Poster: Optimization based Data De-anonymization

Shouling Ji†, Weiqing Li†, Jing (Selena) He§, Mudhakar Srivatsa‡, and Raheem Beyah†
†Georgia Institute of Technology, §KSU, ‡IBM 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), including social data, mobility traces, etc. We make a DA practice by presenting a novel *single-phase cold start Optimization based DA* (ODA) algorithm followed by theoretical and experimental analysis. Experimental resutls of ODA show that about $77.7\% - 83.3\%$ of the users in Gowalla (0.2M users and 1M edges) [5] are de-anonymizable, which implies optimization based DA is implementable and powerful in practice. Furthermore, We discuss the future research directions of this project.**

I. Introduction and System Model

In this poster, we focus on the De-Anonymization (DA) attack on anaonymized 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 $N_i^a = \{j | \exists e_{i,j}^a \in E^a\}$ 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*: $\sigma : V^a \to V^u$. 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* $\overline{f_n(i)}$ is a β*-dimensional vector* $(d_1^i, d_2^i, \cdots, d_\beta^i)$, where d_k^i $(1 \le k \le \beta)$ is the k-th largest degree in $\{|N_j^a| | j \in N_i^a\}$ (resp., $\{|N_j^u| | j \in N_i^u\}$). In the case that $|N_i^a| < \beta$ (resp., $|N_i^u| < \beta$), we set $d_{|N_i^a|+1}^i = d_{|N_i^a|+2}^i = \cdots = d_\beta^i = \Delta^a$ (resp., $d_{|N_i^u|+1}^i = d_{|N_i^u|+2}^i = \cdots = d_\beta^i = \Delta^u$), where $\Delta^a = \max\{|N_i^a| | i \in V^a\}$ (resp., $\Delta^u = \max\{|N_i^u| | i \in V^u\}$) 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* $\overline{f_K(i)}$ is a K*-dimensional vector* $(h_1^i, h_2^i, \cdots, h_K^i)$, where h_k^i $(1 \le k \le K)$ 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 $U_L^u = \{u_1, u_2, \cdots, u_L | u_k \in V^u\}$ under some σ with $\sigma(v_k) = u_k$ $(1 \le k \le L)$. Then, for $i \in V^a \backslash V_L^a$ (resp., $V^u \backslash U_L^u$), we define its *landmark reference distance feature* $\overline{f_l(i)} = (h_1^i, h_2^i, \cdots, h_L^i)$, where h_k^i $(1 \le k \le L)$ 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 without inducing 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 $f_c(i) = \sum\limits_{j \in S^a \backslash \{i\}} \frac{1}{h(i,j)}$ (resp., $f_c(i) = \sum\limits_{j \in S^u \backslash \{i\}} \frac{1}{h(i,j)}$), where $h(i,j)$ is the distance from i to j.

1 Define $\Lambda^a = \Lambda^u = \emptyset$;
2 **while** ***true*** **do**
3 $\quad \Lambda^a$ = GetTopDegree(V^a, α), Λ^u = GetTopDegree(V^u, α);
4 $\quad$ for every $i \in \Lambda^a$, compute a *candidate mapping set* $\mathcal{C}(i)$ = GetTopSimilarity(i, Λ^u, γ);
5 $\quad$ apply the *consistent rule* and *pruning rule* to find the de-anonymization scheme $\sigma(\Lambda^a) \in \prod\limits_{i \in \Lambda^a} (i \times \mathcal{C}(i))$ which induces the least DE $\Psi_{\sigma(\Lambda^a)}$, denoted by $\sigma^*(\Lambda^a) = \{(i_1, j_1), (i_2, j_2), \cdots, (i_\alpha, j_\alpha)\}$;
6 $\quad$ for each $(i,j) \in \sigma^*(\Lambda^a)$, **if** $\phi(i,j) \ge \theta$ **then**
7 $\quad\quad$ accept the mapping (i,j);
8 $\quad\quad V^a = V^a \setminus \{i\}$, $V^u = V^u \setminus \{j\}$;
9 $\quad$ if no mapping in $\sigma^*(\Lambda^a)$ is accepted, ***break***;

Fig. 1. ODA.

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

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

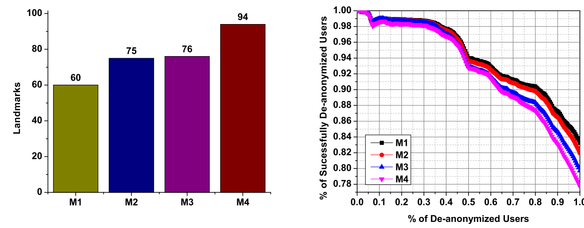
Since a perfect DA tends to induce the least DE according to graph theory [4], based on Ψ_σ, we give the framework of ODA as shown in Fig. 1. In ODA, $\Lambda^a \subseteq V^a$ is the target DA set and $\Lambda^u \subseteq V^u$ is the possible mapping set of Λ^a. GetTopDegree(X, y) is a function to return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

From ODA, it de-anonymizes 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 Λ^a and Λ^u ($|\Lambda^a|, |\Lambda^u| \le \alpha$). In Line 4, we compute a *candidate mapping set* $\mathcal{C}(i)$ for each $i \in \Lambda^a$. $\mathcal{C}(i)$ consists the γ most similar users of i in Λ^u. Here, we define $\mathcal{C}(\cdot)$ mainly for reducing the computational complexity. In stead of trying every mapping from i to Λ^u, we only consider to map i to some user in $\mathcal{C}(i)$. In Line 5, we find a DA scheme $\sigma^*(\Lambda^a)$ on Λ^a such that $\Psi_{\sigma^*(\Lambda^a)} = \min\{\Psi_{\sigma(\Lambda^a)} | \sigma(\Lambda^a) \in \prod_{i \in \Lambda^a} (i \times \mathcal{C}(i))\}$, i.e., $\sigma^*(\Lambda^a)$ 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 $\sigma(\Lambda^a)$ consistent, i.e., no *mapping confliction* 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 $\sigma^*(\Lambda^a)$, we accept the mappings in $\sigma^*(\Lambda^a)$ with similarities scores no less than a *threshold value* θ (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 γ be some constant value, $\alpha = \Theta(\log n)$, and Γ 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 $O(nd^{k-1}) = O(n^k)$, where d is maximum degree of G^a/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 $V^a = V^u$. In practice, it is possible that $V^a \neq V^u$. In this case, if V^a 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 technique in [5] on the mobility trace to construct four graphs with different *recalls* and *precisions*, denoted by $M1, M2, M3$, and $M4$, respectively (recall $= \frac{\text{true positive}}{\text{true positive}+\text{false negative}}$ and precision $= \frac{\text{true positive}}{\text{true positive}+\text{false positive}}$). Particularly, the recall and precision of $M1$ are 0.6 and 0.865, of $M2$ are 0.72 and 0.83, of $M3$ are 0.75 and 0.78, and of $M4$ are 0.8 and 0.72, respectively.



(a) Landmark identification (b) DA result

Fig. 2. De-anonymize Gowalla.

As we mentioned earlier, ODA itself can work as a *landmark identification algorithm*. Let $V_L^a = U_L^u = \emptyset$. We run ODA for Gowalla to identify landmarks as shown in Fig. 2 (a). The results show that we can de-anonymize the first 60-94 users in Gowalla perfectly. Based on the identified landmarks, we then employ ODA to de-anonymize Gowalla as shown in Fig. 2 (b), where x-axis represents the *accumulated percentage of de-anonymized users* and y-axis represents the *accumulated percentage of successfully de-anonymized users*. From Fig. 2 (b), we can see that the successful DA rate is higher for large-degree users than that of small-degree users. The reason is that large-degree users carry more structural information and thus be more accurately de-anonymizable. In summary, the results show that $77.7\% - 83.3\%$ of the users in Gowalla can be de-anonymized, which implies optimization based DA is implementable and powerful in practice.

III. Conclusion and Future Work

In this poster, we present a novel *cold start single-phase Optimization based DA* (ODA) algorithm. We also analyze ODA theoretically and experimentally.

The future work of this project can be conducted as follows. (i) In this poster, we conduct an optimization based DA practice. This motivates us to quantify the de-anonymizability of structural data theoretically, which is also an open problem; (ii) Data utility is another important concern. We propose to study how to quantify the tradeoff between privacy and utility followed by proposing privacy protection schemes with utility preservation; and (iii) Finally, due to the importance of secure data publishing, we propose to develop a *secure data publishing platform* in the future, which is expected to be invulnerable to both semantics based and structure based DA attacks.

References

[1] L. Backstrom, C. Dwork, and J. Kleinberg, *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, WWW 2007.

[2] A. Narayanan and V. Shmatikov, *De-anonymizing Social Networks*, S&P 2009.

[3] M. Srivatsa and M. Hicks, *Deanonymizing Mobility Traces: Using Social Networks as a Side-Channel*, CCS 2012.

[4] B. Bollobás, *Random Graphs (Second Edition)*, Cambridge U. Press, 2001.

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