Co-exposure maximization in online social networks

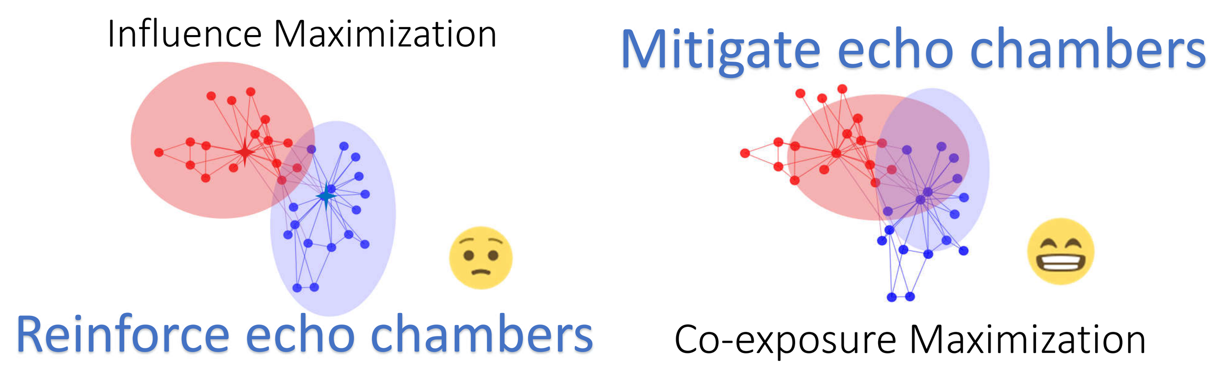
Sijing Tu [1] · Cigdem Aslay [2] · Aristides Gionis [1]

[1] KTH Royal Institute of Technology, Sweden · [2] Aarhus University, Denmark

Motivation

Traditional **viral-marketing** campaigns:

- Aim at reaching out to the maximum number of users;
- Prioritize revenue, thus, relevance;
- Reinforce users' existing viewpoints;
- Result in biased and imbalanced campaigning, leading to echo chambers and polarization.



Main assumptions:

- A centralized authority (host) responsible for allocating seed nodes to campaigns, e.g., Facebook, Twitter;
- Two campaigns supporting the opposing side of a controversial social issue;
- Propagation of campaigns follow the independent cascade (IC) model.

Allocation problem:

- Strategically allocate seed users to campaigns to maximize the expected number of users who are co-exposed to both campaigns.

Problem Formulation

Problem setting

- A directed social network $G = (V, E)$, with two opposing campaigns, denoted by r and b;
- Campaign-specific propagation probabilities p^r_{uv} and p^b_{uv} for all graph edges $(u, v) \in E$;
- Campaigners have seed set budgets: the social-metwork host assigns seed sets S_r and S_b with at most k_r and k_b seed nodes, respectively.

Possible-world semantics

- We define a *directed edge-colored multigraph* $\tilde{G} = (V, \tilde{E}, \tilde{p})$, for any possible world $w \sqsubseteq \tilde{G}$:

$$\Pr[w] = \prod_{i \in \{r,b\}} \prod_{(u,v)_i \in w} p^i_{uv} \prod_{(u,v)_i \in \tilde{E} \setminus w} (1 - p^i_{uv});$$

- Let $I_w(S_r)$ and $I_w(S_b)$ denote the set of nodes reachable from S_r and S_b, respectively, in a possible world w.
- Expected number of users co-exposed to both campaigns is defined as

$$\mathbb{E}[C(S_r, S_b)] = \sum_{w \sqsubseteq \tilde{G}} \Pr[w] |I_w(S_r) \cap I_w(S_b)|.$$

> **Problem** (Co-Exposure Maximization (CoEM)) *Given two positive integers k_r and k_b, find two disjoint seed sets S_r and S_b, such that $|S_r| \le k_r$ and $|S_b| \le k_b$ and $\mathbb{E}[C(S_r, S_b)]$ is maximized.*

CoEM is **NP**-hard to approximate within $1 - \frac{1}{e} + o(1)$ (reduction from Maximum Coverage);

The objective function $\mathbb{E}[C(S_r, S_b)]$ is not (bi-)submodular, and it can have submodularity ratio of 0 in certain problem instances.

Approximation Algorithm

- *set-of-pairs* **system** $(\mathcal{E}, \mathcal{I})$:
  - $(\mathcal{E}, \mathcal{I})$ is a set system, where $\mathcal{E}$ is a set of all ordered pairs of nodes, $\mathcal{I}$ is a collection of subsets of $\mathcal{E}$;
  - Define $X_r = \bigcup\{r \mid (r, b) \in X\}$ and $X_b = \bigcup\{b \mid (r, b) \in X\}$;
  - For any set $X \in \mathcal{I}$, the following conditions hold: (i) $|X_r| \le k_r$; (ii) $|X_b| = |X| \le k_b$; (iii) $X_r \cap X_b = \emptyset$; and (iv) $|\bigcup\{b \mid (r_0, b) \in X\}| \le \lceil \frac{k_b}{k_r} \rceil$, for each $r_0 \in X_r$.
- $(\mathcal{E}, \mathcal{I})$ is a $2\lceil \frac{k_b}{k_r} \rceil$-system.
- Define the function $f(X) = |I(X_r) \cap I(X_b)|$; an equivalent univariate formulation of CoEM is:

$$\max_{X \in \mathcal{I}} \quad \mathbb{E}[f(X)].$$

- Define function $g(X) = |\cup_{(r,b) \in X} (I(r) \cap I(b))|$: $\mathbb{E}[g]$ is submodular and monontune. The greedy algorithm provides an approximation guarantee $\left(1 + 2\lceil \frac{k_b}{k_r} \rceil\right)^{-1}$ [3];
- $\mathbb{E}[f(X)] \le k_r \mathbb{E}[g(X)]$ for any $X \in \mathcal{I}$;
- It follows that the greedy algorithm for $g(X)$ is an approximation algorithm for the CoEM problem with guarantee $\left((1 + 2\lceil \frac{k_b}{k_r} \rceil) k_r\right)^{-1}$.

> Pairs-Greedy($G = (V, E, p)$, $(\mathcal{E}, \mathcal{I})$.)
> - **Initialize:** $\mathcal{X}^G \leftarrow \emptyset$;
> - While $\mathcal{E} \ne \emptyset$
>   - $y = \arg\max_{x: X^G \cup \{x\} \in \mathcal{I}} \mathbb{E}[g(X^G \cup \{x\})] - \mathbb{E}[g(X^G)]$
>   - $\mathcal{E} \leftarrow \mathcal{E} \setminus \{y\}$
>   - $X^G = X^G \cup \{y\}$
> - Return X^G

Fast Algorithm

It is **#P-hard** to compute $\mathbb{E}[g(X)]$ for any given X.

Sample random RRP-sets (generalizing *reverse-reachable* sets [2]):

- A random RRP-set R in possible world w is defined as: $R = \{(r, b) : v \in I_w(r) \cap I_w(b)\}$.

Let $\mathcal{R}$ be a collection of RRP-sets, define $F_{\mathcal{R}}(X) = \sum_{R \in \mathcal{R}} \mathbb{1}[R \cap X \ne \emptyset]/|\mathcal{R}|$. Then:

- $\mathbb{E}[g(X)] = n\, \mathbb{E}[F_{\mathcal{R}}(X)]$ with randomness in $v \sim V$ and $w \sim \tilde{G}$.
- We can estimate $\mathbb{E}[g(X)]$ by estimating $\mathbb{E}[F_{\mathcal{R}}(X)]$.

Let $\mathcal{I}_{base} \subseteq \mathcal{I}$ be the set of maximal independent sets of $(\mathcal{E}, \mathcal{I})$ and let $\lambda = 4n/\epsilon^2(\epsilon/3 + 2)(\ell \ln n + \ln 2 + \ln|\mathcal{I}_{base}|)$.

Theorem. *Assume $\mathcal{R}$ is such that $|\mathcal{R}| \ge \lambda/\mathrm{OPT}$. Then, it holds $|nF_{\mathcal{R}}(X) - \mathbb{E}[g(X)]| < \frac{\epsilon}{2}\mathrm{OPT}$, for any $X \in \mathcal{I}_{base}$, with probability at least $1 - n^{-\ell}/|\mathcal{I}_{base}|$, and the algorithm* RR-Pairs-Greedy *returns an approximate solution to the problem* CoEM *with guanrantee $\left((1 + 2\lceil \frac{k_b}{k_r} \rceil)^{-1} k_r^{-1} - \epsilon\right)$, with probability at least $1 - n^{-\ell}$.*

- To ensure $|\mathcal{R}| \ge \lambda/\mathrm{OPT}$, we estimate a lower bound of OPT using martingale theory [1, 5]

Adaptive estimation of OPT.

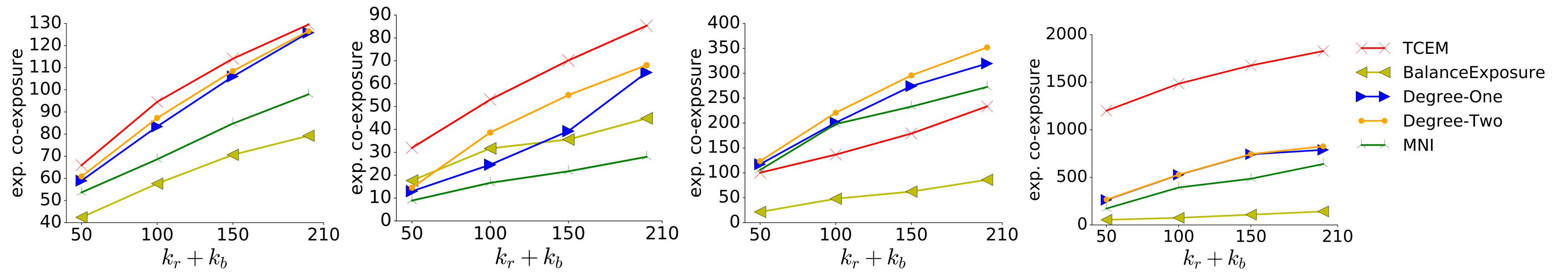
- For the i-th iteration, define $y = n/2^i$, and $\theta_i = \epsilon_2^{-2}(2\epsilon_2/3 + 2)(\ell \ln n + \ln \log_2 n + \ln|\mathcal{I}_{base}|)n/y$;
- Execute algorithm RR-Pairs-Greedy on a sample of size θ_i: if $nF_{\mathcal{R}}(\tilde{X}^G_i) \ge (1 + \epsilon_2)\, y$, then set $LB = \frac{nF_{\mathcal{R}}(\tilde{X}^G_i)}{1+\epsilon_2}$.

Theorem. *With probability at least $1 - n^{-\ell}$, algorithm* Sampling *returns a sample $\mathcal{R}$ such that $|\mathcal{R}| \ge \lambda/\mathrm{OPT}$.*

> RR-Pairs-Greedy($\mathcal{R}$, $(\mathcal{E}, \mathcal{I})$)
> - **Initialize:** $X \leftarrow \emptyset$;
> - $x = \arg\max_{x:\{x\} \cup X \in \mathcal{I}} F_{\mathcal{R}}(X \cup \{x\}) - F_{\mathcal{R}}(X)$
> - While $x \ne \emptyset$
>   - $X = X \cup \{x\}$
>   - $x = \arg\max_{x:\{x\} \cup X \in \mathcal{I}} F_{\mathcal{R}}(X \cup \{x\}) - F_{\mathcal{R}}(X)$
> - Return X

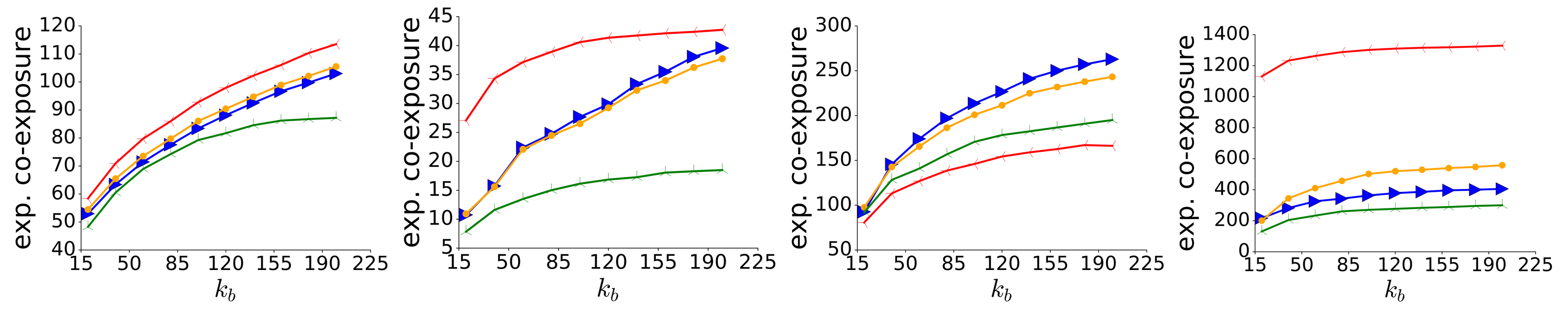
> Sampling($\tilde{G}$, λ, β, ϵ_2, $\tilde{I}$)
> - **Initialize:** $\mathcal{R} \leftarrow \emptyset$, $LB \leftarrow LB_0$;
> - for $i = 1, \dots, \log_2 n - 1$
>   - $y \leftarrow n/2^i$, $\theta_i = \frac{\beta}{y}$
>   - while $|\mathcal{R}| \le \theta_i$
>   - $\mathcal{R} \leftarrow \mathcal{R} \cup$ GenerateRRP-Set
>   - $\tilde{X}_i \leftarrow$ RR-Pairs-Greedy$(\mathcal{R}, \tilde{I})$
>   - if $n\, F_{\mathcal{R}}(\tilde{X}_i) \ge (1 + \epsilon_2)\, y$,
>   - $LB \leftarrow \frac{n\, F_{\mathcal{R}}(\tilde{X}_i)}{1+\epsilon_2}$, break
> - $\theta \leftarrow \lambda/LB$
> - while $|\mathcal{R}| \le \theta$
>   - $\mathcal{R} \leftarrow \mathcal{R} \cup$ GenerateRRP-Set
> - Return $\mathcal{R}$

Experiments



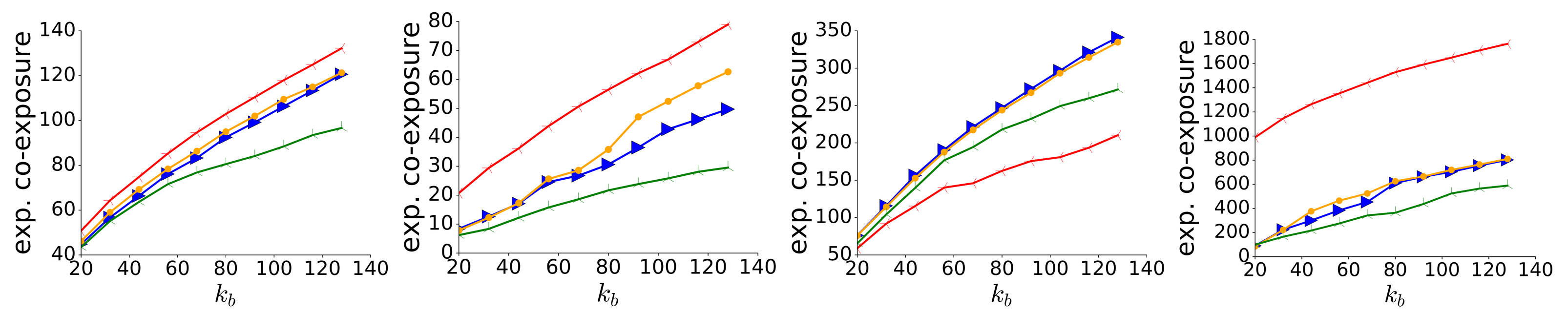
(a) Last.FM_het (b) NetHEPT_het (c) WikiVote_wc (d) Flixster_hom

Vary $k_r + k_b$



(a) Last.FM_het (b) NetHEPT_het (c) WikiVote_wc (d) Flixster_hom

Fix $k_r = 20$



(a) Last.FM_het (b) NetHEPT_het (c) WikiVote_wc (d) Flixster_hom

Fix $\tau = 2$

Datasets

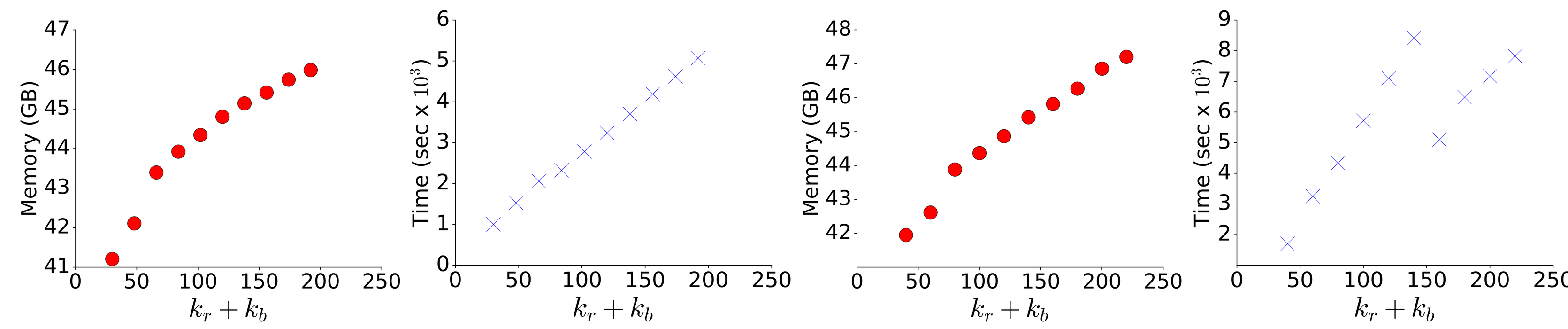
- Flixster, Last.FM, NetHEPT, WikiVote.

Methods to assign independent-cascade parameters

- *weighted-cascade model* (prefix _wc);
- homogeneous (_hom) and heterogeneous (_het) *trivalency model* randomly drawing from $\{0.1, 0.01, 0.001\}$;

Baselines The first two baselines consider the nodes in decreasing order of out-degree.

- Degree-One: k_r seeds are assigned to one campaign and k_b to the other;
- Degree-Two: seeds are assigned in a round-robin fashion;
- *Maximum neighborhood intersection* (MNI) solves $\arg\max_{X \in \mathcal{I}} |N'(X_r) \cap N'(X_b)|$, where $N'(X_i)$ is the union of the nodes in X_i and their out-neighbors;
- BalanceExposure is the greedy method proposed by Garimella et al. [4], which we use without initial seeds.



(a) Memory ($\tau = 2$) (b) Time ($\tau = 2$) (c) Memory ($k_r = 20$) (d) Time ($k_r = 20$)

Results

- TCEM outperforms the baselines in all the datasets under the homogeneous and heterogeneous propagation models;
- For the weighted-cascade model, the local algorithms that use out-degree information may perform better than TCEM as observed in WikiVote dataset, although this behavior is not robust;
- Memory and time increase linearly, or better.

References

[1] C. Aslay, A. Matakos, E. Galbrun, and A. Gionis. Maximizing the diversity of exposure in a social network. In *ICDM*, 2018.

[2] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. In *SODA*, 2014.

[3] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 2011.

[4] K. Garimella, A. Gionis, N. Parotsidis, and N. Tatti. Balancing information exposure in social networks. In *NeurIPS*, 2017.

[5] Y. Tang, Y. Shi, and X. Xiao. Influence maximization in near-linear time: A martingale approach. In *SIGMOD*, 2015.