A Criterion for the Measurement of the Privacy of User Profiles

David Rebollo-Monedero, Javier Parra-Arnau, Jordi Forné

Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain

{david.rebollo, javier.parra, jforne}@entel.upc.edu

Abstract

In previous work, we presented a novel information-theoretic privacy criterion for query forgery in the domain of information retrieval. Our criterion measured privacy risk as a divergence between the user's and the population's query distribution, and contemplated the entropy of the user's distribution as a particular case. In this work, we make a twofold contribution. First, we thoroughly interpret and justify the privacy metric proposed in our previous work, elaborating on the intimate connection between the celebrated method of entropy maximization and the use of entropies and divergences as measures of privacy. Secondly, we attempt to bridge the gap between the privacy and the information-theoretic communities by substantially adapting some technicalities of our original work to reach a wider audience, not intimately familiar with information theory and the method of types.

1 Introduction

During the last two decades, the Internet has gradually become a part of everyday life. One of the most frequent activities when users browse the Web is submitting a query to a search engine. Search engines allow users to retrieve information on a great variety of categories, such as hobbies, sports, business or health. However, most of them are unaware of the privacy risks they are exposed to [1].

As a concrete example, from November to December of 2008, 61% of adults in the U.S. looked for online information about a particular disease, a specific treatment, an alternative medicine, and other related topics [2]. Such queries could disclose sensitive information and be used to profile users about potential diseases. In the wrong hands, such private information could be the cause of discriminatory hiring, or could seriously damage someone's reputation.

The fact is that the literature of information retrieval abounds with examples of user privacy threats. Those include the risk of user profiling not only by an Internet search engine, but also by location-based service (LBS) providers, or even corporate profiling by patent and stock market database providers. In this context, query forgery, which consists in accompanying genuine with forged queries, appears as an approach, among many others, to preserve user privacy

to a certain extent, if one is willing to pay the cost of traffic and processing overhead.

In our previous work [3], we presented a novel information-theoretic privacy criterion for query forgery in the domain of information retrieval. Our criterion measured privacy risk as a divergence between the user's and the population's query distribution, and contemplated the entropy of the user's distribution as a particular case. In this work, we make a twofold contribution. First, we thoroughly interpret and justify the privacy metric proposed in our previous work, elaborating on the intimate connection between the celebrated method of entropy maximization and the use of entropies and divergences as measures of privacy. Secondly, we attempt to bridge the gap between the privacy and the information-theoretic communities by substantially adapting some technicalities of our original work to reach a wider audience, not intimately familiar with information theory and the method of types.

Sec. 2 examines some fundamental concepts related to information theory which will help to better understand the essence of this work. Inspired by the maximum entropy method, we put forth an information-theoretic criterion to measure the privacy of user profiles in Sec. 3. Conclusions are drawn in Sec. 4.

2 Statistical and Information-Theoretic Preliminaries

This section establishes notational aspects, and, in order to make our presentation suited to a wider audience, recalls key information-theoretic concepts assumed to be known in the remainder of the paper. The measurable space in which a *random variable* (r.v.) takes on values will be called an *alphabet*, which, with a mild loss of generality, we shall always assume to be finite. We shall follow the convention of using uppercase letters for r.v.'s, and lowercase letters for particular values they take on. The *probability mass function* (PMFs) p of an r.v. X is essentially a *relative histogram* across the possible values determined by its alphabet. Informally, we shall occasionally refer to the function p by its value $p(x)$. The *expectation* of an r.v. X will be written as $\mathrm{E}\, X$, concisely denoting $\sum_x x\, p(x)$, where the sum is taken across all values of x in its alphabet.

We adopt the same notation for information-theoretic quantities used in [4]. Concordantly, the symbol H will denote entropy and D relative entropy or Kullback-Leibler (KL) divergence. We briefly recall those concepts for the reader not intimately familiar with information theory. All logarithms are taken to base 2. The *entropy* $\mathrm{H}(p)$ of a discrete r.v. X with probability distribution p is a measure of its uncertainty, defined as

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

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 as

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

The KL divergence is often referred to as *relative entropy*, as it may be regarded as a generalization of entropy of a distribution, relative to another. Conversely,

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 is not a distance in the mathematical sense of the term, because it is neither symmetric nor satisfies the triangle inequality, it does provide a measure of discrepancy between distributions, in the sense that $\mathrm{D}(p\,\|\,q) \geq 0$, with equality if, and only if, $p = q$. On account of this fact, relation (1) between entropy and KL divergence implies that $\mathrm{H}(p) \leqslant \log n$, with equality if, and only if, $p = u$. Simply put, *entropy maximization* is a special case of *divergence minimization*, attained when the distribution taken as optimization variable is identical to the *reference distribution*, or as "close" as possible, should the optimization problem appear accompanied with *constraints* on the desired space of candidate distributions.

3 Entropy and Divergence as Measures of Privacy

In this paper we shall interpret entropy and KL divergence as privacy criteria. For that purpose, we shall adopt the perspective of Jaynes' celebrated *rationale on entropy maximization methods* [5], which builds upon the *method of types* [4, §11], a powerful technique in large deviation theory whose fundamental results we proceed to review.

The first part of this section will tackle an important question. Suppose we are faced with a problem, formulated in terms of a model, in which a probability distribution plays a major role. In the event this distribution is unknown, we wish to assume a feasible candidate. What is the most likely probability distribution? In other words, what is the "probability of a probability" distribution? We shall see that a widespread answer to this question relies on choosing the distribution *maximizing the Shannon entropy*, or, if a reference distribution is available, the distribution *minimizing the KL divergence* with respect to it, commonly subject to feasibility constraints determined by the specific application at hand.

In the second part of this section, the key idea is to model a user profile as a histogram of relative frequencies across categories of interest, regard it as a probability distribution, apply the maximum entropy method to measure the likelihood of a user profile either as its entropy or as its divergence with respect to the population's average profile, and finally take that likelihood as a measure of anonymity.

3.1 Rationale behind the Maximum Entropy Method

A wide variety of models across diverse fields have been explained on the basis of the intriguing principle of entropy maximization. A classical example in physics is the Maxwell-Boltzmann probability distribution $p(v)$ of particle velocities V in a gas [6, 7] of known temperature. It turns out that $p(v)$ is precisely the probability distribution maximizing the entropy, subject to a constraint on the temperature, equivalent to a constraint on the average kinetic energy, in turn

equivalent to a constraint on $\mathrm{E}\,V^2$. Another well-known example, in the field of electrical engineering, of the application of the maximum entropy method, is Burg's spectral estimation method [8]. In this method, the power spectral density of a signal is regarded as a probability distribution of power across frequency, only partly known.

Having motivated the maximum entropy method, we are ready to proceed to describe Jaynes' attempt to justify by reviewing the method of types of large deviation theory, a beautiful area lying at the intersection of statistics and information theory. Let $X_1, \ldots, X_k$ be a sequence of k i.i.d. drawings of an r.v. uniformly distributed in the alphabet $\{1, \ldots, n\}$. Let k_i be the number of times symbol $i = 1, \ldots, n$ appears in a sequence of outcomes $x_1, \ldots, x_k$, thus $k = \sum_i k_i$. The *type* t of a sequence of outcomes is the relative proportion of occurrences of each symbol, that is, the *empirical distribution* $t = \left(\frac{k_1}{k}, \ldots, \frac{k_n}{k}\right)$, not necessarily uniform. In other words, consider tossing an n-sided fair dice k times, and seeing exactly k_i times face i. In [5], Jaynes points out that

$$\mathrm{H}(t) = \mathrm{H}\left(\frac{k_1}{k}, \ldots, \frac{k_n}{k}\right) \simeq \frac{1}{k} \log \frac{k!}{k_1! \cdots k_n!} \quad \text{for } k \gg 1.$$

Loosely speaking, for large k, the size of a *type class*, that is, the number of possible outcomes for a given type t (permutations with repeated elements), is approximately $2^{k\,\mathrm{H}(t)}$ in the exponent. The fundamental rationale in [5] for selecting the type t with maximum entropy $\mathrm{H}(t)$ lies in the approximate equivalence between entropy maximization and the maximization of the number of possible outcomes corresponding to a type. In a way, this justifies the infamous *principle of insufficient reason*, according to which, one may expect an approximately equal relative frequency $k_i/k = 1/n$ for each symbol i, as the uniform distribution maximizes the entropy. The principle of entropy maximization is extended to include constraints also in [5].

Obviously, since all possible permutations count equally, the argument only works for uniformly distributed drawings, which is somewhat circular. A more general argument [4, §11], albeit entirely analogous, departs from a prior knowledge of an arbitrary PMF $\bar{t}$, not necessarily uniform, of such samples $X_1, \ldots, X_k$. Because the empirical distribution or type T of an i.i.d. drawing is itself an r.v., we may define its PMF $p(t) = \mathrm{P}\{T = t\}$; formally, the PMF of a random PMF. Using indicator r.v.'s, it is straightforward to confirm the intuition that $\mathrm{E}\,T = \bar{t}$. The general argument in question leads to approximating the probability $p(t)$ of a type class, a fractional measure of its size, in terms of its relative entropy, specifically $2^{-k\,\mathrm{D}(t\,\|\,\bar{t})}$ in the exponent, i.e.,

$$\mathrm{D}(t\,\|\,\bar{t}) \simeq -\frac{1}{k} \log p(t) \quad \text{for } k \gg 1,$$

which encompasses the special case of entropy, by virtue of (1). Roughly speaking, the likelihood of the empirical distribution t exponentially decreases with its KL divergence with respect to the average, reference distribution $\bar{t}$.

In conclusion, the most likely PMF t is that minimizing its divergence with respect to the reference distribution $\bar{t}$. In the special case of uniform $\bar{t} = u$, this is equivalent to maximizing the entropy, possibly subject to constraints on t that reflect its partial knowledge or a restricted set of feasible choices. The application of this idea to the establishment of a privacy criterion is the object of the remainder of this work.

3.2 Measuring the Privacy of User Profiles

We are finally equipped to justify our proposal to adopt Shannon's entropy and KL divergence as measures of the privacy of a user profile. In the context of this paper, an intuitive justification in favor of entropy maximization is that it boils down to making the apparent user profile as uniform as possible, thereby hiding a user's particular bias towards certain categories of interest. But a much richer argumentation stems from Jaynes' rationale behind entropy maximization methods [5,9], more generally understood under the beautiful perspective of the method of types and large deviation theory [4, §11], which we motivated and reviewed in the previous subsection.

Under Jaynes' rationale on entropy maximization methods, the entropy of an apparent user profile, modeled by a relative frequency histogram of categorized queries, may be regarded as a measure of privacy, or perhaps more accurately, anonymity. The leading idea is that the method of types from information theory establishes an approximate monotonic relationship between the likelihood of a PMF in a stochastic system and its entropy. Loosely speaking and in our context, the higher the entropy of a profile, the more likely it is, and the more users behave according to it. This is of course in the absence of a probability distribution model for the PMFs, viewed abstractly as r.v.'s themselves. Under this interpretation, entropy is a measure of anonymity, *not* in the sense that the user's identity remains unknown, but only in the sense that higher likelihood of an apparent profile, believed by an external observer to be the actual profile, makes that profile more common, hopefully helping the user go unnoticed, less interesting to an attacker assumed to strive to target peculiar users.

If an aggregated histogram of the population were available as a reference profile, the extension of Jaynes' argument to relative entropy, that is, to the KL divergence, would also give an acceptable measure of privacy (or anonymity). Recall from Sec. 2 that KL divergence is a measure of discrepancy between probability distributions, which includes Shannon's entropy as the special case when the reference distribution is uniform. Conceptually, a lower KL divergence hides discrepancies with respect to a reference profile, say the population's, and there also exists a monotonic relationship between the likelihood of a distribution and its divergence with respect to the reference distribution of choice, which enables us to regard KL divergence as a measure of anonymity in a sense entirely analogous to the above mentioned.

4 Conclusion

In our previous work [3], we presented an information-theoretic privacy criterion for query forgery in the domain of information retrieval, which arose from the formulation of the privacy-redundancy compromise. Inspired by the work in [10], the privacy risk was measured as the KL divergence between the user's apparent query distribution, containing dummy queries, and the population's. Our formulation contemplated, as a special case, the maximization of the entropy of the user's distribution.

In this work, we make a twofold contribution. First, we thoroughly interpret and justify the privacy metric proposed in our previous work, elaborating on the intimate connection between the celebrated method of entropy maximiza-

tion and the use of entropies and divergences as measures of privacy. Measuring privacy enables us to optimize it, drawing upon powerful tools of convex optimization. The entropy maximization method is a beautiful principle amply exploited in fields such as physics, electrical engineering and even natural language processing.

Secondly, we try to bridge the gap between the privacy and the information-theoretic communities by substantially adapting some technicalities of our original work to reach a wider audience, not intimately familiar with information theory and the method of types. As neither information theory nor convex optimization are fully widespread in the privacy community, we elaborate and clarify the connection with privacy in far more detail, and hopefully in more accessible terms, than in our original work.

References

[1] D. Fallows, "Search engine users," Pew Internet and Amer. Life Project, Res. Rep., Jan. 2005.

[2] S. Fox and S. Jones, "The social life of health information," Pew Internet and Amer. Life Project, Res. Rep., Jun. 2009.

[3] 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.

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

[5] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, Sep. 1982.

[6] L. Brillouin, *Science and Information Theory.* New York: Academic-Press, 1962.

[7] E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics.* Dordrecht: Reidel, 1982.

[8] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford Univ., 1975.

[9] E. T. Jaynes, "Information theory and statistical mechanics II," *Phys. Review Ser. II*, vol. 108, no. 2, pp. 171–190, 1957.

[10] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190