On the rationale of KL divergence as a measure of the privacy of user profiles

David Rebollo-Monedero
Dept. Computer Science and Mathematics
Universitat Rovira i Virgili

Javier Parra-Arnau
Institute of Telematics
Karlsruhe Institute of Technology

Jordi Forné
Dept. Telematics
Universitat Politècnica de Catalunya

***Abstract*—This work elaborates on the rationale of using Shannon's entropy and KL divergence as measures of the privacy of a user profile. We justify their use by building upon Stein's lemma in the context of a privacy attacker who wishes to distinguish a user among the population, in a hypothesis-testing sense.**

***Index Terms*—privacy metrics, Kullback-Leibler divergence, Stein's lemma**

I. Introduction

Evaluating the privacy provided by a privacy-enhancing technology is crucial to determine its benefit, compare its effectiveness with other technologies, and ultimately to improve it. Further, quantifiable measures of the privacy gained and the cost incurred enable system designers to devise and optimize privacy-enhancing mechanisms in terms of their privacy-utility trade-off; say, maximizing privacy for a given, acceptable cost.

It is therefore not surprising that a great deal of research has been devoted to the investigation of both privacy and utility metrics. The vast majority of these metrics have emerged from the mature fields of statistical disclosure control [2] and anonymous-communication systems [3]. In applications involving user profiles, however, the literature is significantly less prolific and the Shannon's entropy and the Kullback-Leibler (KL) divergence (also known as relative entropy) are the most popular choices [5].

Although these information-theoretic quantities date back to the fifties, it was in [4] where —for the first time— their usage as privacy metrics was rigorously justified in *a context of user profiles*. Previously, the choice of Shannon's and relative entropy as measures of profiles was founded merely on uncertainty arguments: the privacy of an observed profile is high if its uncertainty so is. However, several flavors of entropy (e.g., min-entropy, Hartley, Rényi [7]) can be found in the literature, and the question is —what is the fundamental reason for choosing one instead of the other?

The cited work justified Shannon's and the relative entropies by elaborating on Jaynes' rationale behind entropy-maximization methods. In this work, we provide an additional riveting justification of these two quantities as privacy metrics, by building upon the famous Stein's lemma.

II. Information-Theoretic Concepts

Probability mass functions (PMFs) are denoted by p, subindexed by the corresponding r.v. in case of ambiguity risk. Similarly, we use the notations $p_{X|Y}$ and $p(x|y)$ equivalently.

We adopt the same notation for information-theoretic quantities used in [1]. Specifically, the symbol H will denote entropy and D relative entropy or KL divergence. We briefly recall those concepts for the reader not intimately familiar with information theory. For simplicity, we use natural logarithms.

- The *entropy* $\mathrm{H}(X)$ of a discrete r.v. X with probability distribution p is a measure of its uncertainty, and it is defined as

$$\mathrm{H}(X) = -\,\mathrm{E}\,\ln p(X) = -\sum_x p(x)\ln p(x),$$

  where E is the expectation operator. This operator is replaced by the integral when p is a PDF.

- Given two probability distributions $p(x)$ and $q(x)$ over the same alphabet, the *KL divergence* or *relative entropy* $\mathrm{D}(p\|q)$ is defined in the discrete case as

$$\mathrm{D}(p\|q) = \mathrm{E}_p\,\ln\frac{p(X)}{q(X)} = \sum_x p(x)\ln\frac{p(x)}{q(x)}.$$

The KL divergence is often referred to as *relative entropy*, as it may be regarded as a generalization of the Shannon entropy of a distribution, relative to another. Conversely, Shannon's entropy is a special case of KL divergence, as for a uniform distribution u on a finite alphabet of cardinality n,

$$\mathrm{D}(p\,\|u) = \log n - \mathrm{H}(p). \tag{1}$$

Although the KL divergence does not satisfy the symmetric property and the triangle inequality, it gives us a measure of distance or discrepancy between distributions, in the sense that $\mathrm{D}(p\|q) \geqslant 0$, with equality if, and only if, $p = q$.

This intuitive sense of distance is made more evident by reviewing Stein's lemma. Suppose we observe a sequence of k independent and identically distributed (i.i.d.) random variables (r.v.'s), and that we need to evaluate whether they have been drawn according to a probability distribution p_1, hypothesis $\mathcal{H}_1$, or p_2, hypothesis $\mathcal{H}_2$. Given these two hypothesis, we define the *acceptance region* $\mathcal{A}_k$ as the set of sequences that, when observed, $\mathcal{H}_1$ is accepted. Analogously, we define the complement of this set, $\bar{\mathcal{A}}_k$, as the set of sequences that lead us to conclude $\mathcal{H}_2$. Next, we contemplate the following probabilities of error:

- the probability of a false negative α_k, defined as the probability of accepting $\mathcal{H}_2$ when $\mathcal{H}_1$ is true,
- and the probability of a false positive β_k, defined as the probability of accepting $\mathcal{H}_1$ when $\mathcal{H}_2$ is true.

Suppose we choose an acceptance region aimed to minimize β_k while not allowing α_k to exceed a certain threshold value ϵ. Loosely speaking, Stein's lemma states that the optimal error rate, β_k^ϵ, is approximately $e^{-k\,\mathrm{D}(p_1\|p_2)}$, for large k and small ϵ.

III. KL Divergence as a Privacy Criterion in Query Forgery

Next, we describe an application scenario where the KL divergence may be employed as a privacy criterion. The chosen scenario is that of a user who would like to perturb their original search query profile to avoid disclosing an accurate view of their interests to a Web search engine (WSE). The way the perturbation is introduced is by submitting some dummy or fake queries. In the sequel, we model this scenario mathematically. We adopt the same formulation of [6].

User *queries* are regarded as r.v.'s, which take on values in a common, finite alphabet. User profiles are modeled accordingly as their corresponding PMFs. We define p as the distribution of the *population*'s queries, q as the distribution of legitimate queries of a particular *user*, and r as the distribution of queries *forged* by that user. In addition, we introduce a query *redundancy* parameter $\rho \in [0, 1)$, which represents the ratio of forged queries to total queries. Concordantly, we define the user's *apparent* query distribution as the convex combination $s = (1-\rho)\,q + \rho\,r$, which will actually be the distribution the WSE will observe.

In this context, we define *privacy risk* as the KL divergence between the user's authentic profile and the population's distribution, and the *privacy-redundancy* function as

$$\mathcal{R}(\rho) = \min_r \ \mathrm{D}((1-\rho)\,q + \rho\,r\|p), \tag{2}$$

which poses the optimal trade-off between privacy (risk) and redundancy. The minimization variable is the PMF r representing the optimum profile of forged queries, for a given ρ.

Notice that entropy maximization is a special case of divergence minimization when the reference distribution is uniform (see Eq.1). Accordingly, we may regard $\mathrm{H}(s)$ as a measure of privacy gain, rather than risk.

IV. Divergence Minimization

We operate under the mild assumption that an attacker is able to estimate the apparent query distribution s of a given user from previous activity. We suppose further that the attacker observes a sequence of new k i.i.d. queries, and tries to guess whether they have been submitted by this particular user, perhaps working from a different system, or not. More precisely, the attacker considers the binary hypothesis test between two alternatives, namely whether the queries have been drawn according to the user's apparent distribution s, hypothesis $\mathcal{U}$, or the general population's distribution p, hypothesis $\mathcal{P}$.

At this point, we consider two mutually exclusive strategies for an attacker. The former assumes the attacker is interested in bounding the probability of a false negative, that is, $\mathrm{P}(\mathcal{P}|\mathcal{U})$, assuming this attacker does not want the user to go unnoticed. From Stein's lemma, we find that the probability $\mathrm{P}(\mathcal{U}|\mathcal{P})$ of a false positive is approximately $e^{-k\,\mathrm{D}(s\|p)}$ for large k. Consequently, the minimization of $\mathrm{D}(s\|p)$ in the definition of the privacy-redundancy function (2) involves the maximization of the exponent in the error rate of false positives. In other words, the optimal distribution of forged queries r^* hinders the attacker in their efforts to recognize a user among the population, and therefore compromise user's privacy.

Suppose now that, rather than fixing the probability of a false negative, the attacker's objective is to minimize the overall probability of error

$$P_\mathrm{T} = \mathrm{P}(\mathcal{U})\,\mathrm{P}(\mathcal{P}|\mathcal{U}) + \mathrm{P}(\mathcal{P})\,\mathrm{P}(\mathcal{U}|\mathcal{P}).$$

Taking advantage of the fact that the global population's activity is much higher than that of a single user, the privacy attacker is primarily concerned with bounding $\mathrm{P}(\mathcal{U}|\mathcal{P})$ and doing their best to minimize $\mathrm{P}(\mathcal{P}|\mathcal{U})$. It turns out that the probability of a false negative given by Stein's lemma is approximately $e^{-k\,\mathrm{D}(p\|s)}$, which justifies an alternative definition of the privacy-redundancy function given by the inversion of the two arguments of the KL divergence. According to this observation, the query forgery strategy r^* minimizing $\mathrm{D}(p\|s)$, leads to the maximization of the attacker's overall probability of error and contributes to protect user's privacy.

V. Conclusion

Our contribution is a justification of the KL divergence and Shannon's entropy as quantifiable measures of the privacy of a user profile. Our justification relies upon Stein's lemma in the context of an attacker who wishes to distinguish a user among the population, in a hypothesis-testing sense.

Acknowledgment

J. Parra-Arnau is the recipient of an Alexander von Humboldt postdoctoral fellowship.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[2] T. B. Jabine, "Statistical disclosure limitation practices at united states statistical agencies," *J. Official Stat.*, vol. 9, no. 2, pp. 427–454, 1993.

[3] C. Kuhn, M. Beck, S. Schiffner, E. Jorswieck, and T. Strufe, "On privacy notions in anonymous communication," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*. Springer-Verlag, 2019.

[4] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Gen. Comput. Syst. (FGCS), Special Issue Data, Knowl. Eng.*, vol. 33, pp. 53–63, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.01.001

[5] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, *Advanced Research in Data Privacy*. Springer-Verlag, 2015, ch. Privacy-Enhancing Technologies and Metrics in Personalized Information Systems, pp. 423–442.

[6] D. Rebollo-Monedero and J. Forné, "Optimal query forgery for private information retrieval," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4631–4642, 2010.

[7] D. Rebollo-Monedero, J. Parra-Arnau, J. Forné, and C. Diaz, "Optimizing the design parameters of threshold pool mixes for anonymity and delay," *Compu. Netw.*, pp. 180–200, 2014.