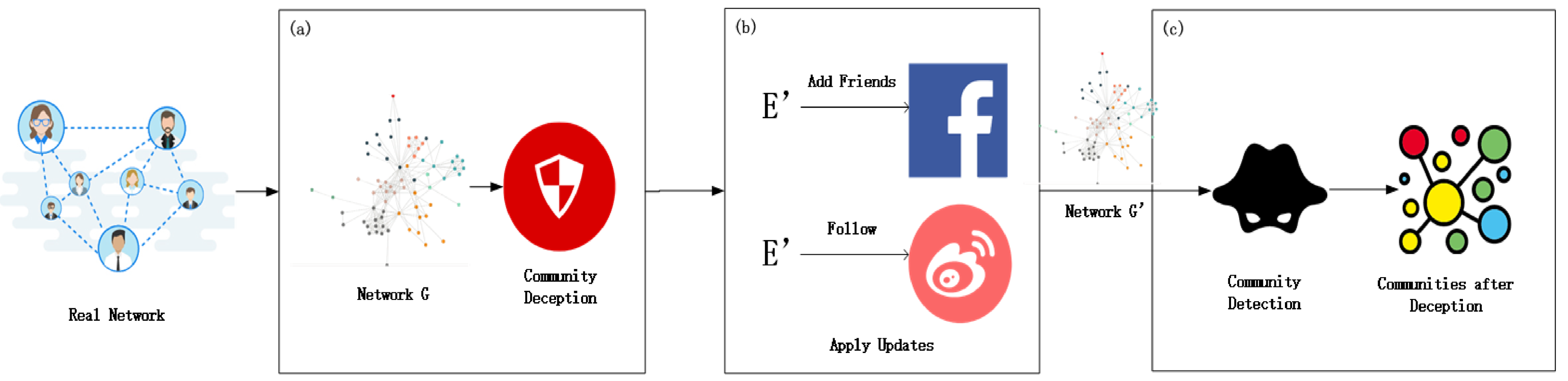
REM: From Structural Entropy To Community Structure Deception

Yiwei Liu[1,2], Jiamou Liu[3], Zijian Zhang[1,3], Liehuang Zhu[1], Angsheng Li[4]
[1]Beijing Institute of Technology, Beijing, China; [2]Zhejiang University, Zhejiang, China
[3]The University of Auckland, Auckland, New Zealand; [4]Beihang University, Beijing, China
{yiweiliu, zhangzijian, liehuangz}@bit.edu.cn, jiamou.liu@auckland.ac.nz, angsheng@buaa.edu.cn

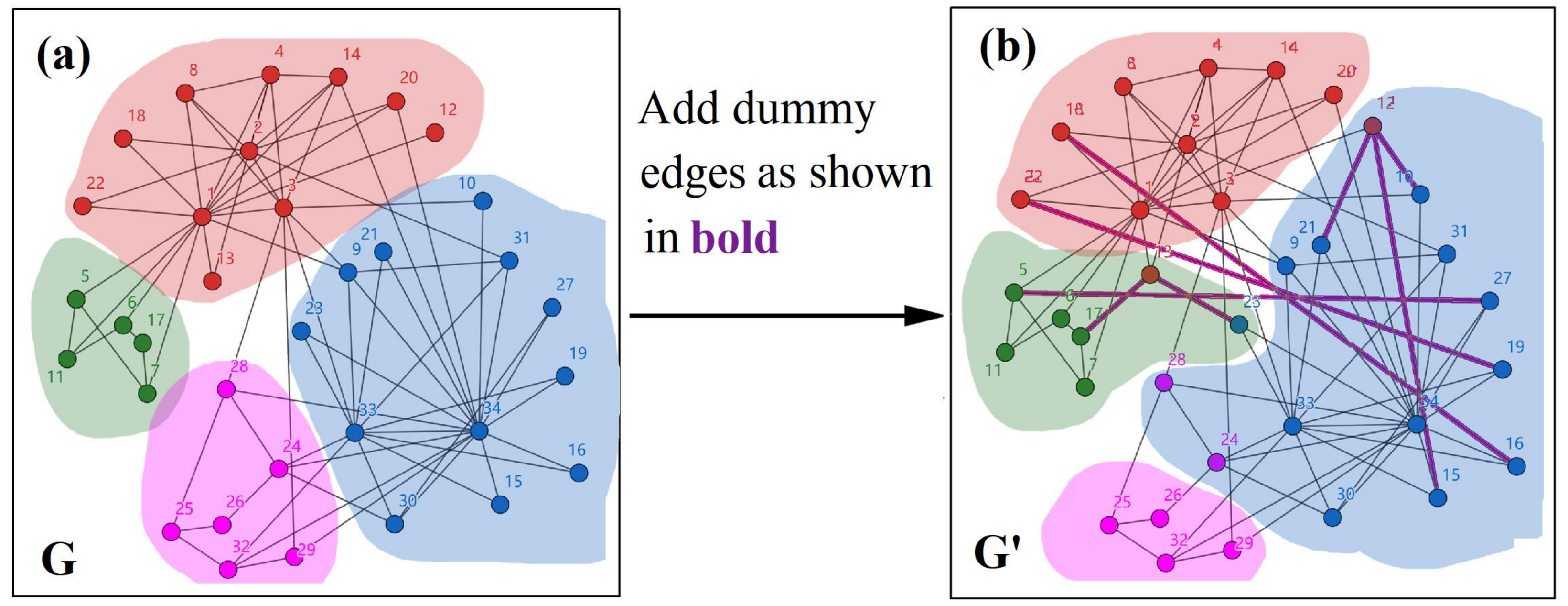
Background

The disclosure of the users' community affiliations leads to privacy leak. This raises the problem of ***community structure deception (CSD)***, which asks for ways to minimally modify the network so that a given community structure maximally hides itself from community detection algorithms.



Problem Formulation

- A social network $\longrightarrow$ An undirected connected graph $G = (V, E)$.
- The community structure $\longrightarrow$ a partition $P = \{X_1, X_2, \cdots, X_L\}$ of V.
- obfuscating $P \longrightarrow$ by **adding a number of "dummy edges"**.



We measure the similarity between P and P' by three metrics:

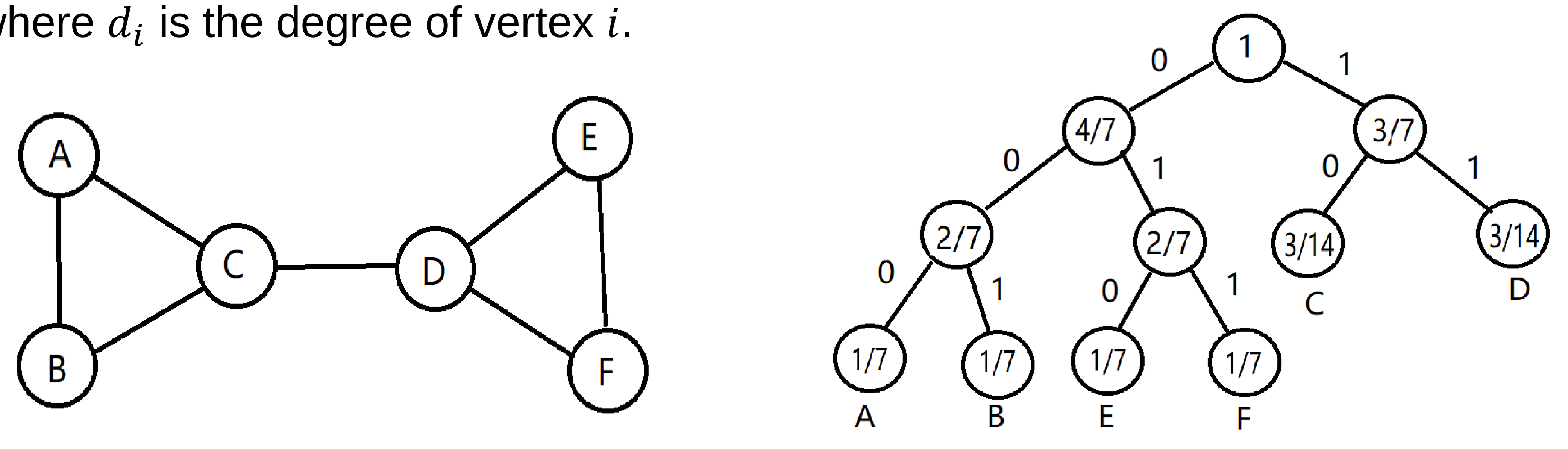
- $\boldsymbol{J(P, P')}$: J is ***jaccard index***
- $\boldsymbol{D(P, P')}$: D is ***normalized mutual information***
- $\boldsymbol{R(P, P')}$: R is the ***recall***

Residual Entropy-based CSD

Definition 1. (Shannon entropy) $\mathcal{H}(G)$ captures the average number of bits needed to encode the n vertices in a lossless way:

$$\mathcal{H}(G) := -\sum_{i=1}^{|V|} \frac{d_i}{2|E|} \log_2 \frac{d_i}{2|E|},$$

where d_i is the degree of vertex i.

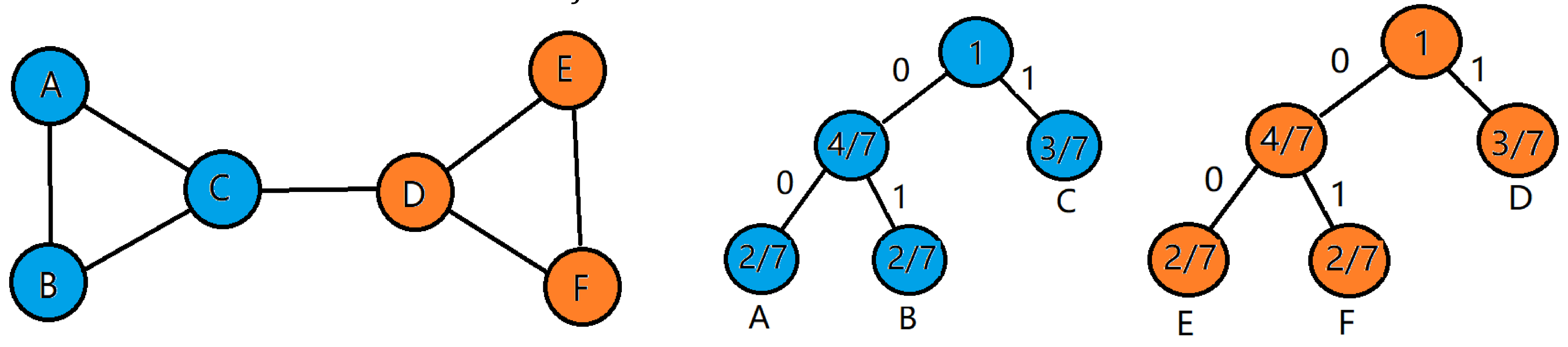


A: 000 B: 001 C: 10 D: 11 E: 010 F: 011 $\boldsymbol{\mathcal{H}(G) = 5.11}$

Definition 2. The structural entropy of $\boldsymbol{G}$ relative to $\boldsymbol{P}$ is:

$$\mathcal{H}_P(G) := -\sum_{j=1}^{L} \frac{v_j}{2|E|} \mathcal{H}(G \upharpoonright X_j) - \frac{g_j}{2|E|} \log_2 \frac{v_j}{2|E|},$$

where the subgraph **G $\upharpoonright$ $\boldsymbol{X_j}$ is induced by $\boldsymbol{X_j}$** and g_j denotes the number of edges with exactly one end point in X_j.



A: 00 B: 01 C: 1 D: 1 E: 00 F: 01 $\boldsymbol{\mathcal{H}_P(G)}$ **=3.40**

Definition 3. The normalized residual entropy of $\boldsymbol{P}$ is

$$\rho_P(G) := (\mathcal{H}(G) - \mathcal{H}_P(G))/\mathcal{H}(G).$$

REM: Algorithm and optimization

Theorem 1. There exists a critical non-edge $\{u, v\}$ is RE-minimizing.
Theorem 2. Examining the critical edges implements in $\boldsymbol{\mathcal{O}(L|V|)}$.

Algorithm 1: An efficient REM deceptor

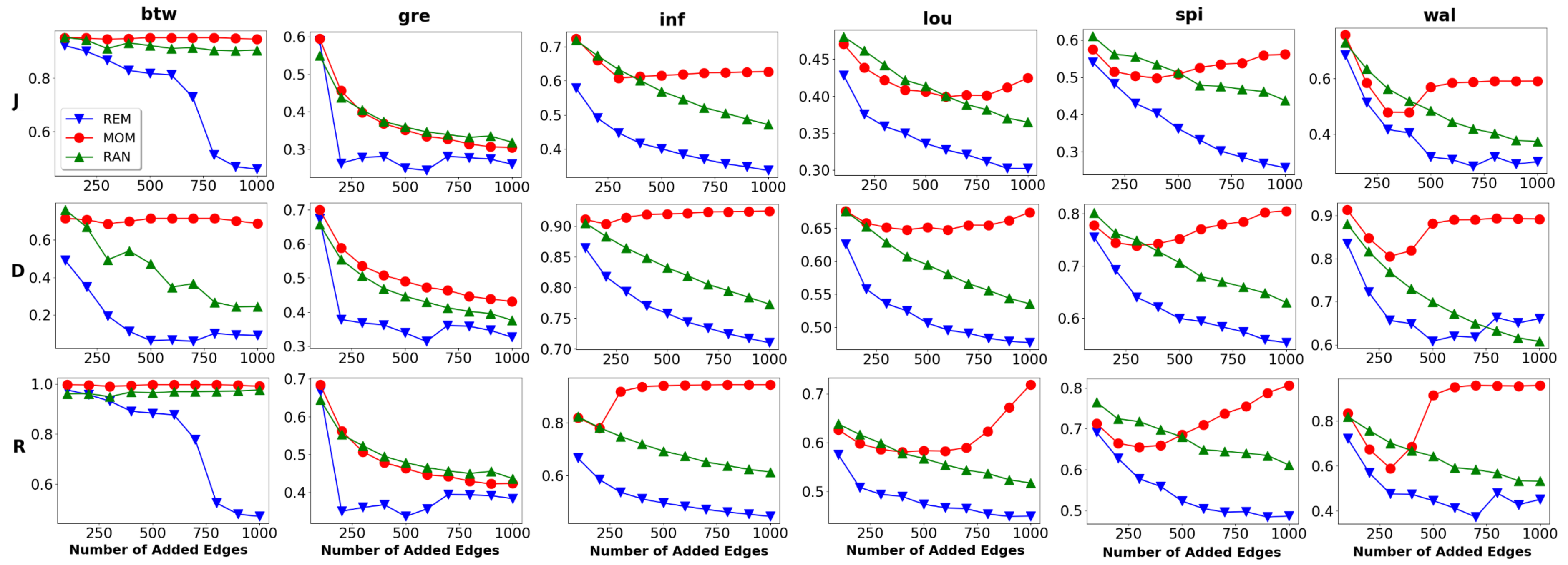
Input: Graph $G = (V, E)$, $\mathcal{P} = \{X_1, X_2, \ldots, X_L\}$
Output: A non-edge $\{u^*, v^*\}$

Initialize $\rho^* \leftarrow 1$;
for $s \leftarrow 1$ *to* L *and* $t \leftarrow s$ *to* L **do**
  if $X_s \times X_t$ *contains no non-edge* **then**
    continue;
  for *all critical non-edge* $\{u, v\}$ *in* $X_s \times X_t$ **do**
    Set $\rho_{u,v} \leftarrow (\mathcal{H}(G \oplus \{u, v\}) - \mathcal{H}_{\mathcal{P}}(G \oplus \{u, v\}))/\mathcal{H}(G \oplus \{u, v\})$;
    if $\rho_{u,v} < \rho^*$ **then**
      Set $\rho^* \leftarrow \rho_{u,v}$, $u^* \leftarrow u$, and $v^* \leftarrow v$;
return $\{u^*, v^*\}$;

A crude implementation runs in $\mathcal{O}(|V|^2)$ time. We present an $\mathcal{O}(L|V|)$-implementation by only examining the critical edges.

Experiments

Dataset: 9 real-world networks. **Adversary:** $\{btw, gre, inf, lou, spi, wal\}$.
Evaluation: 3 metrics, $\{J, D, R\}$. **Benchmark:** $\{MOM, RAN\}$.



The preservation of the data after applying REM:

Dataset	$\lvert E'\rvert$	Jaccard	Clustering coefficient	Mean shortest path length	10% Pagerank	10% Betweenness
Dol	10	$1 \to 0.44$	$0.308 \to 0.298$	$3.357 \to 2.996$	$1 \to 0.833$	$1 \to 0.833$
Jaz	250	$1 \to 0.48$	$0.520 \to 0.498$	$2.23 \to 2.070$	$1 \to 0.895$	$1 \to 0.842$
Eml	100	$1 \to 0.39$	$0.166 \to 0.166$	$3.606 \to 3.577$	$1 \to 0.991$	$1 \to 0.982$
PGP	400	$1 \to 0.45$	$0.378 \to 0.377$	$7.485 \to 7.279$	$1 \to 0.979$	$1 \to 0.930$
CAI	1000	$1 \to 0.49$	$0.007 \to 0.007$	$3.875 \to 3.869$	$1 \to 0.983$	$1 \to 0.977$
Bri	1000	$1 \to 0.44$	$0.111 \to 0.111$	$4.858 \to 4.854$	$1 \to 0.993$	$1 \to 0.971$

Conclusions

- Utilize community based structural entropy to the CSD problem
- Propose a residual minimization (REM) algorithm.
- Reduce search space to critical edges to optimize REM.
- validate the performance of our algorithm.