UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>

Publisher
UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by Novática <http://www.ati.es/novatica/>, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*, <http://www.ati.es/>)

UPGRADE monographs are also published in Spanish (full version printed; summary, abstracts and some articles online) by Novática

UPGRADE was created in October 2000 by CEPIS and was first published by Novática and INFORMATIK/INFORMATIQUE, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>)

UPGRADE is the anchor point for UPENET (UPGRADE European NETwork), the network of CEPIS member societies' publications, that currently includes the following ones:
- InfoRewiew, magazine from the Serbian CEPIS society JISA
- Informatica, journal from the Slovenian CEPIS society SDI
- Informatik-Spektrum, journal published by Springer Verlag on behalf of the CEPIS societies GI, Germany, and SI, Switzerland
- ITNOW, magazine published by Oxford University Press on behalf of the British CEPIS society BCS
- Mondo Digitale, digital journal from the Italian CEPIS society AICA
- Novática, journal from the Spanish CEPIS society ATI
- OCG Journal, journal from the Austrian CEPIS society OCG
- Pliroforiki, journal from the Cyprus CEPIS society CCS
- Tölvumál, journal from the Icelandic CEPIS society ISIP

Editorial correspondence: Llorenç Pagés-Casas <pages@ati.es>
Advertising correspondence: <novatica@ati.es>

UPGRADE Newslist available at
<http://www.upgrade-cepis.org/pages/editinfo.html#newslist>

ISSN 1684-5285

Monograph of next issue (April 2010)

"Information Technology inTourism Industry"

(The full schedule of **UP**GRADE is available at our website)

CEPIS UPGRADE
The European Journal for the Informatics Professional
http://www.upgrade-cepis.org

Vol. XI, issue No. 1, February 2010

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>.

How Do we Measure Privacy?

David Rebollo-Monedero and Jordi Forné

We survey the state of the art on the metrics of privacy in perturbative methods for statistical disclosure control. While the focus is on data microaggregation, these methods also address a wide variety of alternative applications such as obfuscation in location-based services. More specifically, we examine k -anonymity and some of its enhancements. Motivated by the vulnerability of these measures to similarity and skewness attacks, we compare three recent criteria for privacy based on information-theoretic concepts that attempt to circumvent this vulnerability.

Keywords: k -Anonymity, t -Closeness, δ -Disclosure, l -Diversity, Information Privacy, Information Theory, Microdata Anonymization, Statistical Disclosure Control,.

Authors

David Rebollo-Monedero received his MSc and PhD degrees in electrical engineering from Stanford University, USA, in 2003 and 2007. Previously, from 1997 to 2000, he was an information technology consultant for PricewaterhouseCoopers, in Spain. He is currently carrying out postdoctoral research on privacy in information systems with the Information Security Group of the *Universitat Politècnica de Catalunya* (UPC), also in Spain. <david.rebollo@entel.upc.edu>

Jordi Forné received his MSc and PhD degrees in telecommunications engineering from the *Universitat Politècnica de Catalunya* (UPC), Spain, in 1992 and 1997. Since 1991, he has been a member of the Information Security Group in the Department of Telematics of this university. Currently, he works as an associate professor. His research interests span privacy, network security, e-commerce and public-key infrastructures. <jforne@entel.upc.edu>

1 Introduction

The right to privacy was recognized as early as 1948 by the United Nations in the Universal Declaration of Human Rights, Article 12. With the exponentially accelerating growth of information technologies, and the trend towards the acquisition of a virtual identity on the Internet by nearly every person, object or entity, privacy will undeniably become increasingly crucial. With this in mind, we wish to design services where user privacy is properly protected. Naturally, we also wish to assess the weaknesses of those services against privacy attacks in an objective, systematic, scientific fashion. But to turn reality into science, we must cross the bridge between the qualifiable and the quantifiable. Thus, the question is inevitable: how do we measure privacy?

The purpose of this paper is precisely to survey the state of the art on metrics of privacy in perturbative methods for statistical disclosure control. These methods consist of perturbing user data in an optimal manner to maximize privacy, while preserving data utility to an acceptable degree. To this end, powerful concepts and techniques from statistics and information theory, among other fields, are exploited.

While the focus is on data microaggregation, these perturbative methods for privacy are applicable to a wide variety of alternative scenarios, such as obfuscation in location-based services, Internet search and P2P networks. More specifically, we briefly examine k -anonymity and some of its enhancements. Motivated by the vulnerability of these measures to similarity and skewness attacks, we compare three recent criteria for privacy based on information-theoretic concepts that attempt to circumvent this vulnerability. Namely, we compare the average privacy risk proposed in [1], t -closeness [2] and δ -disclosure [3].

As we have already stated, there is an inherent trade-off between privacy and data utility in any perturbative method for privacy. We would like to remark that, naturally, the complete specification of the optimization problem contemplating this trade-off would also require the specification of a data utility metric. In addition, solving the optimization problem might be far from trivial. In the interest of length and focus, however, we narrow the scope of this survey to privacy metrics.

This rest of this survey is organized as follows. Section 2 describes two application scenarios. Section 3 reviews the state of the art in privacy metrics for SDC. A more in-depth analysis of three of the information-theoretic criteria for measuring privacy is provided in Section 4. Conclusions are drawn in Section 5.

2 Application Scenarios

This section motivates the importance of controlling the disclosure of information with regard to privacy by introducing two related problems, namely microdata anonymization and the private retrieval of location-based information.

2.1 Microdata Anonymization

A *microdata set* is a database table whose records carry information concerning individual respondents, either people or companies. This set commonly contains *key attributes* or *quasi-identifiers*, namely attributes that, in combination, may be linked with external information to re-identify the respondents to whom the records in the microdata set refer. Examples include job, address, age, gender, height and weight. Additionally, the data set contains *confidential attributes* with sensitive information on the respondent, such as salary, religion, political affiliation or health condition. The classification of attributes as key or confidential may ultimately rely on the specific application and the privacy requirements the microdata set is intended for.

Intuitively, perturbation of key attributes enables us to preserve *privacy* to a certain extent at the cost of losing some of the *data utility* with respect to the unperturbed version. k *-Anonymity* is the requirement that each tuple of key records in the data set. This may be achieved through the *microaggregation* approach illustrated by the example shown in Figure 1, where height and weight are regarded as key attributes, and the blood concentration of (low-density lipoprotein) cholesterol as a confidential attribute. Rather than making the original table available, we publish a k -anonymous version containing aggregated records, in the sense that all key attribute values within each group are replaced by a common representative tuple. Despite the fact that k -anonymity as a measure of privacy is not without shortcomings, its simplicity make it a widely popular criterion in *statistical disclosure control* (SDC) literature.

2.2 Privacy in Location-Based Services

The problem of microdata anonymization we have motivated arises, at least conceptually, in a wide range of apparently different applications. An example of particular relevance is *location-based services* (LBSs). The simplest form of interaction between a user and an LBS provider involves a direct message from the former to the latter including a query and the location to which the query refers. An example would be the query "Where is the nearest bank to my home address?", accompanied by the geographic coordinates or simply the address of the user's residence. Under the assumption that the communication system used allows the LBS provider to recognize the user ID, there exists a patent privacy risk. Namely, the provider could profile users according to their locations, the contents of their queries and their activity.

Essentially, a perturbative method analogous to data microaggregation may be used to tackle this privacy risk, as represented in Figure 2. In general, users may contact an untrusted LBS provider directly, perturbing their location information in order to hinder providers in their efforts to compromise user privacy in terms of location, although clearly not in terms of query contents and activity. This approach, sometimes referred to as *obfuscation*, presents the inherent trade-off between data utility and privacy common to any perturbative privacy method. The parallel with microdata anonymization can now be drawn simply by identifying IDs and location information with confidential and key attributes, respectively.

3 k -Anonymity and Some of its Enhancements as Measures of Privacy in Statistical Disclosure Control

We mentioned in Section 2.1 that a specific piece of data on a particular group of respondents is said to satisfy

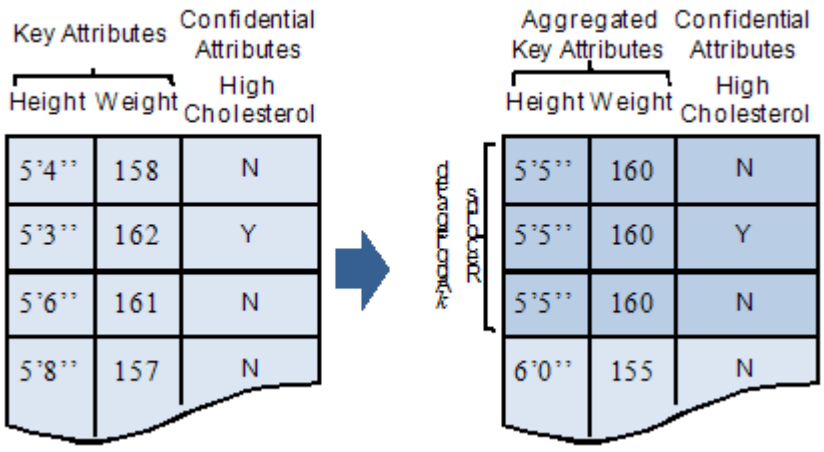

Figure 1: Microaggregation of Values of Key Attributes to Attain k -Anonymity.

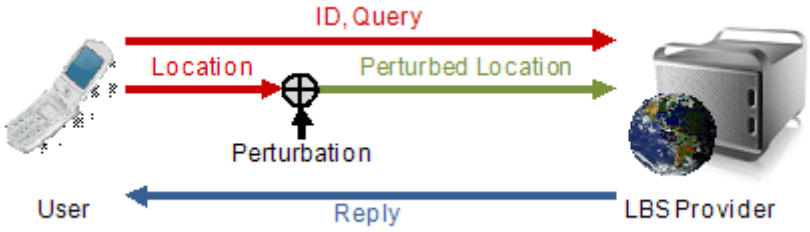

Figure 2. Users May Contact an Untrusted LBS Provider Directly, Perturbing their Location Information to Help Protect Their Privacy.

the k-anonymity requirement (for some positive integer k) if the origin of any of its components cannot be ascertained beyond a subgroup of at least k individuals. We also said that the concept of k-anonymity, originally proposed by the SDC community [4], is a widely popular privacy criterion, partly due to its mathematical tractability.

The original formulation of this privacy criterion, based on generalization and recording of key attributes, was modified into the microaggregation-based approach already commented on, and illustrated in Figure 1, in [5]. Both formulations may be regarded as special cases of a more general one utilizing an abstract *distortion* measure between the unperturbed and the perturbed data, possibly taking on values in rather different alphabets.

Multivariate microaggregation was proved to be NP-hard. A number of heuristic methods have been proposed, which can be categorized into fixed-size and variable-size methods, according to whether all groups but one have exactly k elements with common perturbed key attributes. The maximum distance (MD) algorithm and its less computationally demanding variation, the maximum distance to average vector (MDAV) algorithm [6], are fixed-size algorithms that perform particularly well in terms of the distortion they introduce, for many data distributions. The probability-constrained Lloyd (PCL) algorithm [7] is a recently proposed heuristic that extends the Lloyd-Max algorithm, a celebrated data compression algorithm that often produces optimal clusters.

Unfortunately, while k-anonymity prevents identity disclosure, it may fail to protect against attribute disclosure. Precisely, the definition of this privacy criterion establishes that complete re-identification is unfeasible within a group of records sharing the same tuple of perturbed key attribute values. However, if the records in the group also share a common value of a confidential attribute, the association between an individual linkable to the group of perturbed key attributes and the corresponding confidential attribute remains disclosed, as the example in Figure 3 illustrates. More generally, the main issue with k-anonymity as a privacy criterion is its vulnerability against the exploitation of the difference between the prior distribution of confidential data in the entire population, and the posterior conditional distribution of a group given the observed, perturbed key attributes. For example, imagine that in Figure 1 the proportion of respondents with high cholesterol is much higher than that in the overall data set. This is known as a *skewness attack*.

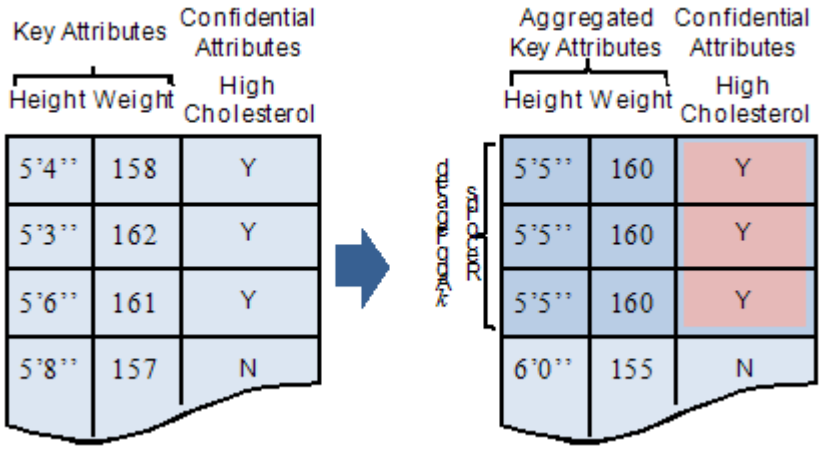

Figure 3. k-Anonymity of Key Attributes Does not Necessarily Guarantee Confidentiality.

This vulnerability motivated the proposal of enhanced privacy criteria, some of which we proceed to sketch briefly, along with algorithm modifications. A restriction of k-anonymity called p-*sensitive* k-anonymity was presented in [8]. In addition to the k-anonymity requirement, it is required that there be at least p different values for each confidential attribute within the group of records sharing the same tuple of perturbed key attribute values. Clearly, large values of p may lead to a huge data utility loss. A slight generalization called l-*diversity* [9] was defined with the same purpose of enhancing k-anonymity. The difference with respect to p-sensitivity is that group of records must contain at least l "well-represented" values for each confidential attribute. Depending on the definition of well-represented, l-diversity can reduce to p-sensitive k-anonymity or be more restrictive. We would like to stress that neither of these enhancements succeeds in completely removing the vulnerability of k-anonymity to skewness attacks. Furthermore, both are still susceptible to *similarity attacks*, in the sense that while confidential attribute values within a cluster of aggregated records might be p-sensitive or l-diverse, they might also very well be semantically similar, for example similar diseases or salaries.

A privacy criterion aimed at overcoming similarity and skewness attacks is t-*closeness* [2]. A perturbed microdata set satisfies t-closeness if for each group sharing a common tuple of perturbed key attribute values, some distance between the posterior distribution of the confidential attributes in the group and the prior distribution of the overall population does not exceed a threshold t. To the extent to which the within-group distribution of confidential attributes resembles the distribution of those attributes for the entire dataset, skewness attacks will be thwarted. In addition, since the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset.

The main limitation of the original t-closeness work [2] is that no computational procedure to reach t-closeness was specified. An information-theoretic privacy criterion, inspired by t-closeness, was proposed in [1]. In the latter work, privacy risk is defined as an information-theoretic measure of discrepancy between the posterior and the prior distributions. Conceptually, the privacy risk defined may be regarded as an averaged version of the t-closeness requirement, over all aggregated groups. It is important to notice as well that the criterion for privacy risk in [1], in spite of its convenient mathematical tractability, as any criterion based on averages, may not be adequate in all applications. A related albeit more conservative criterion, named δ-disclosure privacy, is proposed in [3], and measures the maximum difference between the prior and the posterior distributions. The average privacy risk of [1], t-closeness and δ-disclosure are discussed further in Section 4.

Regarding the parallelism with LBSs we drew in Section 2.1, we would like to remark that a wide variety of perturbation methods for private retrieval of location-based information has been proposed [10]. Not surprisingly, some employ the k-anonymity criterion as a measure of privacy. An illustrative example is that of [11]. Fundamentally, k users add zero-mean random noise to their locations and share the result to compute the average, which constitutes a shared perturbed location sent to the LBS provider. Unfortunately, some of these users may apply noise cancelation to attempt to disclose a slow-changing user's location. A location anonymizer that clusters exact locations to provide k-anonymity in LBSs using PCL is proposed in [7].

4 Information-theoretic Privacy Metrics

The following is a more in-depth discussion on some of the most recently proposed privacy metrics, based on information-theoretic concepts, which attempt to address the vulnerabilities of k-anonymity and its enhancements. Even though the metrics are new, as is the corresponding mathematical formulation of the microdata anonymization problem in terms of these metrics along with their solutions, we shall see that the metrics themselves are strongly related to concepts already proposed by Shannon in the fifties.

Our discussion will be fairly conceptual. Hence, it should suffice to recall that the *entropy* of a random variable (r.v.) is a measure of its uncertainty, that the *mutual information* between two r.v.'s is a measure of the information that one contains about the other, and that the Kullback-Leibler (KL) *divergence* is a measure of discrepancy between probability distributions. Readers interested in the mathematical definition of these information-theoretic quantities are encouraged to consult [12].

4.1 Privacy Risk, Shannon's Equivocation and Information Gain

In the problem of microdata anonymization introduced in Section 2.1 we model (tuples of) confidential attributes by a r.v. W, with probability distribution p_W. (Tuples of) key attributes are represented by a r.v. X, and are perturbed somehow to produce slightly modified (tuples of) data $\widehat{X}$. Rather than making the table, or more generally speaking, the probability distribution containing X and W available, the *sanitized* version with $\widehat{X}$ and W is published instead. Because

tables may be regarded as a specification of an empirical probability distribution, our model is slightly more general. Remember that our objective is to hinder attackers in their efforts to link the respondents' identity with their confidential data.

Consider now, on the one hand, the prior distribution of the confidential attributes W, and on the other, the posterior or conditional distribution of W given the perturbed attributes $\widehat{X}$. Whenever the posterior distribution differs from the prior distribution, we have actually gained some information about individuals statistically linked to the perturbed key attributes, in contrast to the statistics of the general population. In terms of the example illustrated in Figure 1, the probability of high cholesterol of the population might be, say, 25%, whereas the probability of high cholesterol for the group corresponding to a quantized height of 5 feet 5 inches and a quantized weight of 160 pounds is approximately 33%. Intuitively speaking, an individual of known height and weight falling into this category is more likely to have high cholesterol than one could have guessed merely from the entire population distribution. We recognize this situation as a statistical privacy risk, although not as severe as that illustrated by Figure 3.

In order to quantify the previous intuition, we first recall the concept of *equivocation* introduced by Shannon in 1949 [13], namely the conditional entropy of a private message given an observed cryptogram. The application of the principle of Shannon's equivocation to privacy is by no means new. For example, in [4], the degree of anonymity observable by an attacker is measured as the entropy of the probability distribution of possible senders of a given message. Conceptually, and slightly more generally, we shall regard Shannon's equivocation as the entropy of the private, unobserved information, given the public, observed information.

In terms of our formulation, we concordantly compare the entropy of W associated with the prior distribution of the confidential attributes, with the equivocation, that is, the entropy of W given $\widehat{X}$, associated with the posterior distribution given the observed perturbed key attributes. The reduction in uncertainty, that is, the entropy difference, is taken directly as a measure of privacy risk in our work [1]. Moreover, this work shows that this entropy reduction is precisely the mutual information between W and $\widehat{X}$, which in turn matches the KL conditional divergence

$$\left(D(p_{(w|\widehat{x})} \,\|\, p_W\right)$$

between the prior and the posterior distributions.

Remember that a conditional divergence is a divergence between conditional distributions of the conditioned r.v., averaged over the conditioning r.v. In the simpler case of *deterministic microaggregation*, where a value of X is assigned to a single value of $\widehat{X}$, conceptually speaking the privacy risk defined is an average between the discrepancies of the posterior distributions for each group of records sharing a common value of $\widehat{X}$ with respect to the prior distribution.

According to the properties of mutual information and KL divergence, the privacy risk defined is nonnegative, and vanishes if and only if W and $\widehat{X}$ are statistically independent, or equivalently, if the prior and posterior distributions match. Of course, in this extreme case, the utility of the published data would be severely compromised. In the other extreme, leaving the original data undistorted in general compromises privacy, because in general the prior and posterior distributions differ.

We can also trace back to the fifties the information-theoretic interpretation of the divergence between a prior and a posterior distribution, named *(average) information gain* in some statistical fields [15]. In addition to the work already cited, others already used Shannon entropy as a measure of information loss, pointing out limitations affecting specific applications. We would like to stress that we have introduced a KL divergence as a measure of *information disclosure* (rather than loss) consistently with the equivalence between the case when prior and posterior distributions match, and the complete absence of privacy risk.

Perhaps the most interesting property of the privacy criterion of [1] is that it leads to a mathematical formulation of the privacy-utility trade-off that generalizes a well-known, extensively studied information-theoretic problem with half a century of maturity. Namely, the problem of lossy compression of source data with a distortion criterion, first proposed by Shannon in 1959 [12].

4.2 t-Closeness and δ-Disclosure

We mentioned in Section 1 that the privacy criterion of [1] is a conditional divergence, where the conditioned r.v. is W and the conditioning r.v., $\widehat{X}$. In the deterministic microaggregation case, a conditional divergence is a between-group average of within-group divergences, where groups share a common value of $\widehat{x}$. By the definition of KL divergence, it turns out that the within-group divergences are themselves averages of log-ratios between probability values.

This privacy measure is tightly related to the measure of t-closeness of [2]. In terms of the formulation introduced in Section 1 and for the simpler case of discrete dis-

tributions and deterministic clustering, t-closeness may be defined as the between-group maximum among the within-group divergences, themselves averages of log ratios of probabilities. A related, more conservative criterion, named δ-disclosure privacy, is proposed in [3], and measures the maximum difference between the prior and the posterior distributions for each group sharing a common $\hat{x}$.

Simply put, the privacy risk measure in [1], reviewed in Section 1, is a between-group average of within-group averages (thus an average), t-closeness is a between-group maximum of within-group averages, and δ-disclosure is a between-group maximum of within-group maxima (thus a maximum). Hence, these measures range from modelling the average-case to the worst-case scenario.

5 Conclusion

In conclusion, we have motivated the importance of perturbative methods for privacy in microdata anonymization and also obfuscation for LBSs. Regarding the privacy criteria reviewed, we would like to emphasize that despite the shortcomings of k-anonymity and its enhancements as a measure of privacy, it is still a widely popular criterion for SDC, mainly because of its simplicity and its theoretical interest. Nevertheless, due to the vulnerability of k-anonymity and its enhancements to similarity and skewness attacks, privacy metrics based on information-theoretic concepts have been proposed recently.

Concordantly, we have examined three related information-theoretic measures of privacy, namely the average privacy risk of [1], t-closeness and δ-disclosure. First, it is only fair to stress that average-case optimization may not address worst cases properly. In other words, we acknowledge that the average privacy criterion, as any criterion based on averages, may not be adequate in all applications. However, the price of worst-case optimization is, in general, a poorer average, *ceteris paribus*. On the other hand, the work cited shows that the main advantages of the average privacy criterion of [1] are its mathematical tractability, and the fact that it leads to a mathematical formulation of the privacy-utility trade-off that generalizes the problem of lossy compression of source data with a distortion criterion, first proposed by Shannon in 1959 [12].

More generally, we acknowledge that the formulation of any privacy-utility problem relies on the appropriateness of the criteria optimized, which in turn depends on the specific application, on the statistics of the data, on the degree of data utility we are willing to compromise and, last but not least, on the adversarial model and the mechanisms against privacy contemplated. No privacy criterion presents itself as the be-all and end-all of database anonymization [3].

Acknowledgement

This work was partly supported by the Spanish Government through projects CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", TSI2007-65393-C02-02 "ITACA" and TSI2007-65406-C03-01 "E-AEGIS", and by the Government of Catalonia under grant 2009 SGR 1362.

References

[1] D. Rebollo-Monedero, J. Forné, J. Domingo-Ferrer. "From t-closeness-like privacy to postrandomization via information theory", IEEE Trans. Knowl. Data Eng., 2009. <http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190>.

[2] N. Li, T. Li, S. Venkatasubramanian. "t-Closeness: Privacy beyond k-anonymity and l-diversity". Proc. IEEE Int. Conf. Data Eng. (ICDE), Istambul, Turkey, April 2007, pp. 106-115.

[3] J. Brickell, V. Shmatikov. "The cost of privacy: Destruction of data-mining utility in anonymized data publishing". Proc. ACM SIGKDD Int. Conf. Knowl. Disc., Data Min. (KDD), Las Vegas, USA, August 2008.

[4] P. Samarati. "Protecting respondents' identities in microdata release". IEEE Trans. Knowl. Data Eng., vol. 13, núm. 6, pp. 1010-1027, 2001.

[5] D. Defays, P. Nanopoulos. "Panels of enterprises and confidentiality: The small aggregates method,". Proc. Symp. Design, Anal. Longitudinal Surveys, Stat. Canada, Ottawa, Canada, 1993, pp. 195-204.

[6] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz, F. Sebé. "Efficient multivariate data-oriented microaggregation". VLDB J., vol. 15, núm. 4, pp. 355-369, 2006.

[7] D. Rebollo-Monedero, J. Forné, M. Soriano. "Private location-based information retrieval via k-anonymous clustering,". Proc. CNIT Tyrrhenian Int. Workshop Digital Commun., Pula, Sardinia, Italy, September 2-4, 2009.

[8] T. M. Truta, B. Vinay. "Privacy protection: p-sensitive k-anonymity property". Proc. Int. Workshop Privacy Data Manage. (PDM), Atlanta, USA, 2006, p. 94.

[9] A. Machanavajjhala, J. Gehrke, D. Kiefer, M. Venkitasubramanian. "l-Diversity: Privacy beyond k-anonymity". Proc. IEEE Int. Conf. Data Eng. (ICDE), Atlanta, USA, April 2006, p. 24.

[10] M. Duckham, K. Mason, J. Stell y M. Worboys. "A formal approach to imperfection in geographic information". Comput., Environ., Urban Syst., vol. 25, núm. 1, pp. 89-103, 2001.

[11] J. Domingo-Ferrer. "Microaggregation for database and location privacy". Proc. Int. Workshop Next-Gen. Inform. Technol., Syst. (NGITS), ser. Lecture Notes Comput. Sci. (LNCS), Springer-Verlag, vol. 4032, Israel, Jul. 2006, pp. 106-116.

[12] T. M. Cover, J. A. Thomas. Elements of Information Theory, 2ª ed. Nueva York: Wiley, 2006.

[13] C. E. Shannon. "Communication theory of secrecy systems," Bell Syst., Tech. J., 1949.

[14] C. Díaz, S. Seys, J. Claessens, B. Preneel. "Towards measuring anonymity". Proc. Workshop Privacy Enhanc. Technol. (PET), ser. Lecture Notes Comput. Sci. (LNCS), Springer-Verlag, vol. 2482, April 2002.

[15] D. V. Lindley. "On a measure of the information provided by an experiment". Annals Math. Stat., vol. 27, núm. 4, pp. 986-1005, 1956.

