generalization and suppression. DM cost for the three approaches are shown in Figure 3 with different values of the variable L.

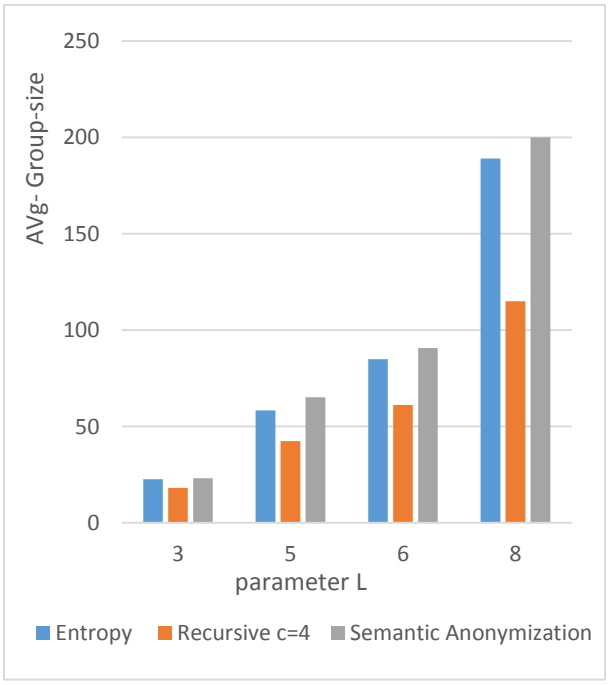

Fig.2. Average size of equivalence class

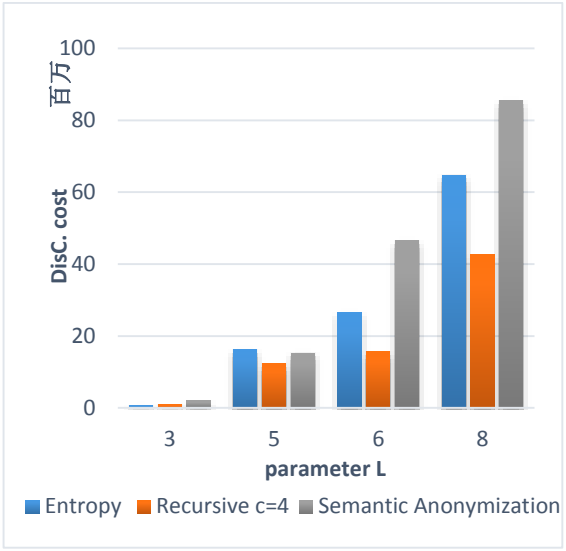

Fig.3. Discernibility Metric

VI. Conclusion

Privacy preserving approaches of the published data such as k-anonymity and l-diversity suffer from the similarity attacks because of the semantic relationship that can be among the sensitive attribute values. In this paper, the domain-based semantic rules and data owner semantic rules are used for anonymization process that overcome the similarity attacks on privacy. The results shows that the increase in privacy level using the proposed approach effects on the utility of the data. For the future work, the semantic anonymization algorithm needs to be optimized to decrease the information loss and a dynamic version is provided based with a deterministic relation between the utility and the privacy level.

References

[1] G. T. Duncan and D. Lambert, "Disclosure-limited data dissemination," *Journal of the American statistical association*, vol. 81, no. 393, pp. 10–18, 1986.

[2] D. Lambert, "Measures of disclosure risk and harm," *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, vol. 9, pp. 313–313, 1993.

[3] P. Samarati, "Protecting respondents identities in microdata release," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 13, no. 6, pp. 1010–1027, 2001.

[4] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *PODS*, 1998, vol. 98, p. 188.

[5] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[6] Y. Zhao, M. Du, J. Le, and Y. Luo, "A survey on privacy preserving approaches in data publishing," in *Database Technology and Applications, 2009 First International Workshop on*, 2009, pp. 128–131.

[7] C. C. Aggarwal, J. Pei, and B. Zhang, "On privacy preservation against adversarial data mining," in

Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 510–516.

[8] P. Gulwani, "Association rule hiding by positions swapping of support and confidence," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 4, no. 4, p. 54, 2012.

[9] T. Dalenius and S. P. Reiss, "Data-swapping: A technique for disclosure control," *Journal of statistical planning and inference*, vol. 6, no. 1, pp. 73–85, 1982.

[10] P. Diaconis, B. Sturmfels, and others, "Algebraic algorithms for sampling from conditional distributions," *The Annals of statistics*, vol. 26, no. 1, pp. 363–397, 1998.

[11] G. T. Duncan and S. E. Fienberg, "Obtaining information while preserving privacy: A markov perturbation method for tabular data," in *Joint Statistical Meetings*, 1997, pp. 351–362.

[12] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proceedings of the 31st international conference on Very large data bases*, 2005, pp. 901–909.

[13] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 2005, pp. 217–228.

[14] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005, pp. 49–60.

[15] X. Xiao and Y. Tao, "Personalized privacy preservation," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, 2006, pp. 229–240.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.

[17] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "Framework for scure Cloud Computing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 754–759.

[18] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.," in *ICDE*, 2007, vol. 7, pp. 106–115.

[19] L. H. Cox, "New results in disclosure avoidance for tabulations," *International Statistical Institute Proceedings of the 46th Session*, pp. 83–84, 1987.

[20] L. H. Cox, "Suppression methodology and statistical disclosure control," *Journal of the American Statistical Association*, vol. 75, no. 370, pp. 377–385, 1980.

[21] Y. He and J. F. Naughton, "Anonymization of set-valued data via top-down, local generalization," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 934–945, 2009.

[22] M. Kern, "Anonymity: A Formalization of Privacy-l-Diversity," in *Proceeding zum Seminar Future Internet (FI), Innovative Internet Technologien und Mobilkommunikation (IITM) und Autonomous Communication Networks (ACN)*, 2013, vol. 49.

[23] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 279–288.

[24] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 2005, pp. 217–228.

Authors' Profiles

Dr. Emad ELABD: is working Assistant Professor, Dept. of Information Systems, Menoufia University, Egypt. He got his Ph.D. in the field of Web services compliance over high-level specifications at LIRIS, University Lyon1, France, July 2011. He received bachelor ́s degrees in Electronic Engineering from Menoufia University, Egypt where he did his master ́s studies in computer science also. His research interests include Web services modeling and analysis with access control and time aspects, Web services (specification, composition), Semantic Web, and Information retrieval.

Prof. Hatem Abdulkader is working associate professor in the Information Systems Department, Faculty of Computers and Information; Menoufia University, Egypt. he obtained his BS. and M.SC., both in electrical engineering from the Alexandria University, Faculty of Engineering, Egypt, 1990 and 1995, respectively. He obtained his Ph.D. degree in electrical engineering also from Faculty of Engineering, Alexandria University, Faculty of Engineering, Egypt in 2001. His areas of interest are data security, Web applications and artificial intelligence, and he is specialized in neural networks.

Ahmed Mubark: Master student at Menoufia University, Egypt. He works as a teaching assistant in Education and Computer Sciences faculty of Ibb University, Yemen. His main research interest is Security technologies in cloud computing. He received his BSc in computer science from Ibb University.

How to cite this paper: Emad Elabd, Hatem Abdulkader, Ahmed Mubark,"L–Diversity-Based Semantic Anonymaztion for Data Publishing", International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.10, pp.1-7, 2015. DOI: 10.5815/ijitcs.2015.10.01