Poster: Optimization based Data De-anonymization

Shouling Ji1, Weiqing Li3, Jing (Selena) He5, Mudhakar Srivatsa1, and Raheem Beyah1
1Georgia Institute of Technology. 2KSU. 3IBM T. J. Watson Research Center
Email: {sji, wli64}@gatech.edu, jhe4@kennesaw.edu, msrivats@us.ibm.com, rbeyah@ece.gatech.edu

Abstract—In this poster, we study optimization based structural data De-Anonymization (DA) attack on anonymized structural data, which could be social data, e.g., Google+, and/or mobility data, e.g., the classical longitude-latitude spatiotemporal traces [5], etc.

I. INTRODUCTION AND SYSTEM MODEL

In this poster, we focus on the De-Anonymization (DA) attack on anonymized structural data, which could be social data, e.g., Google+, and/or mobility data, e.g., the classical longitude-latitude spatiotemporal traces [5], etc.

Data Model. We model the anonymized structural data by a graph $G^a = (V^a, E^a)$, where $V^a$ is the user set and $E^a$ is the edge/relationship set. For $i \in V^a$, its neighborhood is defined as $N_i^a = \{j|e_{ij}^a \in E^a\}$ and we denote the cardinality of $N_i^a$ as $|N_i^a|$, i.e., the degree of $i$. The auxiliary data is also assumed to be structural data modeled by a graph $G^u = (V^u, E^u)$, where $V^u$ and $E^u$ are the user set and edge/relationship set, respectively. Similarly, the neighborhood of $i \in V^u$ is defined as $N_i^u$.

DA Attack. Given $G^a$ and $G^u$, a DA attack can be defined as a mapping: $\sigma: V^a \rightarrow V^u$. The objective of a DA attack is to successfully de-anonymize as many users in $V^a$ as possible.

II. OPTIMIZATION BASED DA PRACTICE

ODA Framework. We first define some useful structural features for $i \in V^a$ or $V^u$ as follows. (i) Degree: For $i \in V^a$ (resp., $V^u$), its degree feature $f_d(i)$ is its degree in $G^a$ (resp., $G^u$). (ii) Neighborhood: For $i \in V^a$ (resp., $V^u$), its neighborhood feature $f_n(i)$ is a $d$-dimensional vector $(d_1^i, d_2^i, \cdots, d_d^i)$, where $d_k^i (1 \leq k \leq d)$ is the $k$-th largest degree in $\{|N_j^a| | j \in N_k^a\}$ (resp., $\{|N_j^u| | j \in N_k^u\}$). In the case that $|N_k^a| < \beta$ (resp., $|N_k^u| < \beta$), we set $d_1^i, d_2^i, \cdots, d_{\beta}^i = \Delta^a$ (resp., $d_1^i, d_2^i, \cdots, d_{\beta}^i = \Delta^u$), where $\Delta^a = \max\{|N_j^a| | j \in V^a\}$ (resp., $\Delta^u = \max\{|N_j^u| | j \in V^u\}$) is the maximum degree of $G^a$ (resp., $G^u$). (iii) Top-K reference distance: For $i \in V^a$ (resp., $V^u$), its Top-K reference distance feature $f_k(i)$ is a $K$-dimensional vector $(h_1^i, h_2^i, \cdots, h_k^i)$, where $h_k^i (1 \leq k \leq K)$ is the distance from $i$ to the user with the $k$-th largest degree in $G^a$ (resp., $G^u$). Note that it is possible $h_k^i = \infty$ if the graph is not connected. (iv) Landmark reference distance: Suppose $V_L^a = \{v_1, v_2, \cdots, v_L | v_k \in V^a\}$ is a set of users that has been de-anonymized (evidently, $V_L^a = \emptyset$ initially) to $U_L^a = \{u_1, u_2, \cdots, u_L | u_k \in V^a\}$ under some scheme with $\sigma(v_k) = u_k (1 \leq k \leq L)$. Then, for $i \in V^a \setminus V_L^a$ (resp., $V^u \setminus U_L^u$), we define its landmark reference distance feature $f_t(i) = (h_1^i, h_2^i, \cdots, h_L^i)$, where $h_k^i (1 \leq k \leq L)$ is the distance from $i$ to $v_k \in V_L^a$ (resp., $u_k \in U_L^u$). (v) Sampling closeness centrality: For $i \in V^a$ (resp., $V^u$), we define the sampling closeness centrality feature $f_c(i)$ to characterize its global topological property by sampling too much computational overhead. Formally, we first randomly sample a subset $S^a$ of $V^a$ (resp., $S^u$ of $V^u$) and then define $f_c(i) = \sum_{j \in S^a \setminus \{i\}} \frac{1}{h(i,j)}$ (resp., $f_c(i) = \sum_{j \in S^u \setminus \{i\}} \frac{1}{h(i,j)}$, where $h(i,j)$ is the distance from $i$ to $j$.

According to the features defined for each user, we can quantitatively measure the similarity between an anonymized user $i \in V^a$ and a known user $j \in V^u$. Let $f_d(i)$, $f_n(i)$, $f_k(i)$, $f_t(i)$, $f_c(i)$.

Then, we define the structural similarity between $i \in V^a$ and $j \in V^u$ as $\phi(i,j) = c_1 \cdot s(f_d(i), f_d(j)) + c_2 \cdot s(f_n(i), f_n(j)) + c_3 \cdot s(f_k(i), f_k(j)) + c_4 \cdot s(f_t(i), f_t(j))$, where $c_{1,2,3,4} \in [0, 1]$ are constant weights representing the importance of $c_1 + c_2 + c_3 + c_4 = 1$, and $s(\cdot, \cdot)$ is the Cosine similarity between two vectors.

Furthermore, given a DA scheme $\sigma$, we define the De-Anonymization Error (DE) on a user mapping $(i,j, \sigma) \in \Psi_{i,j} = \{f_d(i) - f_d(j) + (1 - \phi(i,j))\cdot |f_d(i) - f_d(j)|\}$, and the DE on $\sigma$ as $\Psi_{\sigma} = \sum_{(i,j) \in \sigma} \Psi_{i,j}$.

Since a perfect DA tends to induce the least DE according to graph theory [4], based on $\Psi_{\sigma}$, we give the framework of ODA as shown in Fig. 1. In ODA, $\Lambda^a \subseteq V^a$ is the target DA set and $\Lambda^u \subseteq V^u$ is the possible mapping set of $\Lambda^a$. GetTopDegree($X, y$) is a function to return $y$ users with the largest degree values in $X$, i.e., return $\{i| i$ has the Top-$y$ degree in $X\}$. $C(i) \subseteq \Lambda^u$ is the candidate mapping set for $i \in \Lambda^a$, which consists of the $\gamma$ most possible mappings of $i$ in $\Lambda^u$. GetTopSimilarity($i, \Lambda^u, \gamma$) is a function to return $\gamma$ users having the highest similarity scores $\phi(i, \cdot)$ with $i \in \Lambda^a$, i.e., return $\{j|j \in \Lambda^u \land \gamma \text{has the Top-$\gamma$ score $\phi(i,j)$ in $\Lambda^u$}\}$.

From ODA, it de-anonizes $G^a$ iteratively. During each iteration, ODA is trying to de-anonymize a subset of $V^a$ and seeking the sub-DA scheme $\sigma^*(\Lambda^a)$ which induces the least DE. In Line 3, we initialize $\Lambda^a$ and $\Lambda^u$ ($|\Lambda^a|, |\Lambda^u| \leq \alpha$). In Line 4, we compute a candidate mapping set $C(i)$ for each $i \in \Lambda^a$. $C(i)$ consists of the $\gamma$ most similar users of $i \in \Lambda^a$. Here, we define $C(\cdot)$ mainly for reducing the computational complexity. In stead
of trying every mapping from $i$ to $\Lambda^u$, we only consider to map $i$ to some user in $\mathcal{C}(i)$. In Line 5, we find a DA scheme $\sigma^*(\Lambda^u)$ on $\Lambda^u$ such that $\Psi_{\sigma^*(\Lambda^u)} = \min_{\{\sigma \in \mathcal{P}(\Lambda^u)\}} \{\sigma A^u \in \prod_{i \in \Lambda^u} (i \times \mathcal{C}(i))\}$, i.e., $\sigma^*(\Lambda^u)$ causes the least DE. Furthermore, the consistent rule and the pruning rule are applied to remove some unqualified DA schemes in advance, which can speed up ODA. The consistent rule makes any possible DA scheme $\sigma(\Lambda^u)$ consistent, i.e., no mapping conflict which is defined as the situation that two or more anonymized users are mapped to the same known user. The pruning rule is used to remove some DA schemes whose DE is larger than the current known least DE. After obtaining $\sigma^*(\Lambda^u)$, we accept the mappings in $\sigma^*(\Lambda^u)$ with similarities score no less than a threshold value $\theta$ (Lines 6-8). For the mappings that been rejected, they will be re-considered in the following iterations for possible better DAs. If no mapping can be accepted, we stop ODA. Subsequently, we analyze the time and space complexities of ODA in the following theorem (the proof is omitted due to space limitation).

**Theorem 1.** (i) The space complexity of ODA is $O(\min\{n^2, m+n\})$. (ii) Let $\gamma$ be some constant value, $\alpha = \Theta(\log n)$, and $\Gamma$ be the average number of accepted mappings in each iteration of ODA. Then, the time complexity of ODA is $O(m+n \log n + n^{\Theta(1)} \log \gamma + 1/\Gamma)$ in the worst case.

Finally, we make some remarks on ODA as follows. (i) ODA is a cold start algorithm, i.e., we do not need any priori knowledge, e.g., the seed mapping information [1][2][3], to bootstrap the DA process. Furthermore, unlike existing DA algorithms [1][2][3] which consist of two phases, ODA is a single-phase algorithm. Interestingly, ODA itself can act as a landmark identification algorithm. From our experiment, ODA can de-anonymize the 60-94 Top-degree users in Gowalla [5] perfectly. In addition, ODA as a landmark identification algorithm is much faster than that in [2] (with complexity of $O(nk^d \log^{-k}) = O(nk^d)$, where $d$ is maximum degree of $G^d/G^u$ and $k$ is the number of landmarks) and [3] (with complexity of $k!$, could be computationally infeasible for a PC when $k \geq 20$). (ii) ODA is an optimization based DA scheme, which is different from most of existing heuristics based solutions [1][2][3]. In ODA, the objective is to minimize a DE function. Furthermore, ODA has a polynomial time complexity of $O(m+n \log n + n^{\Theta(1)} \log \gamma + 1/\Gamma)$ in the worst case, which is computationally feasible. (iii) In ODA, one implicit assumption is $V^u = V^u$. In practice, it is possible that $V^u \neq V^u$. In this case, if $V^u$ and $V^u$ are not significantly different, ODA is also workable at the cost of some performance degradation. One better solution could be estimating the overlap between $G^a$ and $G^u$ first, and then apply ODA to the overlap to achieve better performance. We take the estimation of the overlap between $G^a$ and $G^u$ as one of the future works.

**Experiments.** We evaluate the performance of ODA on a real world dataset: Gowalla [5]. Gowalla consists of two different datasets. The first dataset is a spatiotemporal mobility trace consisting of 6.44M check-ins generated by .2M users. The second dataset is a social graph (1M edges) of the same .2M users. Now, assume the mobility trace is anonymized. Since the mobility trace does not have an explicit graph structure, supposing the social graph is the ground truth, we apply the tech-

---

**REFERENCES**

5. H. Pham, C. Shahabi, and Yan Liu, EBM - An Entropy-Based Model to Infer Social Strength from Spatiotemporal Data, SIGMOD 2013.