A Graphical Framework for Representing Anonymity Metrics

Rajiv Bagai
Department of Electrical Engineering & Computer Science
Wichita State University, Wichita, Kansas 67260–0083, USA
Email: rajiv.bagai@wichita.edu

***Abstract*—We construct a graphical framework for representing seven well-known anonymity metrics, including the one based on Shannon entropy, and show that all of these metrics sometimes err, as they base their anonymity level measurements on just some small piece of information contained in a probability distribution, while ignoring other useful information. We thereby make a case for taking all information into consideration in arriving at the degree of anonymity associated with a probability distribution. Such a comprehensive approach shows more promise of always resulting in correct measurement.**

***Keywords*-probability distributions; entropy.**

I. Introduction

Suppose an attack attempts to determine the sender of a message M going via an anonymity system with users $u_1, u_2, \ldots, u_n$. Let $D = \langle d_1, d_2, \ldots, d_n \rangle$ be the *distribution* resulting from the attack, where each d_i is the probability of u_i being the sender of M. All values in D are real numbers in the closed interval $[0, 1]$, and their sum is 1. Let Δ_n be the set of all distributions of length n. Two special distributions in Δ_n, namely

$$\widehat{n} = \langle 1, 0, 0, \ldots, 0 \rangle, \qquad \text{and} \qquad \overline{n} = \langle \tfrac{1}{n}, \tfrac{1}{n}, \ldots, \tfrac{1}{n} \rangle,$$

are of interest. The distribution $\widehat{n}$ corresponds to no remaining anonymity, and $\overline{n}$ corresponds to full anonymity.

Structures at the core of our graphical framework are two related profiles, namely the *base-profile* of D, $\mathbf{B}_D : \mathbb{R} \to \mathbb{R}$, and the *norm-profile* of D, $\mathbf{N}_D : \mathbb{R} \to \mathbb{R}$, defined as:

$$\mathbf{B}_D(x) = \textstyle\sum_{i=1}^{n} d_i^x, \qquad \text{and} \qquad \mathbf{N}_D(x) = (\mathbf{B}_D(x))^{1/x}.$$

These profiles arise from a generalization of the concept of *distance* between points in the space $\mathbb{R}^n$, and are described in detail in Bagai and Jiang [2]. Figure 1 shows some properties of $\mathbf{N}_{\widehat{n}}(x)$, $\mathbf{B}_{\widehat{n}}(x)$, $\mathbf{B}_{\overline{n}}(x)$, and profiles $\mathbf{B}_D(x)$ and $\mathbf{N}_D(x)$, for an arbitrary $D \in \Delta_n$. It also depicts that the slopes of $\mathbf{B}_D(x)$ and $\mathbf{N}_D(x)$ are identical at $x = 1$.

Our main observation of the profiles of any distribution $D \in \Delta_n$ is the following:

> ***Observation A:*** $\mathbf{B}_D(x) = \mathbf{B}_{\widehat{n}}(x)$ *corresponds to no anonymity, and* $\mathbf{B}_D(x) = \mathbf{B}_{\overline{n}}(x)$ *corresponds to full anonymity. As the anonymity level underlying D increases, the base-profile* $\mathbf{B}_D(x)$ *moves from* $\mathbf{B}_{\widehat{n}}(x)$ *to* $\mathbf{B}_{\overline{n}}(x)$*. Similarly, for norm-profile* $\mathbf{N}_D(x)$*.*

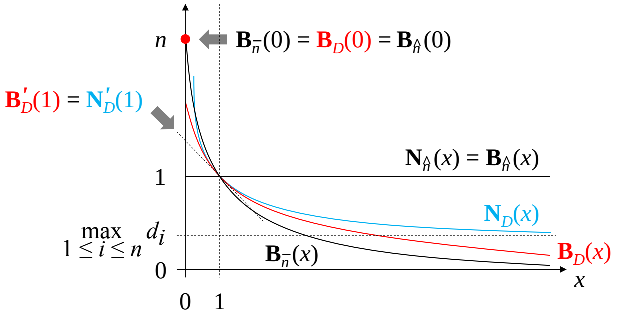

Figure 1. Profiles $\mathbf{N}_{\widehat{n}}(x)$, $\mathbf{B}_{\widehat{n}}(x)$, $\mathbf{B}_{\overline{n}}(x)$, along with profiles $\mathbf{B}_D(x)$ and $\mathbf{N}_D(x)$, for an arbitrary $D \in \Delta_n$.

In order to accurately measure the anonymity level underlying D, it is essential to measure how "close" the *entire* curve $\mathbf{B}_D(x)$ is to that of $\mathbf{B}_{\overline{n}}(x)$, i.e. over *all* values $x \geq 1$.

II. Existing Anonymity Metrics

Well-known anonymity metrics, for a given distribution D, essentially attempt to measure the closeness of the curves $\mathbf{B}_D(x)$ and $\mathbf{B}_{\overline{n}}(x)$ by just looking at some *local* portions of $\mathbf{B}_D(x)$. This can be seen by representing each of these metrics in our graphical framework, as summarized below:

Anonymity set size metric of Chaum [3]: Number of users in the system, n, in our framework, is $\mathbf{B}_D(0)$.

Reduced anonymity set size metric of Kesdogan, Egner and Büschkes [7]: Number of users in the system with a nonzero probability in D, in our framework, is $\lim_{x \to 0} \mathbf{B}_D(x)$.

Shannon entropy based metric of Serjantov and Danezis [8]: This metric can be shown to be the negation of the slope of profiles $\mathbf{B}_D(x)$ and $\mathbf{N}_D(x)$, at $x = 1$, as follows: $S(D) = -\sum_{i=1}^{n} d_i \log d_i = -\mathbf{B}'_D(1) = -\mathbf{N}'_D(1)$.

Normalized Shannon entropy based metric of Diaz et al. [6]: This metric can be shown to be the ratio of slope of profiles of D and $\overline{n}$, at $x = 1$, as follows: $d(D) = S(D) \,/\, \log n = \mathbf{B}'_D(1) \,/\, \mathbf{B}'_{\overline{n}}(1) = \mathbf{N}'_D(1) \,/\, \mathbf{N}'_{\overline{n}}(1)$.

Maximal probability metric of Tóth, Hornák and Vajda [9]: Largest probability in D, namely $\max_{i=1}^{n} d_i$, can be shown in our framework to be $\lim_{x \to \infty} \mathbf{N}_D(x)$.

Rényi entropies based metric family of Clauß and Schiffner [5]: This parametric family of metrics, given by $R_\alpha(D) = \frac{1}{1-\alpha}\log\sum_{i=1}^{n} d_i^\alpha$, where $\alpha \in [0,1) \cup (1,\infty)$ is a real-valued parameter of the family, has as its maximum value $R_0(D) = \log n = -\mathbf{B}'_{\overline{n}}(1) = -\mathbf{N}'_{\overline{n}}(1)$. Its minimum value is $\lim_{\alpha\to\infty} R_\alpha(D) = -\mathbf{B}'_{\overline{1/MAX(D)}}(1) = -\mathbf{N}'_{\overline{1/MAX(D)}}(1)$, where for any real value $\mu > 0$, we define $\mathbf{B}_{\overline{\mu}}(x) = 1/\mu^{x-1}$, and $\mathbf{N}_{\overline{\mu}}(x) = (\mathbf{B}_{\overline{\mu}}(x))^{1/x}$.

Euclidean distance metric of Andersson and Lundin [1]: The Euclidean distance between D and $\overline{n}$, given by $\sqrt{\sum_{i=1}^{n}\left(d_i - \frac{1}{n}\right)^2}$, can be shown to be $\sqrt{\mathbf{B}_D(2) - \lim_{x\to\infty}\mathbf{N}_{\overline{n}}(x)}$.

III. Shortcomings of Existing Metrics

By placing the existing metrics in our graphical framework, Section II showed that each of these metrics attempts, in its own approximate way, to essentially measure the closeness of $\mathbf{B}_D(x)$ and $\mathbf{B}_{\overline{n}}(x)$, as required by Observation A. For example, the metrics based on Shannon entropy look at the slope of $\mathbf{B}_D(x)$ at $x = 1$, while the Euclidean distance metric looks at the value of $\mathbf{B}_D(x)$ at $x = 2$, etc. Such approximate ways usually work, except when profiles of two distributions intersect.

As a simple example, consider the distributions $D = \langle\frac{7}{16}, \frac{7}{16}, \frac{1}{8}\rangle$, and $E = \langle\frac{13}{24}, \frac{11}{48}, \frac{11}{48}\rangle$. Then, as shown in Figure 2, $\mathbf{B}_D(x)$ and $\mathbf{B}_E(x)$ intersect at $x = 2$, i.e. $\mathbf{B}_D(2) = \mathbf{B}_E(2)$, because: $\left(\frac{7}{16}\right)^2 + \left(\frac{7}{16}\right)^2 + \left(\frac{1}{8}\right)^2 = \left(\frac{13}{24}\right)^2 + \left(\frac{11}{48}\right)^2 + \left(\frac{11}{48}\right)^2$. Entropy-based metrics declare E as the distribution with higher anonymity because $\mathbf{B}'_E(1)$ is closer to $\mathbf{B}'_{\overline{n}}(1)$, while the Euclidean distance metric assigns the same anonymity level to D and E. But it is clear that the attack resulting in E is stronger than the one resulting in D, because the most suspicious user in E is in a class by itself and is exposed more than the two most suspicious users in D. Thus, D should be assigned a higher anonymity.

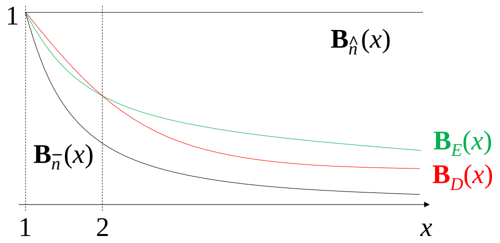

Figure 2. Base-profiles intersecting at $x = 2$.

In general, if $D, E \in \Delta_n$ are distinct distributions, neither of which is any of the extreme distributions, $\widehat{n}$ or $\overline{n}$, then their base-profiles, $\mathbf{B}_D(x)$ and $\mathbf{B}_E(x)$, may intersect an arbitrary number of times. Which of these two intersecting profiles is closer to $\mathbf{B}_{\overline{n}}(x)$ changes at each intersection.

IV. A New, Global Anonymity Metric

We believe a global metric that takes *entire* profiles into account, rather than just some of their *local* aspects, will result in more accurate anonymity measurement in all situations. One such metric, based on the area swept under the entire base-profile curves, was proposed by Bagai and Jiang [2]: $R(D) = 1 \,/ \int_1^\infty \mathbf{B}_D(x)\; dx$, which simplifies to: $R(D) = 1 \,/\, (\sum_{i=1}^{n} d_i \,/\, (-\log d_i))$. This metric possesses an intriguing duality with the Shannon entropy based metric, $S(D) = \sum_{i=1}^{n} d_i(-\log d_i)$ of Serjantov and Danezis [8]. If each d_i value is interpreted as the weight of its corresponding $-\log d_i$ value, then $S(D)$ is the weighted *arithmetic* mean of all the $-\log d_i$ values, whereas $R(D)$ is their weighted *harmonic* mean. It is well-known that in many situations, harmonic mean is the truer measure of average (see, for example, Chou [4]). Perhaps also for the task of anonymity measurement?

References

[1] C. Andersson and R. Lundin, "On the fundamentals of anonymity metrics," in *The Future of Identity in the Information Society*, ser. IFIP International Federation for Information Processing, S. Fischer-Hübner, P. Duquenoy, A. Zuccato, and L. Martucci, Eds. Springer, 2008, vol. 262, pp. 325–341.

[2] R. Bagai and N. Jiang, "Measuring anonymity by profiling probability distributions," in *Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TRUSTCOM)*, Liverpool, UK, 2012, pp. 366–374.

[3] D. Chaum, "The dining cryptographers problem: Unconditional sender and recipient untraceability," *Journal of Cryptology*, vol. 1, no. 1, pp. 65–75, 1988.

[4] Y. Chou, *Statistical analysis for business and economics*. Elsevier Publishing, 1989.

[5] S. Clauß and S. Schiffner, "Structuring anonymity metrics," in *Proceedings of the ACM Workshop on Digital Identity Management*, 2006, pp. 55–62.

[6] C. Diaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Proceedings of the 2nd Privacy Enhancing Technologies Workshop*, ser. Lecture Notes in Computer Science, vol. 2482, Springer-Verlag, R. Dingledine and P. Syverson, Eds., San Francisco, USA, 2002, pp. 54–68.

[7] D. Kesdogan, J. Egner, and R. Büschkes, "Stop-and-go- MIXes providing probabilistic anonymity in an open system," in *Proceedings of the International Information Hiding Workshop*. Lecture Notes in Computer Science - 1525, 1998, pp. 83–98.

[8] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proceedings of the 2nd Privacy Enhancing Technologies Workshop*, ser. Lecture Notes in Computer Science, vol. 2482, Springer-Verlag, R. Dingledine and P. Syverson, Eds., San Francisco, USA, 2002, pp. 41–53.

[9] G. Tóth, Z. Hornák, and F. Vajda, "Measuring anonymity revisited," in *Proceedings of the 9th Nordic Workshop on Secure IT Systems*, 2004, pp. 85–90.