# Towards internal privacy and flexible K-anonymity

Hussein Hellani, Rima Kilany
Faculty of electrical engineering
Université Saint Joseph
Beirut, Lebanon
Hussein.hellani@hotmail.com, rima.kilany@usj.edu.lb

Maria Sokhn
Department of Computer Science
HES-SO Valais Wallis
Techopole 3, switzerland
Maria.sokhn@hevs.com

*Abstract*—**The focus on k-anonymity enhancements along the last decade, definitely allows this method to be elected as the point of start for any research. In this paper we propose « Contra-Loss », the first anonymization approach applied at internal level, to enhance local privacy where data is at rest and « Flexible k- anonymity » which aims to apply k-anonymity in most situations by defining a semantic ontology which distinguishes between scanty and abundant quasi-identifiers to achieve adequate k-blocks.**

## I. Introduction

Usually, any anonymization process passes through one stage only, which mainly addresses the people outside the organization while totally neglecting the data repository from being accessed by any employee or attacker that can reach this data and violate individuals' privacies. According to most of the statistics and surveys1, 70% of all security incidents come from insiders, whereof misleading the internal users is as important as the external ones and the anonymization at this level becomes an urgent need to limit the risk of internal threat.

We use *Comiqual* database[2], the internet quality measurement for mobile and ADSL users, to demonstrate the ability of apply our internal anonymization approach. Equally and based on k-anonymity [1], we propose Flexible k-anonymity approach, an enhanced anonymization method to be applied at the external level, by inferring aggregation levels from the ontology in order to be able to use different k-anonymity values and build appropriate k-blocks.

## II. Related Works

### A. Internal anonymization

In the middle tier, where data is generated and prepared to be transferred, a lot of miscellaneous attacks threatening data privacies can occur before anonymization. The main constraints in the case of internal data anonymization are: 1) we cannot remove the PII, since their existence is a must for authentication purposes. 2) Generalization in the meantime, isn't a good idea and should be applied only to the second use of data in order to guarantee high data precision and utility. The p-sensitivity approach [3] and [4] are among the few solutions that have been proposed. These approaches doesn't provide any kind of protection against operative intrusion.

### B. External Anonymization

Experiments show that constructing a k-block of multiple quasi-identifiers is not always crowned by success due to the nature of data that often contains some sparse values. This adds more challenges to the grouping of similar items during the anonymization. To enhance the anonymization methods, many researches such as p-sensitivity [3], vocabulary k-anonymity [5], ontology k-anonymity [2] and ontological semantics technology [6] are based on semantic ontology. The idea is to add an ontology layer on top of the k-anonymity method in order to create a more robust privacy-enforcing system [2]. Such methods do not apply well on measurement applications like Comiqual, because such measurement platforms do not store any sensitive attributes. The hardness of applying k-anonymity in a sparsely data environment and the eventual benefit of using semantics encouraged us to use a semantic ontology and change the core of the k-anonymity process.

## III. Internal Anonymization

Prior to applying the new internal anonymization method, it is mandatory to hash the personal identifier information (PII) of each record in order to avoid direct individual identification. PII will be encrypted in the second step to be stored simultaneously in a text file and a temporary database in the third step. The received records which are composed of data and secure PII, will reside in the temporary database for a period of time to increase the chance in achieving k-anonymity via accumulating the arrival records. The approach consists of adding fake records and fake PII to the data with the aim of deceiving the internal workers by k similar information. Thereby if records are similar enough, membership information is protected because the adversary cannot differentiate original tuples from fake ones. The combined data within internal database synchronize instantaneously with new entries in order to remove the fake records that exceed the k records, or remove all fake records for a specific k-block, in case the new arrivals with combination of the stored ones, satisfy the k-anonymity. In that way no need to add fake records. The golden rule is to prevent fake records number to exceed k-1 in any situation so as to avoid the unnecessary huge data. The main challenge of this approach is that combining fake and real records into one dataset, definitely leads to data inconsistency. To address such impairment, PII of each record that was hashed, encrypted and stored in a secured text file will be used to filter out the non-

[1] http://www.theguardian.com/uk
[2] Comiqual: Collaborative measurement of internet quality http://comiqual.usj.edu.lb/

fake records in order to be re-anonymized and published to external database. We call this method the “CONTRA-LOSS” because it protects the data from being suppressed and/or generalized.  In a word, contra-loss allows us to apply k-anonymity without using suppression or generalization while enforcing 0% data loss in the first use of data. e.g. for k=5, assume the system detects three iPhone 6 only, so it will clone two more records with random fake PII and store the 5-block internally at time t1. Suppose the system detects new similar entry “iPhone 6” at t2, it will remove one of the fake records and synchronize with internal dataset to limit the number of those fake records so as their number is always less than k-1.

## IV. External Anonymization

Our semantic model consists of two main hierarchies: the sparse and the normal attributes branches as displayed in the figure 1. In order to select k similar mobiles, we scroll up the sparse attributes hierarchy to infer the aggregation level from the ontology that will enable us to build the appropriate k-block. Keeping sparse attributes (e.g. “mobile model”) out of consideration during the construction of k-anonymity blocks can lead to eventual individual identification, while dealing with such attributes, in the same way as for non-sparse ones, could end up with the anonymization process failure. Flexible k-anonymity is based on splitting the quasi-identifiers attributes into sparse and normal classes towards having sparse data partially contribute in the creation of k-block, by k of sparse value “ks”, where ks < k. This method allows us to use different k values for the same dataset rather than fix a static one. The level of contribution depends mainly on the percentage of sparse data available within the dataset which is called sparse probability P(s) that indicate how much "sparse data" should contribute in k-block construction. P(s) represents the direct relation between k and ks:

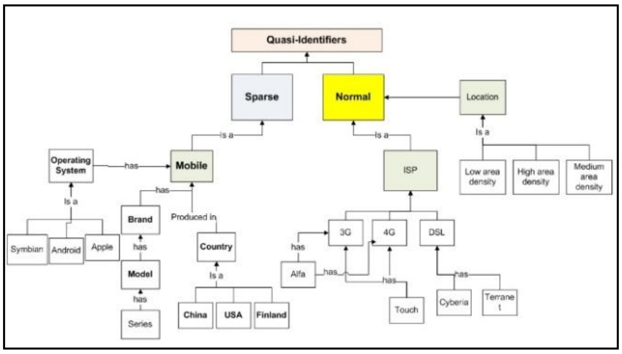

**Figure 1. Semantic Ontology model for Comiqual data**

a) $Ks = P(s) \times K$  b) $Ks: = \begin{cases} K \; for \; P(s) \ll \\ otherwise; Ks < K \end{cases}$

The number of the remaining tuples to be filled into the dataset will be inferred from ontology through the detection of an appropriate aggregation level by going up in the hierarchy of sparse attribute classes as we can see in Figure 1.

In flexible k-anonymity, defining the k value is done intuitively by estimating the occurrences of similar or approximate records of the non-sparse attribute whereas k of sparse “Ks” represents portion of this predefined k value, defined by sparse data proportion of the whole dataset.

## V. Sanitization Process

Contra-loss and flexible k-anonymity are two independents anonymization methods. The below demonstrates the integration process:

- Hash the PII of each entry to avoid direct identification.
- Encrypt the hashed PII (used for filter the true data)
- Encrypted PII will be stored in a secure text file.
- Data burst are reserved in a temporary database for t time.
- If records’ number is less than k, synchronize them with the internal database content (to avoid the growths of fake record size and achieve k-anonymity with minimum cost).
- Check k value after data synchronization, if it is satisfied, then data will be stored in the internal database
- If the records’ number is still less than k, contra-loss method will be applied thru adding fake records with fake PII -that are hashed and encrypted - until satisfying the k-anonymity (0% data loss).
- Fake & real records will be published to internal database and the temporary location will be cleared after t time.
- Filter the real records by means of true PII and preserve them in a new temporary database.
- Check k satisfaction, publish immediately to external database if records’ number satisfy k-condition, and the temporary location will be cleared after a t time.
- If records’ number is less than k, flexible k-anonymity will be applied to differentiate sparse and non-sparse attribute and enforce $K_S$ & K simultaneously based on probability of sparse P(s). First generalize normal and sparse attributes to satisfy k & $k_S$ then use ontology to fill the remaining records ($k-k_S$) based on most common criterion.
- Ensure K and $K_S$ satisfaction (using semantic ontology).
- Publish the anonymized data to the external website

## VI. Conclusion & Future Work

Accordingly, we can say that standard anonymization is not completely efficient in protecting individual privacy. In this paper we enlarge the scope of research to cover the internal stored data as well as that published to the external, by introducing contra-loss and flexible k-anonymity approaches.

As for future work, we would like to extend our study to investigate and improve the continuous flow of data as well as the behavior of new arrival entries.

## References

[1] L.Sweeney“k-anonymity: A model for protecting privacy.” International Journal of Uncertainty, Fuzziness 2002.

[2] E.Omran, “A K-anonymity Based Semantic Model for Protecting Personal Information and Privacy,” 2009 IEEE.

[3] Z. Xiao, X. Meng “p-Sensitivity: A Semantic Privacy-Protection Model for Location-based Services. MDMW 2008.

[4] Daubert, J.“Internal attacks in anonymous publish-subscribe P2P overlays.” In NetSys2015 International Conference.

[5] J., Liu, K.,Wang, “Enforcing Vocabulary k-Anonymity by Semantic Similarity Based Clustering, in Data Mining.” (ICDM), 2010.

[6] T., Ringenberg, J., Taylor, Semantic “*Anonymization of Medical Records*”. IEEE October 2014 San Diego, CA, USA.