2014 IEEE 27th International Symposium on Computer-Based Medical Systems

The Role of Inference in the Anonymization of Medical Records

Athanasios Zigomitros[1], Agusti Solanas[2], Constantinos Patsakis[3]
[1]Institute for the Management of Information Systems, "Athena" Research Center, Greece &
Department of Informatics, University of Piraeus, Greece
[2]Smart Health Research Group, Dept. Computer Engineering & Mathematics, Universitat Rovira i Virgili, Tarragona. Spain
[3]Distributed Systems Group, School of Computer Science and Statistics , Trinity College, Dublin, Ireland
azigomit@imis.athena-innovation.gr,agusti.solanas@urv.cat,patsakik@scss.tcd.ie

***Abstract*—The quality of life has been significantly improved and one of the main reasons is the medical advances of the past decades. Nevertheless, to further advance the research and services in the field, practitioners, researchers and health organizations should share more information. While this need is indisputable, the sensitivity of the information demands that it is preprocessed, so that the published data are anonymized and individuals cannot be identified.**

The scope of this work is to highlight the difficulties in providing automated anonymization approaches for medical records without consulting experts in the field. One of the major problems that is going to be highlighted is that Quasi-Identifiers ($\mathcal{QI}$) are not independent. It is well known that combinations of $\mathcal{QI}s$ can be used to infer other relevant information. Nevertheless, this work tries to exploit the other way of information flow, we show how sensitive attributes can be exploited to derive information about the $\mathcal{QI}s$, leading to many privacy hazards for the patients whose records are shared. To this extent, we illustrate some relevant examples and discuss probable counter-measures.

***Index Terms*—Privacy, data anonymization, medical records**

I. Introduction

The great advances that have been achieved in the past decades in health and medicine have managed to make many diseases that plagued humanity for thousands years obsolete. People that were expected to live short or had severe injuries not only survive, but manage to live without significant changes in their daily living. As a result, the quality of life of people has been drastically improved and the average life expectancy has peaked. Definitely there are more things to be done, nevertheless, we have reached to a point that we can discuss about personalized medical services and diagnoses. While there is a lot of knowledge and a lot of information shared among researchers, it is understood that in order to proceed we need to share even more information. By fusing this information hidden patterns are expected to be detected leading to new discoveries.

The sensitivity of medical records though does not allow jeopardizing, therefore the information before it is shared, has to be preprocessed to anonymize the records of individuals. This will allow the post-processing of the information from others, while simultaneously stop adversaries from extracting the identities of individuals. For instance, if we assume that we have a database with medical records, the first step is to remove names and surnames. Nevertheless, individuals could be identified by combination of other fields such as gender, birthday and zip code [1], therefore special algorithms [2], [3], [4] have been introduced to obfuscate the published information, increasing the uncertainty of possible attackers to desirable levels. It becomes apparent that the published information is corrupted, several fields might be suppressed, generalised, perturbated, or even added with noise, decreasing this way the utility of the information Therefore, the balance between anonymity and utility play a central role in picking which methods are going to be applied on each dataset.

The scope of this work is to highlight the significant challenges in anonymizing medical records. As we are going to show, there is another type of attack that can be launched using this type of data that we call "**inference of $\mathcal{QI}s$ attack**". The attack stems from the nature of medical records, but it can be applied to other data as well. The attack has been overlooked by current state of the art, mainly because the fields of the tables to be anonymized are considered independent, which is not the case for medical records and other datasets. Additionally, most of the attacks are trying to expose the sensitive attributes through the combinations of $\mathcal{QI}s$. The attack that we introduce follows another way. It exploits the knowledge derived from the sensitive attributes in some records to recover generalised and/or suppressed $\mathcal{QI}s$, or to break the anonymization guarantees of anatomized or sliced data, increasing this way the re-identification risk. Therefore, sensitive information about other records can be exposed.

The rest of the article is organized as follows. The next section gives a brief overview of current state of the art in anonymization methods and major attacks. Section 3 introduces the concept of inference attacks by illustrating how it can be applied on a previously anonymized table which adheres the known requirements. Section 4 discusses some probable counter-measures to these attacks and in section 5 we discuss some related work. Finally, the article concludes summarizing our contributions and proposing ideas for future work.

II. Attacking and Defending Anonymized Datasets

A. Attacks on Anonymized Datasets

The attributes in a dataset can be classified in four main categories. Firstly, we have **Explicit Identifiers**, which are

1063-7125/14 $31.00 © 2014 IEEE
DOI 10.1109/CBMS.2014.88

attributes such as Social Security Number, that can identify a person uniquely. A **Quasi-Identifier** ($\mathcal{QI}$) is an attribute that cannot be used to identify a person uniquely by itself, such as birthday, gender and zip code. However, by combining quasi-identifiers one could re-identify a person by narrowing down the possible identities of a specific record. We call **Sensitive attributes** (SA) the information which the adversary tries link to his victim, examples of such attributes are the salary, in case of financial records, or the illness in a medical dataset. Any attribute that does not qualify to any of these three categories and has no importance when disclosed, is called a **Non-Sensitive Attribute**.

Each group that shares the same values on every $\mathcal{QI}$ is called Equivalence Class (EC).

Some of the most well-known attacks on anonymized datasets are the following:

Record Linkage Attack: In this attack scenario, an adversary tries to link a record of the anonymized database to a person. Using the quasi-identifiers of a victim the attacker can form a small group of possible successful links on the released database. Using his knowledge, an adversary can uniquely identify his target as shown by Sweeney [1].

Attribute Linkage Attack: In attribute linkage attack, the adversary may not uniquely identify a person, but he could gain additional knowledge about the target's sensitive attributes. If there is not any diversity on the sensitive values of each group, when groups are formatted based on $\mathcal{QI}$s, an adversary could infer the sensitive value of a person, even if he cannot point which record belongs to him.

Table Linkage Attack: On the two aforementioned attacks, it is assumed that the adversary already knows that his target record is on the released table. However, this is not always the case. Sometimes the presence or the absence of an individual from a released table can reveal sensitive information about him [5]. If an adversary can confidently infer for the presence or absence of victim in a released table, then he has sucessfully carried what we call a *table linkage* attack.

Background Knowledge Attack: In background knowledge attack, as described in [3], the adversary uses his knowledge about the victim to infer with great probability sensitive information about it. To illustrate their attack impact, give an example in which an adversary manages to deduce that a patient has a specific viral infection and not a heart related disease. This is achieved by exploiting the background knowledge that heart diseases are rare to Japanese. Thus, the adversary combines the knowledge about his victim with the SA and prunes many possible tuples.

Skewness Attack: Skewed distribution of a SA can be used to expose other attributes, as shown in [4]. To understand how this can be achieved, assume that a SA has a skewed overall distribution with two distinct values, with probability of the first one being 1% and 99% for the other. In an EC if both values have 50% probability, then the probability of the first value has been increased from 1% to 50%, resulting to a privacy breach for the individuals in that EC.

Similarity Attack: When the SA values in an EC are semantically similar to each other, then an adversary can infer important information about his target [4]. Assume than in an EC all the SA values have an heart related disease. The adversary does not know which disease his target has, but he can infer that is in the target's heart.

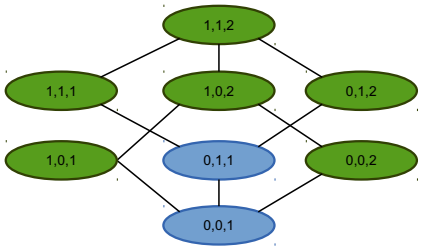

Fig. 1. Lattice

B. k-anonymity

Samarati and Sweeney [1],[6] introduced the notion of k-anonymity. A dataset is called k-anonymous if for any query, based on any set of quasi-identifiers, returns at least k records. In other words, a record must be indistinguishable from at least $k-1$ other records with respect to quasi-identifiers.

A formal definition of k-anonymity, as given by Machanavajjhala et al. in [3], is the following:

Definition 1 (k-anonymity): A table T is k-anonymous if for every record (tuple) $t \in T$ there exist $k-1$ other records $t_{i_1}, t_{i_2}, ..., t_{i_{k-1}} \in T$ such that $t[\mathcal{C}] = t_{i_1}[\mathcal{C}] = t_{i_2}[\mathcal{C}] = ... = t_{i_{k-1}}[\mathcal{C}],\ \forall \mathcal{C} \in \mathcal{QI}$

One way to successfully apply the k-anonymization method,is transforming the data by performing generalization and/or suppression. Generalization replaces all values of a $\mathcal{QI}$ attribute in an EC by a more general value that contains them. For example, $\mathcal{QI}$ gender with values "male" and "female" can be generalized to "person", $\mathcal{QI}$ "age" with values 22, 27 and 35 can be generalized to [22 - 35] and $\mathcal{QI}$ zip code with values 50100, 50120 and 50140 can be represented with the value 501** . The suppression deletes some $\mathcal{QI}$ values or even tuples, which could also be generalized to the most general state of the $\mathcal{QI}$. Notice that the $\mathcal{QI}$ Gender when is generalized to the value person does not provide any information about the gender, so if we supress this value, there is no additional gain in terms of privacy. Suppression can be considered as an extreme case of generalization. An example of a Table with Explicit Identifiers, Quasi-Identifiers and Sensitive Attributes is illustrated in Table I.

k-anonymity is currently considered a well-known standard in data anonymization, however it can only be considered as a baseline as suffers from a serious limitation. While it takes into consideration the $\mathcal{QI}$ attributes, it ignores the SA attributes. This flaw makes k-anonymity method vulnerable to several attacks, such as the already discussed *Attribute Linkage* attack. ℓ-diversity [3] adds a requirement of at least ℓ different values to be in each EC. However, the ℓ-diversity model failed to protect record privacy against the *Skewness*

TABLE I
CLASSIFICATION OF ATTRIBUTE - EXAMPLE

EI	QI			SA
Name	**Gender**	**Zip Code**	**Age**	**Disease**
Alice	F	50100	22	Uterine Cancer
Bob	M	50100	45	Flu
Carolain	F	50100	33	Mastitis
Dennis	M	50100	25	Stomach cancer
Ethan	M	50100	56	HIV
Fay	F	50100	65	Coronary heart disease
George	M	50120	34	Hepatitis
Heather	F	50120	18	Obesity
Ian	M	50120	54	Diabetes
John	M	50120	73	Prostate Cancer
Keith	F	50120	88	Alzheimer
Lea	F	50120	14	Juvenile idiopathic arthritis

TABLE II
K=3 LATTICE 1,0,1

Gender	Zip Code	Age	Disease
*	50100	14-34	Uterine Cancer
*	50100	45-88	Flu
*	50100	14-34	Mastitis
*	50100	14-34	Stomach cancer
*	50100	45-88	HIV
*	50100	45-88	Coronary heart disease
*	50120	14-34	Hepatitis
*	50120	14-34	Obesity
*	50120	45-88	Diabetes
*	50120	45-88	Prostate Cancer
*	50120	45-88	Alzheimer
*	50120	14-34	Juvenile idiopathic arthritis

Attack. To overcome the *Skewness Attack*, Li et al. proposed the notion of t-closeness [4]. t-closeness guarantees that the cumulative difference of SA values inside a EC is not more than a given threshold t when compared to the overall dataset. Afterwards, Brickell and Shmatikov proposed the notion of δ-disclosure privacy [7]. A table is called δ-disclosure private if the distribution of the SA values within each $\mathcal{QI}$ class is roughly the same compared with their distribution in the entire table. The δ-disclosure has the advantage to correctly model disclosures when some value of the SA occurs in certain $\mathcal{QI}$ classes, but not in others. ℓ-diversity, t-closeness and δ-disclosure are additional requirements that are based on k-anonymity. Therefore, they can be considered as extensions and not replacement of the original concept. Nevertheless, as it going to be shown, none of these extensions can prevent the attack that is illustrated in the next section.

TABLE III
K=3 LATTICE 0,0,2

Gender	Zip Code	Age	Disease
F	50100	*	Uterine Cancer
M	50100	*	Flu
F	50100	*	Mastitis
M	50100	*	Stomach cancer
M	50100	*	HIV
F	50100	*	Coronary heart disease
M	50120	*	Hepatitis
F	50120	*	Obesity
M	50120	*	Diabetes
M	50120	*	Prostate Cancer
F	50120	*	Alzheimer
F	50120	*	Juvenile idiopathic arthritis

III. INFERENCE OF QIS ATTACK

Medical datasets very often contain the disease of the record holder as their SA attribute. Definitely, all the aforementioned attacks in Section V are relevant, nevertheless, our attack follows a different information flow. While most of the attacks try to combine to $\mathcal{QI}$s to expose the SA, our attack goes the other way round. At this point, it has to be highlighted that due to the nature of the SA in medical records, diseases are very of often age or gender dependent, therefore, the value of the SA can expose the value of a generalized or suppressed $\mathcal{QI}$ of the anonymized dataset.

Taking advantage of the last remark, we introduce a new attack, called *inference of $\mathcal{QI}s$ attack*, which is based on the SA values and can break the given anonymity guarantees, such as k-anonymity. To clarify the attack, we should understand that the fields of most of the tables are going to be anonymized can be considered independent. For instance, a table might contain the following fields, gender, age and zip code. These three columns are independent in the sense that there is no logical constraint to assume that a man X years old leaves in $XVille$, or that a woman X' years old leaves in $XVille'$, unless this is a background knowledge. However, as already implied, this is not the case for medical records, something that is illustrated in the following examples.

A. Examples

From the original dataset in Table I, many anonymized tables can be produced. However, in what follows, we will assume that the needed anonymization requirements are 3-anonymity and 3-diversity. The generalization lattice in Figure 1 shows all possible combinations of the generalized domains. The six green nodes satisfy the required guarantees while the two blue do not. Since we want to maximize the data utility, we select the tables which have the least information loss, see Table IV, which in this case are two, Tables II and III. In our attack model, we assume that an adversary knows the $\mathcal{QI}s$, as they can be considered known from other databases.

TABLE IV
POSSIBLE GENERALIZED DOMAINS AND THE REGARDING INFORMATION LOSS. RESULTS PRODUCED USING FLASH [8]

Transf.	Anonymity	Min. Info. Loss	Max. Info. Loss
$[0,0,1]$	Not Anonymous	0.0 [0%]	43.0196 [64, 19%]
$[1,0,1]$	Anonymous	43.0196 [64, 19%]	43.0196 [64, 19%]
$[0,0,2]$	Anonymous	43.0196 [64, 19%]	43.0196 [64, 19%]
$[0,1,1]$	Not Anonymous	0.0 [0%]	67.0196 [100%]
$[1,1,1]$	Anonymous	43.0196 [64, 19%]	67.0196 [100%]
$[1,0,2]$	Anonymous	43.0196 [64, 19%]	67.0196 [100%]
$[0,1,2]$	Anonymous	43.0196 [64, 19%]	67.0196 [100%]
$[1,1,2]$	Anonymous	67.0196 [100%]	67.0196 [100%]

Example 1 Table V is an EC of the original dataset. The Gender $\mathcal{QI}$ has been suppressed in order to satisfy the 3-anonymity and the 3-diversity requirement. With a closer look on the table we notice that using the values of the SA Disease, an adversary can infer the value of the Gender $\mathcal{QI}$ for two out of three record holders. Uterine cancer and mastitis are female specific diseases, while stomach cancer is not gender specific. Due 3-anonymity, not all records can belong to females, therefore, the record with stomach cancer belongs to a male patient. The latter means that Dennis can be re-identified from the original table, since he is the only male that fits in this EC. Additionally, in the other two tuples the EC became from 3-anonymous to 2-anonymous, thus the possibility of a successful linking a tuple to a target has been increased from 33% to 50%.

TABLE V
EXAMPLE 1 LATTICE 1,0,1

Gender	Zip Code	Age	Disease
*	50100	14-34	Uterine Cancer
*	50100	14-34	Mastitis
*	50100	14-34	Stomach cancer

Example 2 Table VI is derived when the path (0,0,2) of the lattice is chosen for anonymization. Therefore, the field Age is suppressed to satisfy the anonymization requirements. In this example there are two age related diseases. In this context, the value "Juvenile idiopathic arthritis" obviously refers to a person less than 16 years old. From the latter, we can deduce that the record belongs to Lea, who is the only below 16. Moreover, the Alzheimer disease, most often, is diagnosed in people over 65 years of age. Therefore, one can deduce that the record belongs to Keith. This means that the final record belongs to Heather. Nevertheless, we should note here that "obesity" can not be linked to a specific age, but if the adversary knows anything about his target appearance, he could easily deduce whether this record is related to a possible candidate for his target or not.

TABLE VI
EXAMPLE 2 LATTICE 0,0,2

Gender	Zip Code	Age	Disease
F	50120	*	Obesity
F	50120	*	Alzheimer
F	50120	*	Juvenile idiopathic arthritis

B. Comparison with other attacks

One of the main differences of the proposed attack compared to others is the knowledge extraction from the SA value, which in the previous examples is the disease. By exploiting the knowledge derived from the SA value, we can infer some $\mathcal{QI}$s of the record holder, which will then lead to further exposure.

While one can claim that the attack is a disguised background knowledge attack, we argue that there are a lot of similarities, nevertheless, the information flow is from the SA values towards $\mathcal{QI}$s. The scope of the background knowledge attack, as stated throughout [3], is to use the demographics of $\mathcal{QI}$s to infer the SA value of a record. In contrast, our attack exploits the knowledge derived from the SA values to recover the respective $\mathcal{QI}$s, which will later be used to expose further information. So even though they are very similar, they operate in a completely different way.

One of the attacks that can be considered close to our approach, is the similarity attack. The attack is based on the similarity between the SA values inside an EC. However, this attack tries to group them and infer the SA of a victim related to a specific type of disease i.e uterine cancer and stomach cancer can be generalized as knowledge to "victim has cancer". The proposed attack though is trying to exploit the additional knowledge from the disease to infer for instance the $\mathcal{QI}$ gender when the latter is generalized to provide the privacy guarantees. It is quite straightforward that the uterine cancer can only be linked to records where the gender is Female.

It should be highlighted that the attack that is presented in this work bares no resemblance with inference control [9]. In that attack scenario, users execute multiple queries on the database to correlate their results, creating "inference channels" that will disclose other sensitive information. Typically, an adversary will split the query to several others that their intersection will recover a sensitive information. However, the proposed attack is focused on already published and anonymized tables; and the inference is initially made on record level by exposing hidden $\mathcal{QI}s$ based on the values of SA.

It becomes apparent that the proposed attack illustrates a novel way of re-identifying information by using the SA values to expose $\mathcal{QI}s$, reversing the wide adopted attack scenarios, where $\mathcal{QI}s$ are used to exposed SA values. However, the impact of the attack is related to nature of the underlying dataset. This means that the more dependent some $\mathcal{QI}$s are to the SA and the more times such values appear, the more data can be linked from the anonymized published table.

IV. SOLUTIONS

Since the attack is based on the semantics of the records, it is quite straight forward that out of the box solutions that simply generalize or suppress records cannot prevent the attack. An obvious way is having an expert to analyze the output set in order to see whether there are $\mathcal{QI}s$ that can be inferred from the anonymized table. While the latter looks quite tempting, depending on the size of the table, it is high likely that it is not practical. Therefore, there is a need for automated solutions. It is obvious that one cannot defend against arbitrary knowledge of the adversary, nevertheless, as it has been shown for the case of medical records, there are many correlations that can be bypassed.

In this context, we believe that a decision support system could be used to provide additional privacy protection. The system will have on the backend a database which contains all

the links between age, gender, ethnicity and diseases. Based on that, the system will parse the table before processing, marking the records that contain such data. These records will then have to be either suppressed, generalised on another field, or the linked values will have to be simultaneously generalised to a predefined value in the database. The decision support system, depending on the utility impact of each approach, will apply the according method. This way, the balance between utility and privacy will be at a good level for both data consumers and individuals.

V. Related Work

Apart from the Generalization and Suppression methods that where previously discussed, datasets can be anonymized with other techniques as well, some of which are listed below.

1) Anatomization: Anatomization [10] which is also known as *bucketization*, de-associates the relationship between the $\mathcal{QI}$ and the SA, without any modification on them. The method succeeds this by releasing the $\mathcal{QI}$ and the SA in separate tables having only a common attribute, the group ID. This way, all the records with the same group ID in the $\mathcal{QI}$ table are linked to all values in the SA table with that group ID. Compared to generalization approach the anatomized tables, because of the unchanged values, gives a more accurate answer to aggregation queries that involve $\mathcal{QI}$ values.

Our attack affects this technique because the groups that are formed can be further split into more groups i.e based on the age related diseases.

Example 3 In tables VII and VIII we give an example the anatomization of our original Table I. Anatomization has less information loss than k-anonimity since the $\mathcal{QI}$s are not generalized or suppressed. It offers the same guarantee as 3-anonymity, since any tuple from the group 1 has 33% probability of successful link to its corresponding SA value.

Using our attack we show that group 4 does not provide the 33% guarantee. Alzheimer and Juvenile idiopathic arthritis are age related diseases, therefore, an adversary could easily link the tuples (F,50120,88) and (F,50120,14) with 100% certainty in the last group. The one tuple that is left is obviously linked to obesity, therefore, for group 4 a complete re-identification could be performed.

TABLE VII
Anatomization - $\mathcal{QI}$ Table

GroupID	Gender	Zip Code	Age
1	F	50100	22
1	F	50100	33
1	F	50100	65
2	M	50100	45
2	M	50100	25
2	M	50100	56
3	M	50120	34
3	M	50120	54
3	M	50120	73
4	F	50120	18
4	F	50120	88
4	F	50120	14

TABLE VIII
Anatomization - Sensitive Table

GroupID	Disease (sensitive)	Count
1	Uterine Cancer	1
1	Mastitis	1
1	Coronary heart disease	1
2	Flu	1
2	Stomach cancer	1
2	HIV	1
3	Hepatitis	1
3	Diabetes	1
3	Prostate Cancer	1
4	Obesity	1
4	Alzheimer	1
4	Juvenile idiopathic arthritis	1

2) Permutation: Zhang et al. introduced the permutation method in [11]. This method partitions the data records into groups and afterwards starts shuffling the values of the SA inside the group. It is a method for numerical SA and improves the answers of aggregate queries on such SA. Since the SA of medical data is often categorical it is considered beyond the scope of this work.

3) Perturbation methods: Perturbation methods can be divided in three major categories:

Additive noise In this category we have methods which add noise to numerical SA. Thus, if we declare v_i the value of SA, these methods add a random number r_i to each v_i, following specific distributions; blinding this way the attacker the real SA value.

Data swapping Data swapping techniques [12] are used both for numerical and categorical SA. These methods anonymize the original table by exchanging the SA values among the records.

Synthetic data generation In this techniques, the publisher builds a mathematical model based on the original data and uses it to generate the anonymized table with synthetic records. This way, the published data retain the original features, nevertheless, they do not reflect real data.

In relation to our attack, *Additive noise* can be performed only in numerical SA and the *Synthetic data generation* generates a complete different dataset than the original. *Synthetic data generation* and *Data swapping* do not keep the truthfulness at the record level, which is a requirement for many applications. Moreover, *Data swapping*, unless it takes into consideration the partial $\mathcal{QI}$ dependencies its possible to generate tuples that are obvious not true such as (F, 50120, 18, Prostate Cancer), crippling the utility of the table.

4) Slicing: Anatomization can be considered as a special case of Slicing [13], where there are only two columns, one that contains all the $\mathcal{QI}s$ and the other only the SA. In slicing, columns can be formed with one or more $\mathcal{QI}s$, SA or both.

Example 4 In Table IX, Zipcode and Disease form one column, which helps the data recipient, in contrast to anatomization, to analyze better their correlations, as attribute correlations are considered an important utility in data publishing.

From the sliced table its easy to link the Juvenile idiopathic

arthritis with the only tuple that has the age to fit in the "juvenile" term and that is tuple (F,14). Alzheimer as an age related disease can be linked to Keith (F,50120,88), from the original Table I. The reason is that she is the only female old enough to have Alzheimer, with the zip code 50120. As in Example 3, the remaining tuple is linked to obesity since there is no other female with zip code 50120.

TABLE IX
SLICING - EXAMPLE

Gender & Age	Zip Code& Disease
(F , 14)	(50100 , Coronary heart disease)
(F , 18)	(50120 , Obesity)
(F , 22)	(50120 , Juvenile idiopathic arthritis)
(F , 33)	(50100 , Mastitis)
(F , 65)	(50120 , Alzheimer)
(F , 88)	(50100 , Uterine Cancer)
(M , 25)	(50100 , Flu)
(M , 34)	(50120 ,Prostate Cancer)
(M , 45)	(50120 , Diabetes)
(M , 54)	(50120 , Hepatitis)
(M , 56)	(50100 , HIV)
(M , 73)	(50100 , Stomach cancer)

For more information regarding anonymization methods dedicated to medical records, the interested reader may refer to [14], [15], [16], [17], [18].

VI. CONCLUSIONS

To allow mining medical records, the databases should undergo an anonymization process of their records. In the literature there are many well-known methods that enable this, providing different balance between privacy and utility of the data. Definitely, these methods are subject to what type of attacks from an adversary can be torrelated by the data owners. However, current state of the art is overlooking a very specific property of the $\mathcal{QI}s$, they might be partially dependent on the SA values. Contrary to most of the attacks that are presented in current literature, our attack follows a different path. It tries to exploit the knowledge derived from the SA values to deduce the $\mathcal{QI}s$. As it is shown in this work, focusing in the case of medical records where this phenomenon is quite often, an adversary may infer a lot of information and break anonymization guarantees, leading to re-identification of individuals. In this sense, this work introduces the concept of inference of $\mathcal{QI}s$ attack, highlights the existence of the problem in specific datasets, however, the article does not quantify its extent. As future work, we plan to experiment with anonymized synthetic and real datasets and quantify the real impact of the attack, while simultaneously try to develop a decision support system to throttle such attacks.

ACKNOWLEDGMENT

Agusti Solanas is partly funded by La Caixa Foundation through project "SIMPATIC: Intelligent, Autonomous and Private Monitoring System based on ICT" RECERCAIXA'12, and by the Government of Catalonia under grant 2009 SGR 1135. He is also supported by the Spanish Government through project CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES," and project TIN2011-27076-C03-01 "CO-PRIVACY".

REFERENCES

[1] 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.

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical report, SRI International, Tech. Rep., 1998.

[3] 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.

[4] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 106–115.

[5] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007, pp. 665–676.

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

[7] J. Brickell and V. Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 70–78.

[8] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn, "Flash: efficient, stable and optimal k-anonymity," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 708–717.

[9] C. Farkas and S. Jajodia, "The inference problem: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp. 6–11, 2002.

[10] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 2006, pp. 139–150.

[11] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 116–125.

[12] 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.

[13] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A new approach for privacy preserving data publishing," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 3, pp. 561–574, 2012.

[14] G. Loukides and A. Gkoulalas-Divanis, "Utility-aware anonymization of diagnosis codes," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 1, pp. 60–70, Jan 2013.

[15] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, "Anonymization of electronic medical records for validating genome-wide association studies," *Proceedings of the National Academy of Sciences*, vol. 107, no. 17, pp. 7898–7903, 2010.

[16] G. Szarvas, R. Farkas, and R. Busa-Fekete, "State-of-the-art anonymization of medical records using an iterative machine learning framework," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 574–580, 2007.

[17] P. Ruch, R. H. Baud, A.-M. Rassinoux, P. Bouillon, and G. Robert, "Medical document anonymization with a semantic lexicon." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000, p. 729.

[18] A. Solanas, A. Martínez-Ballesté, and J. M. Mateo-Sanz, "Distributed architecture with double-phase microaggregation for the private sharing of biomedical data in mobile health," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 901–910, 2013.