首届致远学术节
学生科研成果展示

Social Network De-anonymization with Overlapping Communities: Analysis, Algorithm and Experiments

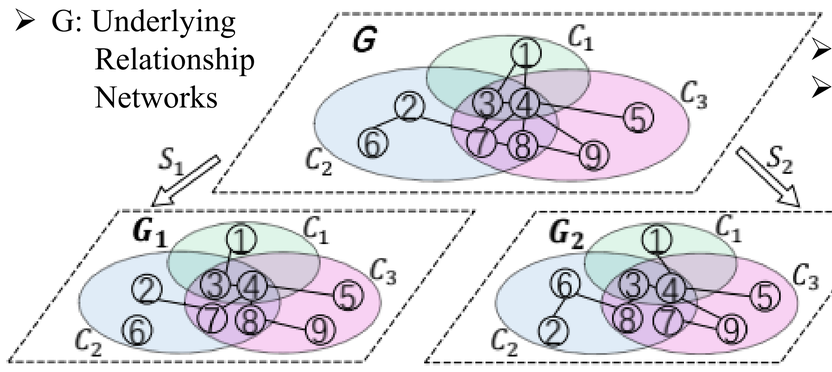
Xinyu Wu, Zhongzhao Hu, Xinzhe Fu, Luoyi Fu, Xinbing Wang, Songwu Lu

Accepted by *IEEE International Conference on Computer Communications* (*INFOCOM*) 2018. Acceptance Rate: **309/1606=19.2%**

Background & Motivations

The advent of social networks poses severe threats on user privacy as adversaries can de-anonymize users' identities by mapping them to correlated cross-domain networks.

- G: Underlying Relationship Networks
- G1: Anonymized Network
- G2: Correlated Cross-domain Unanonymized Network



Adversaries applied G2 to identify the anonymized users in G1.

--De-anonymization

Goals:

- Bring a systematic analysis of this problem in **theory**, **algorithm** and **experiments**.
  - Understanding the conditions under which adversaries can successfully de-anonymize users.
  - Proposing efficient algorithms to solve it.
  - Validating proposed algorithm under real networks.

Methodology:

1. Theory

--Modelling Overlapping Communities by the Overlapping Stochastic Block Model (OSBM).

$$Pr(\boldsymbol{C_i} = \{C_{i1}, C_{i2}, ..., C_{iQ}\}^T) = \prod_{q=1}^{Q} (p_q)^{C_{iq}} (1-p_q)^{1-C_{iq}} \quad Pr\{(i,j) \in E_k\} = \begin{cases} s_k & \text{if } (i,j) \in E, \\ 0 & \text{if } (i,j) \notin E. \end{cases}$$

--Proposing the cost function measuring de-anonymization error based on Minimum Mean Square Error (MMSE) .

$$\hat{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi} \in \mathbf{\Pi}^n} \mathbf{E}_{\mathbf{\Pi_0}}\{d(\mathbf{\Pi}, \mathbf{\Pi_0})\}$$
$$= \arg\min_{\mathbf{\Pi} \in \mathbf{\Pi}^n} \sum_{\mathbf{\Pi_0} \in \Pi^n} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 Pr(\mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta}),$$

NP-hard

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi} \in \mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2,$$

--Transforming minimizing MMSE to a polynomially solvable problem by the restriction of Sequency Inequality.

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi} \in \mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 ||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2$$

Equivalent in expectation

WEMP

$$\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi} \in \mathbf{\Pi^n}} ||\mathbf{\Pi}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi}||_F^2.$$

2. Algorithm

--Optimality of WEMP: Under mild conditions, solving WEMP ensures negligible de-anonymiztion error in large-scale networks.

$$as\ n \to \infty, \frac{||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2}{||\mathbf{\Pi_0}||_F^2} \to 0.$$

Π_0 : Ground-truth mapping
$\tilde{\Pi}$: Estimated mapping by WEMP

--Solvability of WEMP: We proposed a Convex-concave Based De-anonymization Algorithm (CBDA).

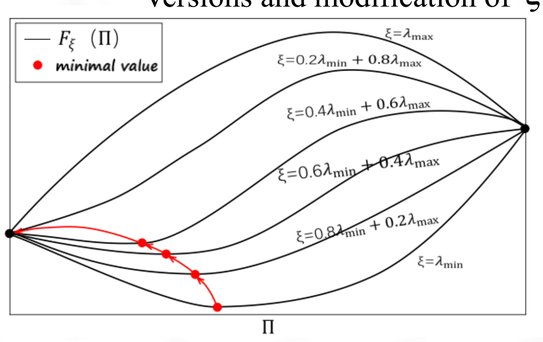
$$F_0(\mathbf{\Pi}) = ||\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi}^{\mathbf{T}}||_{\mathbf{F}}^2 + \mu||\mathbf{\Pi M} - \mathbf{M}||_F^2, \Rightarrow \text{The Equivalent Original Objective}$$

$$F_1(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + \frac{\lambda_{min}}{2}(n - ||\mathbf{\Pi}||_F^2); \Rightarrow \text{Convex Relaxation of the Objective}$$

$$F_2(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + \frac{\lambda_{max}}{2}(n - ||\mathbf{\Pi}||_F^2). \Rightarrow \text{Concave Relaxation of the Objective}$$

$$F(\mathbf{\Pi}) = (1-\alpha)F_1(\mathbf{\Pi}) + \alpha F_2(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + 2\xi(n - ||\mathbf{\Pi}||_F^2),$$
$$\xi = (1-\alpha)\lambda_{min} + \alpha\lambda_{max},\ \xi \in [\lambda_{min}, \lambda_{max}].$$

Combination of 2 Relaxed Versions and modification of ξ



Algorithm 1: Convex-concave Based De-anonymization Algorithm (CBDA)

Input: Adjacent matrices $\mathbf{A}$ and $\mathbf{B}$; Community assignment matrix $\mathbf{M}$; Weight controlling parameter μ; Adjustable parameters δ, $\Delta\xi$.
Output: Estimated permutation matrix $\tilde{\mathbf{\Pi}}$.

1: Form the objective function $F_0(\mathbf{\Pi})$ and $F(\mathbf{\Pi})$.
2: $\xi \leftarrow 0$, $k \leftarrow 1$. Initialize $\mathbf{\Pi_1}$. Set ξ_m, the upper limit of ξ.
3: **while** $\xi < \xi_m$ and $\mathbf{\Pi_k} \notin \Omega_0$ **do**
4: **while** $k = 1$ or $|F(\mathbf{\Pi_{k+1}}) - F(\mathbf{\Pi_k})| \geq \delta$ **do**
5: $\mathbf{X}^\perp \leftarrow \arg\min_{\mathbf{X}^\perp} \mathbf{tr}(\nabla_{\mathbf{\Pi_k}} F(\mathbf{\Pi_k})^T \mathbf{X}^\perp)$, where $\mathbf{X}^\perp \in \Omega$.
//Finding the optimal descent direction
6: $\gamma_k \leftarrow \arg\min_\gamma F(\mathbf{\Pi_k} + \gamma(\mathbf{X}^\perp - \mathbf{\Pi_k}))$, where $\gamma_k \in [0,1]$. //Finding the optimal step size
7: $\mathbf{\Pi_{k+1}} \leftarrow \mathbf{\Pi_k} + \gamma_k(\mathbf{X}^\perp - \mathbf{\Pi_k})$, $k \leftarrow k+1$.
//Estimation Update
8: **end while**
9: $\xi \leftarrow \xi + \Delta\xi$.
10: **end while**

Based on Frank-Wolfe algorithm

CBDA: Overcoming the brute projection by classical convex optimization technique, which may cause large estimation error.

Experiments & Results:

Experiment Settings:

TABLE II: **Main Experimental Parameters**

Notation	Definition	Range
N	Number of Nodes	{500, 1000, 1500, 2000}
s	Sampling Probability ($s_1 = s_2 = s$)	0.3-0.9
η	Community Ratio	{0.05, 0.1}
OL/NOL	Overlapping or Non-Overlapping	{OL, NOL}

Experiment Datasets:

TABLE III: **Datasets in Basic Experiments**

Dataset	Synthetic	Sampled Real Social	Cross-Domain Co-author
Source	OSBM	LiveJournal [1]	MAG [11]
Num. of Nodes	500 ∼ 2000	500 ∼ 2000	3176
Num. of Communities	25 ∼ 1000	25 ∼ 1000	89

Experiments Results:

➢ Synthetic Networks.

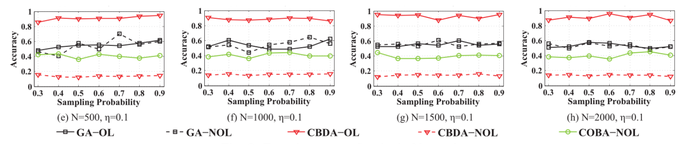


Fig. 2: Experiments on Synthetic Networks.

➢ Sampled Social Networks:

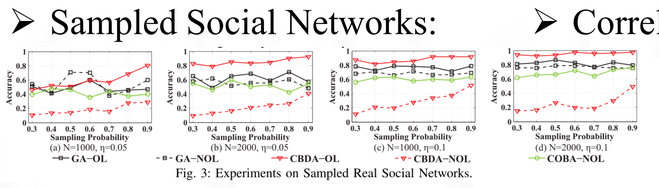
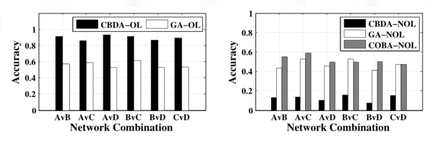


Fig. 3: Experiments on Sampled Real Social Networks.

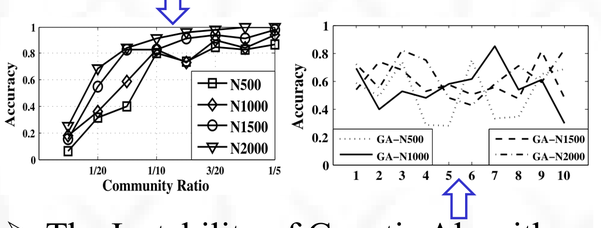
➢ Correlated Cross-Domain Networks:



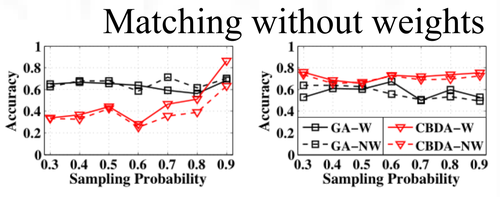
(a) Overlapping (b) Non-Overlapping

Fig. 4: Experiments on Cross-Domain Co-author Networks

➢ The Influence of Community Density



➢ MMSE vs Matching without weights



(a) The effect of $\mathbf{W}$, N=1000 (b) The effect of $\mathbf{W}$, N=2000

➢ The Instability of Genetic Algorithm

Conclusions:

1) CBDA outperforms the Genetic Algorithm and Convex Optimization Based Algorithm under networks with overlapping communities;
2) CBDA outperforms GA and COBA more in larger networks; (Meeting the theoretical results)
3) Overlapping Density positively impacts the de-anonymization accuracy; (Meeting the theoretical results)
4) De-anonymization based on MMSE outperforms the non-weighted cost function in prior art;
5) CBDA promises more practical use than GA on the factor of stability.

Paper Linkes:
9-page: http://wuxinyusjtu.3www.win/Infocom18_De_anonymization.pdf
Full paper: http://wuxinyusjtu.3www.win/IT_Xinyu_Wu.pdf

个人信息： 吴昕宇，2014级，工科
邮　　箱： wuxinyu@sjtu.edu.cn