Privacy preserving publishing of set-values

As technology infiltrates more and more aspects of our lives, each human activity leaves a digital trace in some repository. Vast amounts of personal data are implicitly or explicitly created each day, and rarely one is aware of the extent of information that is kept and processed about him or her. These personal data give rise to significant concerns about user privacy, since important and sensitive details about one's life are collected and exploited by third parties.

The goal of privacy preservation technologies is to provide tools that allow greater control over the dissemination of user data. A promising trend in the field is Privacy Preserving Data Publishing (PPDP), which allows sharing of anonymized data. Anonymizing a dataset is not limited to the removal of direct identifiers that might exist in a dataset, e.g. the name or the Social Security Number of a person. It also includes removing secondary information, e.g. like age, zipcode that might lead indirectly to the true identity of an individual. This secondary information is often referred to as *quasi-identifiers*.

Figure 1 - Abstraction of k-anonymization, k=3

A basic attack scenario against user privacy is as follows: An adversary who knows some characteristics of a person, e.g. age and zipcode, searches the anonymized data for records that match her background knowledge. If a single record is found, then she be can certain that the record refers to the person see knows. The sparser the data are, the more unique combinations exist, and the easier it is for an adversary to locate unique records that correspond to specific users. This makes collections of set-values, which are sparse multidimensional data, very hard to anonymize effectively without compromising user privacy [Aggarwal 2007]. There are several methods in research literature that address the challenges of anonymizing set-valued data. I report the basic techniques classified in three categories: anonymization methods that protect against identity disclosure, methods that protect against attribute disclosure and methods that offer differential privacy.

Protection against identity disclosure

Protection against identity disclosure guarantees that adversaries will not be able to associate specific records with a known individual. The most popular guaranty is *k*-anonymity [Samarati 2001, Sweeney 2002]. *k*-anonymity guarantees that each record will be indistinguishable from other *k*-1 records, with respect to the quasi identifiers. Every combination of quasi identifiers appears 0 or more than *k* times in the anonymized dataset.

Manolis Terrovitis, IMIS, RC Athena, 2011
mter@imis.athena.innovation.gr

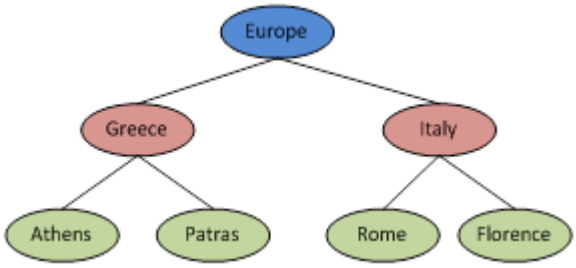

Figure 2 - A generalization hierarchy

In [He et. al. 2009] a method for transforming set-valued data to a *k*-anonymous form, the *Partition* algorithm, is proposed. *Partition* employs generalization for transforming the data to *k*-anonymous form. Generalization is the replacement of a group of original values by one new more abstract one. For example, if the residence area of an individual is reported in terms of cities, the city name can be replaced by the country name in anonymized data as in Figure 2. Partition employs local recoding; not all appearances of an original value are replaced by a generalized one. Partition is a top down algorithm; it starts by considering that all values are generalized to the more generic value of the generalization hierarchy and then drills down the hierarchy until the *k*-anonymity property no longer holds.

Despite using a flexible local recoding transformation, *Partition* has to significantly transform the original data in order to achieve *k*-anonymity. Records of different cardinalities that contain items from a large domain must end up to groups of identical records of size at least *k*. In response to this problem [Terrovitis et. al. 2008, Terrovitis et. al. 2010] propose k^m-anonymity. k^m-anonymity requires that each combination of up to *m* quasi identifiers must appear at least *k* times in the published data. The intuition behind k^m-anonymity is that there is little privacy gain from protecting against adversaries who already know most of the terms of one record, and significant information loss in the effort to do so. In [Terrovitis et. al. 2008, Terrovitis et. al. 2010] algorithms that rely on generalization using both global and local recoding are proposed.

	Meat	Wine	Oranges	Strawberries
Vassilis	X	X		
Manolis	X	X	X	
Eleni			X	
Maria		X	X	
Kostas	X			X

Table 1 - Original data

	Food	Wine	Fruits
Vassilis	X	X	
Manolis	X	X	X
Eleni			X
Maria		X	X
Kostas	X		X

Table 2 – 2-anonymous data

	Food	Wine
Vassilis	X	X
Manolis	X	X
Eleni	X	
Maria	X	X
Kostas	X	

Table 3 – 2^2-anonymous data

In [Loukides et. al. 2011, Gkoulalas et. al. 2011] the authors provide an even more targeted approach. The paper assumes that a set of privacy and utility constraints is known by the data publisher. Such assumption holds in specific domains, like that of medical records [Loukides et. al. 2010/II]. Exploiting this explicit knowledge about privacy dangers and utility the authors propose algorithms that achieve protection against identity disclosure with limited information loss.

Protection against attribute disclosure

The data values in a dataset are not usually equally important as personal information. A common distinction in privacy related literature is between quasi identifiers and sensitive values. Quasi identifiers are usually known through several sources and they do not threaten the privacy of an individual. Sensitive values on the other hand, are not considered available through other sources and they reveal important personal information. If such a distinction holds and it is known by the data publisher, then data must also be protected against the disclosure of sensitive attributes. A common guaranty for protecting against sensitive values is *l*-diversity [Machanavajjhala et. al. 2006]. *l*-diversity guarantees that any adversary cannot associate

her background knowledge with less than *l well represented sensitive values.* Well-represented is usually defined as a probability threshold: an adversary cannot associate her background knowledge with any sensitive value with probability over 1/*l*.

	Meat	Wine	Oranges	Strawberries	Pregnancy test	Viagra
Vassilis	X	X				
Manolis	X	X	X		X	
Eleni			X			X
Maria		X	X			
Kostas	X			X		

Table 4 - Original data. Sensitive values are depicted with different color on the right

	Meat	Wine	Oranges	Strawberries	Sensitive values
Vassilis	X	X			Pregnancy test:1 Viagra:1
Manolis	X	X	X		
Eleni			X		
Maria		X	X		
Kostas	X			X	

Table 5 – Anonymized data

The first anonymization method that provided protection against attribute disclosure in set-valued attributes was proposed in [Ghinita et. al. 2008]. The proposal of [Ghinita et. al. 2008, Ghinita et. al. 2011] relies on separating sensitive values from quasi identifiers as depicted in Tables 4 and 5. The idea of separation was first proposed in the relational context in [Xiao et. al 2006], but it was adjusted and extended in [Ghinita et. al. 2008, Ghinita et. al. 2011] for the set-valued context. The basic idea of proposed the anonymization method is to create clusters of similar records (with respect to quasi identifiers) and then publish at each cluster the quasi identifiers and the sensitive values separately. A benefit of this transformation with respect to generalization and suppression is that it does not require creating groups with identical quasi identifiers. This way the information loss is kept low, even for data of very high cardinality and dimensionality.

Protection against attribute disclosure is also studied in [Cao et. al. 2010]. The proposed guaranty, ρ-uncertainty is similar to as *l*-diversity and requires that quasi identifiers cannot be associated with sensitive values with probability over $1/\rho$. The novelty of the approach is that is considers as quasi identifier *every record subset* that appears in the dataset, including the ones that contain sensitive values. This means that sensitive values can also act as quasi identifiers. The proposed anonymization method relies both on generalization and suppression.

A guaranty that provides protection both from identity and attribute disclosure, the (h,k,p)-coherence, is proposed in [Xu et. al. 2008/I, Xu et. al. 2008/II]. (h,k,p)-coherence. Similarly to k^m-anonymity it protects from adversaries how know up to p terms, by guaranteeing that every combination of items will appear at least k times. Moreover, (h,k,p)-coherence guarantees that combinations of up to p items cannot be associated with a sensitive value with probability over h. The proposed anonymization method relies solely on suppression.

In [Loukides et. al. 2010/I] an anonymization method that can be tailored to specific sensitive inferences is presented. The authors propose the notion of PS-rules, which are sensitive association rules specified by the data owner. The anonymization procedure guarantees that an adversary will not be able to infer these rules with high certainty. The proposed anonymization algorithms are based on generalization.

Differential privacy in for set-valued attributes

Differential privacy [Dwork et. al. 2006] is a more generic guaranty than *k*-anonymity and *l*-diversity and the protection it offers is broader than protection against identity and attribute disclosure. It does not make any

assumptions about the adversary's background knowledge (although anonymization methods that provide differential privacy might make some implicit ones [Kifer et. al. 2011]), and it requires that any analysis of the anonymized data is not significantly affected by the addition or removal of one record in the original data. The trade-off for providing such a strong anonymity guaranty is that the information quality of the result if significantly affected. Only the results of specific analysis can be provided, by adding noise in a non-deterministic procedure.

A first approach to the anonymization of set-values by using differential privacy appears in [Korolova et. al. 2009], in the context of web search query logs. Each record is the web search history of a single user, which can be modeled as a set value, but the result of the anonymization is not set-values. Only the frequent query terms and they support (with the addition of Laplace noise) is published. A more recent approach that publishes anonymized set values under differential privacy and provides set-values as a result, appears in [Chen 2011]. The paper proposes *DiffPart*, a top-down algorithm that decides which frequent itemsets can be published, without violating differential privacy. The supports of all published items are distorted by adding Laplace noise.

References

[Aggarwal 2007]	Charu C. Aggarwal. On Randomization, Public Information and the Curse of Dimensionality. In Proceedings of ICDE'2007.
[Cao et. al. 2010]	Cao, P. Karras, C. Raïssi, and K.-L. Tan. *ρ-uncertainty: Inference-Proof Transaction Anonymization*. PVLDB 2010.
[Chen 2010]	Rui Chen, Noman Mohammed, Benjamin C. M. Fung, Bipin C. Desai, and Li Xiong. Publishing Set-Valued Data via Differential Privacy. PVLDB 4(11), 2011.
[Dwork et. al. 2006]	C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference, pages 265–284, 2006.
[Ghinita et. al. 2008]	Ghinita G., Tao Y., Kalnis P., *On the Anonymization of Sparse High-Dimensional Data*, ICDE, 2008.
[Ghinita et. al. 2011]	Ghinita G., Kalnis P., Tao Y., "Anonymous Publication of Sensitive Transactional Data", IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), 23(2), 161-174, 2011.
[Gkoulalas et. al. 2011]	PCTA: Privacy-constrained Clustering-based Transaction Data Anonymization. A. Gkoulalas-Divanis, G. Loukides, 4th International Workshop on Privacy and Anonymity in the Information Society, ACM, 2011.
[He et. al. 2009]	Y. He and J. F. Naughton. Anonymization of set-valued data via top-down, local generalization. In VLDB, pages 934–945, 2009.
[Kifer et. al. 2011]	Daniel Kifer, Ashwin Machanavajjhala. No free lunch in data privacy. SIGMOD Conference'2011.
[Korolova et. al. 2009]	Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, Alexandros Ntoulas. Releasing search queries and clicks privately. WWW'2009.
[Loukides et. al. 2010/I]	Grigorios Loukides, Aris Gkoulalas-Divanis, Jianhua Shao: Anonymizing Transaction Data to Eliminate Sensitive Inferences. DEXA (1), 2010.
[Loukides et. al. 2010/II]	Anonymization of Electronic Medical Records for Validating Genome-Wide Association Studies. G Loukides, A Gkoulalas-Divanis, B Malin, Proceedings of the National Academy of Sciences, National Acad Sciences, 2010.
[Loukides et. al. 2011]	Grigorios Loukides, Aris Gkoulalas-Divanis, Bradley Malin: COAT: COnstraint-based anonymization of transactions. Knowl. Inf. Syst. 28(2): 251-282 (2011)

Manolis Terrovitis, IMIS, RC Athena, 2011
mter@imis.athena.innovation.gr

[Machanavajjhala et. al. 2006]	Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M. l-Diversity: Privacy Beyond k-Anonymity. ICDE, 2006.
[Samarati 2001]	P. Samarati. *Protecting Respondents' Identities in Microdata Release*. IEEE TKDE, 13(6), 2001.
[Sweeney 2002]	L. Sweeney. *k-Anonymity: A Model for Protecting Privacy*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.
[Terrovitis et. al. 2008]	M. Terrovitis, N. Mamoulis and P. Kalnis. *Privacy-preserving Anonymization of Set-valued Data*. PVLDB, 1(1): 115-125, 2008.
[Terrovitis et. al. 2010]	M. Terrovitis, N. Mamoulis and P. Kalnis. *Local and Global Recoding Methods for Anonymizing Set-valued Data*. The VLDB Journal, (to appear), 2010.
[Xu et. al. 2008/I]	Yabo Xu, Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, Jian Pei: Publishing Sensitive Transactions for Itemset Utility. ICDM 2008: 1109-1114
[Xu et. al. 2008/II]	Yabo Xu, Ke Wang, Ada Wai-Chee Fu, Philip S. Yu: Anonymizing transaction databases for publication. KDD 2008: 767-775
[Xiao et. al 2006]	X. Xiao and Y. Tao, “Anatomy: Simple and Effective Privacy Preservation,” in Proc. of VLDB, 2006, pp. 139–150.