Assessing Disclosure Risk in Anonymized Datasets

Alexei Kounine* and Michele Bezzi†
*Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
Email: alexei.kounine@epfl.ch

†Accenture Technology Labs
Sophia Antipolis, France
Email: michele.bezzi@accenture.com

***Abstract*—Sharing of log data is a valuable step towards the improvement of network security. However, logs often contain sensitive information and organizations are hesitant to share them. Anonymization methods are used for increasing protection, lowering the disclosure risk to a level considered safe. Accordingly, a metric for anonymity is necessary to quantitatively assess the risk before releasing log data. In this paper, we propose a general framework for estimating disclosure risk using conditional entropy between the original and the anonymized datasets. We demonstrate our approach using network log files.**

I. Introduction

Log data analysis is a powerful tool for improving network security. Typically, each organization uses only their own logs, but, with the increasing number of coordinated attacks, sharing log information between organizations is becoming essential [1]. The problem is that organizations are often reluctant to share their logs, since the information contained in them can be sensitive. Anonymization methods are used for limiting disclosure risk in releasing such sensitive datasets. Anonymizing data increases protection, lowering the disclosure risk, but, it also decreases the quality of the data and hence its utility [2]. Finding the optimal trade off between risk and utility is the main scope of the anonymization process. Both quantities are hard to define, and strongly depend on context variables, e.g., data usage, level of knowledge of the attacker, amount of data released. In this paper we focus on the evaluation of disclosure risk (Section II). The main contribution of this paper is to introduce a general measure of disclosure risk, which is applicable to any set of masking transformations. Unlike previous measures, we do not assume any specific masking algorithm. Moreover, our measure provides a robust estimation of risk both at single record level (local risk) and at global level, i.e. for the whole dataset (Section III). Our model is therefore general and can be applied for quantitatively comparing different anonymization policies. Furthermore, it is directly related to the measure of the information lost in the anonymization procedure. We implemented this risk estimator using the FLAIM framework [3] and tested on network log files (Section IV).

II. Privacy in public datasets

Data holders, such as national statistical institutes, often have to release data files containing information on individual people or firms (micro-data) for research purpose. At the same time they have to preserve the privacy of individuals. This problem also occurs for sharing log files, since they may contain personal information which cannot be released in its original form (IP addresses, port numbers, timestamps, quantities...). Consequently, these data holders need to anonymize their databases before release, using data masking algorithms such as: generalizing the data, i.e., recoding variables into broader classes (e.g., releasing only the first two digits of the zip code or removing the last octet of an IP address), suppressing part of or entire records (also known as black marker [3]), randomly swapping some fields among original data records, applying permutations (one-to-one mapping on a defined set) or perturbative masking, i.e., adding random noise to numerical data values.

When masking methods have been applied, data holders have to quantitatively assess the disclosure risk (or anonymity level), to verify whether it is below a defined threshold, in which case it is assumed to be acceptable. To this scope, various measures for estimating disclosure risk have been proposed so far [4], [5], [6]; their validity strongly depends on the application scenarios considered, but still, there is a consensus that the risk of disclosure cannot be reduced to zero (but removing all the information). Thus, in general, a threshold should be determined to decide whether to release a dataset or not. Broadly speaking, there are two different approaches for assessing disclosure risk: estimating the *rareness* in the sample or population, or estimating the probability of re-identifying a masked record using some external information.

Let us examine these two methods in detail. In a typical scenario an attacker has knowledge about some variables, which may identify a record in the dataset. Considering the example of a medical database, the attacker may know a few attributes (age, gender, marital status) from an external public register (census data) or some private source of information (e.g., knowing age and address of his neighbor). He then tries to

(a) Original log file $\mathcal{S}$

SrcIP	SrcPort	DestIP	DestPort	Packets
168.125.253.23	80	147.81.124.173	3157	40
39.109.219.43	7310	142.68.227.108	59959	126
35.187.130.82	161	213.48.191.68	55867	83

(b) Anonymized log file $\mathcal{R}$

SrcIP	SrcPort	DestIP	DestPort	Packets
168.125.253.0	1023	10.1.1.1	65535	42
39.109.219.0	65535	10.1.1.1	65535	132
35.187.130.0	1023	10.1.1.1	65535	81

(c) Background knowledge $\widehat{\mathcal{S}}$

SrcIP	SrcPort	DestIP	DestPort	Packets
39.109.219.43	7310	142.68.0.0	—	—

TABLE I: Example of original ($\mathcal{S}$) and anonymized log files ($\mathcal{R}$). In the anonymization process, the least-significant 8 bits of the SrcIP are blacked out, BM(8) (replaced with 0s). SrcPort and DestPort are partitioned in two classes (1023 and 65535), called binary classification (C). DestPrt is completely blacked out, BM(32). Packets are perturbed with random Gaussian noise.

match these variables (*keys*) with the partly altered records in the released database. In the case of log files, an attacker may inject some information (e.g., scanning some specific ports), with the goal of later recognizing them in the anonymized logs. When a unique record matches a combination of key variables, the intruder can re-identify the masked record, assuming he is certain that the record is in the dataset. In fact, even if there is more than a unique match, but the number of linked records characterized by that combination of keys is still low (say it does not exceed a threshold *k*), these records have a high risk of re-identification. This rule is known as k-anonymity [7]. This approach has some limitations: it does not consider intruder's knowledge explicitly, and, in case of continuous variables the number of population uniques could be extremely large, especially when these data are randomly perturbed during the masking process.

The second approach consists of estimating the probability of re-identification. As in the previous case, the attacker aims at linking pairs of records in the released database with his background information [8], [4], [9]). This method permits to assess the risk in both categorical and continuous data: a record is considered at risk if this probability exceeds a fixed threshold. The main issue with this approach is finding a reliable strategy to compute these probabilities, since in case there are many records with similar, close to threshold, probabilities of re-identification, the risk estimation can be strongly affected by random fluctuations.

III. Entropy based risk estimator

The protection model we propose here creates a measure of disclosure risk for micro-data release, which combines together the two approaches described above. This allows us to develop a measure applicable in general cases (i.e. for any kind of data transformation, as when using the probability of re-identification method) and, at the same time, it considers the whole distribution of original records (as in k-anonymization). The basic idea is to use Shannon entropy as a measure of disclosure risk for a single record. Entropy metrics have previously been proposed for computing information loss [10], and, more recently for estimating disclosure risk for tabular data [11] and in network communications [12], [13].

In this section, we briefly review the theoretical framework and analyze its mathematical features. We refer the reader to [14] for a more extended discussion on the topic.

Let us consider a dataset $\mathcal{S}$ containing some sensitive data, e.g., network log files (Table I(a)). Each entry $s \in \mathcal{S}$ of this dataset is transformed using a data masking procedure, for example one or more of the ones mentioned in the previous section. The final result is an anonymized version of $\mathcal{S}$ dataset, which we call $\mathcal{R}$ (Table I(b)).

The attacker aims at re-identifying released data by linking them with some external information or background knowledge $\widehat{\mathcal{S}}$ (Table I(c)), which has some overlapping attributes with the released dataset. If the attacker is able to reconstruct some attribute values of the original record, we have a privacy breach. Because the data holder does not know in advance which records and attributes might be available to the attacker, it must run the risk analysis on the whole released dataset $\widehat{\mathcal{S}} \equiv \mathcal{S}$ and assume a set of key attributes (called *quasi-identifiers* in the k-anonymity framework) the attacker might know and use for re-identification. These key attributes can coincide with the whole set of attributes. The re-identification procedure consists of estimating for each $\hat{s} \in \widehat{\mathcal{S}}$ the probability of linking it with a record $r \in \mathcal{R}$: $P(r|\hat{s})$. Because we are assuming $\widehat{\mathcal{S}} \equiv \mathcal{S}$, thereafter we will consider the $P(r|s)$ instead of $P(r|\hat{s})$.

We can estimate this probability assuming the attacker simulates the data masking transformations [15], uses the information released by data holders (such as the structure of the noise added) or defines a distance function between records [16]. Intuitively, the more uncertain the mapping $P(r|s)$, the lower the disclosure risk. Shannon's entropy can be used to estimate this uncertainty. By applying it to the conditional probability $P(r|s)$, the conditional entropy is obtained:

$$H(\mathcal{R}|s) = -\sum_{r\in\mathcal{R}} P(r|s)\log_2 P(r|s) \tag{1}$$

This quantity measures the risk at the level of single record s. It represents the average number of binary question we have to ask to identify the corresponding r given s. Low entropy values indicate an almost deterministic mapping, and high risk accordingly, whereas large entropy is associated to low disclosure risk.

For example, in the case where a selected record s can be linked to exactly k_s indistinguishable records in $\mathcal{R}$ (as in k-anonymity [7]), we have a uniform distribution over the k_s records: and the corresponding specific entropy, Eq. (1) is:

$$H(\mathcal{R}|s) = \log_2 k_s \tag{2}$$

The k-anonymity condition over the whole dataset can be written as:

$$k \geq min_{s\in\mathcal{S}} 2^{H(\mathcal{R}|s)} \tag{3}$$

Global identification risk, that is at the dataset level, can be derived from the local risk measures, Eq. (1). One possible choice (see [14] for other options) is to calculate the expected number of correct matches (E_{CM}, herein):

$$E_{CM} = \sum_{s \in \mathcal{S}} \frac{1}{2^{H(\mathcal{R}|s)}} \tag{4}$$

E_{CM} is the *average number* of correct matches considering the intruder is randomly guessing according to $P(r|s)$. In fact, the entropy $H(\mathcal{R}|s)$ represents the average number of binary questions required to determine r, given s [14].

E_{CM} differs from the *estimated* number of correct matches, called N_{TM} herein, typically used for global risk assessment (see [15], [9]). These two measures differ because N_{TM} is based on maximum likelihood, which implies verifying a posteriori whether a match is correct, whereas E_{CM} is the average number of correct matches considering a random guess according to $P(r|s)$. So, the latter lacks the *decoding* part (i.e., the maximum likelihood step) and relies on the shape of the distribution only. In practice, they coincide when $P(r|s)$ has a single sharp peak, that is an almost deterministic one to one mapping. In contrast, they may strongly deviate in presence of multiple peaks and/or a smooth distribution. In addition, because E_{CM} depends on the shape of the whole distribution (not only on its peak value), it is less sensitive to random fluctuations [14]. Lastly, note that conditional entropy is directly linked to the mutual information between $\mathcal{S}$ and $\mathcal{R}$, and it can be used as an estimation of the information lost in the anonymization transformation [14].

IV. Measuring risk on anonymized network logs

We tested the entropy-based risk estimator on a publicly available Netflow log file [1]. In this analysis, we only use a limited set of fields and records. In addition we do not consider the utility of the masked dataset. Consequently, results presented here should be viewed as a proof of concept and recommendations for selecting a specific anonymization policy are not provided.

To run these tests, we developed a risk-estimating module based on FLAIM (Framework for Log Anonymization and Information Management) [3]. FLAIM is a modular and scalable framework for anonymizing log files which includes an anonymization engine with various anonymization primitives (BlackMarker, Permutation, Enumeration, etc ...). We developed a component, RiskEngine (see Figure 1), capable of estimating disclosure risk (Eq. (4)) by comparing the original and anonymized log files. As the other FLAIM components, the risk estimator works on streamed data, allowing us to process very large datasets.

A. Results

To illustrate the previously described method and its implementation in FLAIM, seven different anonymization scenarios are presented. As testing dataset we used the sample Netflow file available on the FLAIM website. The nfdump module provided in FLAIM is used for parsing the log file. We considered a subset of the available fields: the source and destination IPs, the source and destination ports and the number of bytes in a flow. The seven scenarios are summarized in Table II.

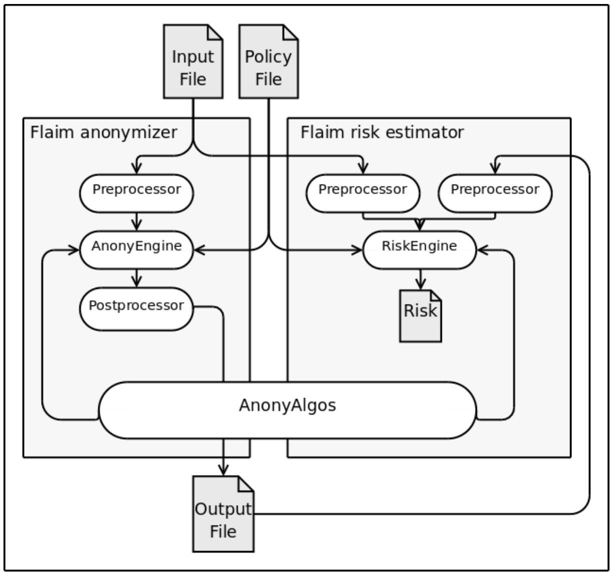

Fig. 1: The structure of the risk estimation component (RiskEngine). It is implemented as a subclass of AnonyEngine. The BasicPreprocessor and BasicPostprocessor classes were extended with interfaces to the RiskEngine. AnonyAlg and each of its subclasses now implement a method for estimating the probability $P(r|s)$ for each anonymization primitive

	SRC_IP	**DST_IP**	**SRC_PRT**	**DST_PRT**	**BYTES**
S1	None	None	None	None	None
S2	BM(16)	BM(16)	None	None	None
S3	BM(16)	BM(16)	C	C	None
S4	BM(16)	BM(16)	C	C	NA(10%)
S5	BM(24)	BM(24)	C	C	NA(10%)
S6	BRP	BRP	C	C	NA(10%)
S7	BM(32)	BM(32)	C	C	NA(10%)

TABLE II: List of the 7 anonymization scenarios discussed in the main text, in order of increasing anonymization *strength*. Legend: BM(16) (BM(24)): Black Marker applied on the 16 (24) least-significant bits. C: Classify: bins ports below 1024 in one bin and ports greater or equal to 1024 in another. NA(10%): Noise Addition: adds zero averaged Gaussian noise with a standard deviation equal to 10% of the value to anonymize. BRP: Binary Random permutation: maps each IP into a randomly generated IP in a consistent way (all IPs equal in the original log file are also equal in the anonymized log file). For more details about these transformations see Ref. [3].

Each anonymization primitive has its corresponding function for calculating the probability $P(r|s)$. For the sake of simplicity we assumed that the different fields are independent. Therefore, in the example above, $P(r|s)$ reads:

$$\begin{aligned} P(r|s) \quad &= \quad P(r|s)_{SRC\_IP} \cdot P(r|s)_{DST\_IP} \cdot \\ &\quad\;\; \cdot P(r|s)_{SRC\_PRT} \cdot P(r|s)_{DST\_PRT} \cdot P(r|s)_{BYTES} \end{aligned}$$

[1] Available at http://flaim.ncsa.uiuc.edu/downloads/flaim/sample.nfdump.log

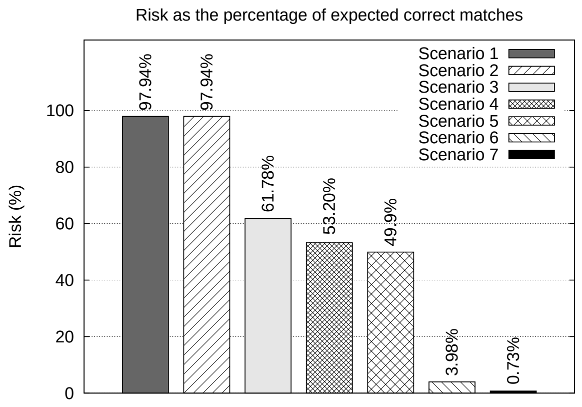

Fig. 2: Entropy-based risk for the 7 anonymization scenarios described in Table II

Figure 2 shows the expected number of correct matches, E_{CM}, as a percentage of the total number of records for the seven scenarios. Intuitively, increasing the number and strength of the anonymization methods leads to a reduced disclosure risk. In more details, we observed that removing the last 8 bits in the IP addresses have no impact on the estimated risk (Scenarios 1 and 2). Similarly suppressing the 16 or the 24 least-significant bits of the IP addresses lead to similar risk values (Scenarios 4 and 5). This indicates that, in this sample, most of the IP addresses sharing the same first octet are actually the same address (however, the corresponding port is not necessarily the same). In other words, due to a lack of diversity, most of the IPs can be identified by their first 8 most-significant bits. By generalizing the port number (Scenario 3), we observed a $\simeq 36\%$ decrease in the risk, suggesting that port re-coding could be a valuable anonymization strategy in this context. Adding random noise on the number of packets transmitted gives a further $\simeq 8\%$ decrease in the risk (Scenario 4). To obtain low risk values, we needed to remove most of the information contained in IP addresses by either using a one-to-one mapping into a predefined set (binary random permutation, scenario 6) or black marking all the 32 bits of the address (BM(32) in Scenario 7).

V. Summary

The advantage of using Shannon's entropy as a measure of disclosure risk for log file release is twofold: First, it can be applied to any general masking transformation, unlike k-anonymity measure, which is limited to non-perturbative masking transformations. Second, it only depends on the shape of the probability distribution; thus it is less sensitive to random fluctuations than measures where decoding of the masked record is needed.

The main technical issue is that computing the probability of re-identification can be hard for complex masking transformations [15]. Furthermore, these probabilities depend on the attack scenarios (attacker's knowledge, data sensitivity, etc ...), that are often difficult to model and application-specific. In the simple example we presented here, we could easily derive these probabilities under the assumptions of independence among fields and records. Both these hypotheses are unrealistic in many real world scenarios, such as in port scanning attack, where multiple ports are scanned in sequence on a single target host. Further analysis is needed to investigate the viability of this approach in realistic settings.

Acknowledgments

Alexei Kounine contributed to this study during his internship at Accenture Technology Labs in Sophia Antipolis. We thank Kiran Lakkaraju and Adam J. Slagall, for their help with FLAIM.

References

[1] A. Slagell and W. Yurcik, "Sharing computer network logs for security and privacy: A motivation for new methodologies of anonymization," 2005. [Online]. Available: citeseer.ist.psu.edu/slagell05sharing.html

[2] G. Duncan, S. Keller-McNulty, and S. Stokes, "Disclosure risk versus data utility: The RU confidentiality map," *Technical paper, Los Alamos National Laboratory, Los Alamos, NM*, 2001.

[3] A. J. Slagell, K. Lakkaraju, and K. Luo, "Flaim: A multi-level anonymization framework for computer and network logs," in *LISA*. USENIX, 2006, pp. 63–77.

[4] G. Duncan and D. Lambert, "The risk of disclosure for microdata," *Journal of Business & Economic Statistics*, vol. 7, p. 207, xx 1989, 10.2307/1391438. [Online]. Available: http://dx.doi.org/10.2307/1391438

[5] T. M. Truta, F. Fotouhi, and D. Barth-Jones, "Assessing global disclosure risk in masked microdata," in *WPES '04: Proceedings of the 2004 ACM workshop on Privacy in the electronic society*. New York, NY, USA: ACM Press, 2004, pp. 85–93.

[6] R. Benedetti and L. Franconi, "Statistical and technological solutions for controlled data dissemination," *Pre-proceedings of New Techniques and Technologies for Statistics*, vol. 1, pp. 225–232, 1998.

[7] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[8] W. E. Yancey, W. E. Winkler, and R. H. Creecy, "Disclosure risk assessment in perturbative microdata protection." in *Inference Control in Statistical Databases*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer, Ed., vol. 2316. Springer, 2002, pp. 135–152.

[9] C. J. Skinner and M. J. Elliot, "A measure of disclosure risk for microdata," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 855–867, 2002. [Online]. Available: http://www.blackwell-synergy.com/doi/abs/10.1111/1467-9868.00365

[10] L. Willenborg and T. de Waal, *Elements of statistical disclosure control*. Springer New York, 2001.

[11] A. Oganian and J. Domingo-Ferrer, "A posteriori disclosure risk measure for tabular data based on conditional entropy," *SORT*, vol. 2, pp. 175–190, 2003.

[12] A. Serjantov and G. Danezis, "Towards an Information Theoretic Metric for Anonymity," *Privacy Enhancing Technologies: Second International Workshop, PET 2002, San Francisco, CA, USA, April 14-15, 2002*, 2003.

[13] C. Diaz, S. Seys, J. Claessens, B. Preneel, and K. ESAT-COSIC, "Towards Measuring Anonymity," *Privacy Enhancing Technologies: Second International Workshop, PET 2002, San Francisco, CA, USA, April 14-15, 2002: Revised Papers*, 2003.

[14] M. Bezzi, "An entropy-based method for measuring anonymity," in *Proceedings of the IEEE/CreateNet SECOVAL Workshop on the Value of Security through Collaboration*, Nice, France, September 2007.

[15] J. P. Reiter, "Estimating risks of identification disclosure in microdata," *Journal of the American Statistical Association*, vol. 100, pp. 1103–1112, December 2005, available at http://ideas.repec.org/a/bes/jnlasa/v100y2005p1103-1112.html.

[16] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," Oct 2006. [Online]. Available: http://arxiv.org/abs/cs/0610105