





Calculating Accumulative Privacy Leakage

Majid Rafiei¹  , Gamal Elkoumy² , and Wil M.P. van der Aalst¹ 

¹ Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany
`{majid.rafiei,wvdaalst}@pads.rwth-aachen.de`

² University of Tartu, Tartu Estonia
`gamal.elkoumy@ut.ee`

Cao et al. [1] show that the accumulative privacy leakages can be formulated as an optimization problem where the objective function is a ratio of two linear functions and the constraints are linear equations. Since the method for analyzing both backward and forward forms of accumulative privacy leakage are the same, in the following, we demonstrate the calculation of $AL_B(\cdot)$.

We rely on transition systems to obtain temporal correlations. Thus, the knowledge of temporal correlations is bounded to the traces in the state space of the transition system. For the traces that are not included in the state space, we consider the worst case w.r.t. the knowledge of correlations, i.e., the probability is 0. Thus, backward and forward privacy leakages are calculated based on the traces included in the transition system of an event log.

Given $TS_{L,state_{hd}}()=(S,A,T)$ as a transition system based on an event log L , let $q=(q_1, q_2, \dots, q_n)$ and $d=(d_1, d_2, \dots, d_n)$ be two vectors representing backward correlations of two different arbitrary states of a case at release point i . For instance, assume $\sigma_c^i=s_5 \in S$ and $\sigma_c^i=s_3 \in S$. Then, $q_1=Pr(\sigma_c^{i-1}=s_1|\sigma_c^i=s_5)$, $q_2=Pr(\sigma_c^{i-1}=s_2|\sigma_c^i=s_5)$, etc., and $d_1=Pr(\sigma_c^{i-1}=s_1|\sigma_c^i=s_3)$, $d_2=Pr(\sigma_c^{i-1}=s_2|\sigma_c^i=s_3)$, etc. We denote V as the universe of such vectors obtained from the given transition system. Let $x=(x_1, x_2, \dots, x_n)^T$ be a vector representing $Pr(\tilde{L}'^1, \dots, \tilde{L}'^{i-1} | \tilde{L}'^{i-1} \uplus \sigma_c^{i-1})$ for different values of $\sigma_c^{i-1} \in S$, e.g., $x_1 = Pr(\tilde{L}'^1, \dots, \tilde{L}'^{i-1} | \tilde{L}'^{i-1} \uplus s_1)$ and $x_2 = Pr(\tilde{L}'^1, \dots, \tilde{L}'^{i-1} | \tilde{L}'^{i-1} \uplus s_2)$, etc. Hence, we can obtain the following:

$$AL_B(BPL(Ad_B^{L^{1..n}}, \mathcal{M}^{i-1})) = \sup_{q,d \in V} \log \frac{q_1 x_1 + q_2 x_2 + \dots + q_n x_n}{d_1 x_1 + d_2 x_2 + \dots + d_n x_n} = \sup_{q,d \in V} \log \frac{qx}{dx} \quad (1)$$

Suppose that $BPL(Ad_B^{L^{1..n}}, \mathcal{M}^{i-1}) = \alpha_B^{i-1}$. Based on the main definition of BPL (Definition 13), α_B^{i-1} is the supremum. Thus, for any $x_n, x_m \in x$, $e^{-\alpha_B^{i-1}} \leq \frac{x_n}{x_m} \leq e^{\alpha_B^{i-1}}$. $AL_B(\alpha_B^{i-1})$ is formalized as the following optimization problem:

$$\begin{aligned} & \text{maximize: } \log \frac{qx}{dx} \\ & \text{subject to: } e^{-\alpha_B^{i-1}} \leq \frac{x_n}{x_m} \leq e^{\alpha_B^{i-1}} \text{ and } 0 < x_n, x_m < 1 \text{ where } x_n, x_m \in x \end{aligned}$$

This optimization problem is a form of *linear-fractional programming* which can be converted into a sequence of *linear programming* problems. Cao et al. [1]

show that optimal solutions always satisfy some conditions, leading to the possibility of designing an efficient algorithm to solve the optimization problem. In the following, we describe such conditions.

Suppose that the variable vector x consists of two parts: x^+ and x^- . Let q^+ , d^+ and q^- , d^- be the corresponding coefficient vectors for x^+ and x^- . Also, let $q = \sum q^+$ and $d = \sum d^+$. According to Theorem 5 in [1], if the following conditions are satisfied, the maximum value of the objective function is $\frac{q(e^{\alpha_B^{i-1}} - 1) + 1}{d(e^{\alpha_B^{i-1}} - 1) + 1}$:

$$\frac{q_j}{d_j} > \frac{q(e^{\alpha_B^{i-1}} - 1) + 1}{d(e^{\alpha_B^{i-1}} - 1) + 1} \text{ where } q_j \in q^+, d_j \in d^+ \quad (2)$$

$$\frac{q_j}{d_j} \leq \frac{q(e^{\alpha_B^{i-1}} - 1) + 1}{d(e^{\alpha_B^{i-1}} - 1) + 1} \text{ where } q_j \in q^-, d_j \in d^- \quad (3)$$

One can find q^+ and d^+ based on Corollary 2 in [1]. According to this corollary, if the above-mentioned conditions are satisfied, then $q_j > d_j$ where $q_j \in q^+$ and $d_j \in d^+$. Based on these analyses, the maximum value of the objective function is the maximum value among 2-permutations of vectors $q, d \in V$. Given a transition system $TS_{L, state_{hd}()} = (S, A, T)$, if we consider $n = |S|$ as the number of states, i.e., the number of vectors, the optimal solution can be found in time $O(n^4)$. Since probability matrices obtained from such transition systems are often sparse matrices, we have implemented an algorithm with time complexity $O(n^2 \times k^2)$ where k is the maximum number of non-zero values in the vectors. We also have the overhead of discovering a transition system and probability matrices per each release. The complexity of discovering transition systems is linear w.r.t. the length of traces. Thus, considering m as the length of the longest trace and $|L|$ as the number of traces, the complexity of discovering a transition system from an event log L is $O(|L| \times m)$. The complexity of calculating the probabilities is $O(n \times k)$ where n is the number of states and k is the maximum number of adjacent states per each state.

References

1. Cao, Y., Yoshikawa, M., Xiao, Y., Xiong, L.: Quantifying differential privacy in continuous data release under temporal correlations. *IEEE Trans. Knowl. Data Eng.* **31**(7), 1281–1295 (2019)