Balancing information exposure in social networks

KIRAN GARIMELLA · ARISTIDES GIONIS
NIKOS PAROTSIDIS · NIKOLAJ TATTI

Aalto University · HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY · Università di Roma Tor Vergata

MOTIVATION

Social media play an important role in the way that people receive news :

- It is estimated that 62% of adults in the US get their news on social media [3].
- Social media provide searching, personalization, and recommendations.

Criticism: social media amplify echo chambers and filter bubbles: users get less exposure to conflicting viewpoints and are isolated in their own informational bubble. This phenomenon is more acute for controversial topics [2].

Need for mechanisms to "burst" the filter bubbles. One approach is to convince a small set of "key" individuals to post information favoring one topic.

ASSUMPTIONS

- The information propagates according to the independent-cascade model.
- Two opposing campaigns, with an initial seeds I_1 and I_2, not necessarily distinct.
- A user is exposed to campaign i via diffusion from the set of seeds I_i, $i = 1, 2$.
- Two cascade settings. **Heterogeneous**: the activation probability of each edge is dependent on the campaign. **Correlated**: the edges are activated with the same probability for the two campaigns.

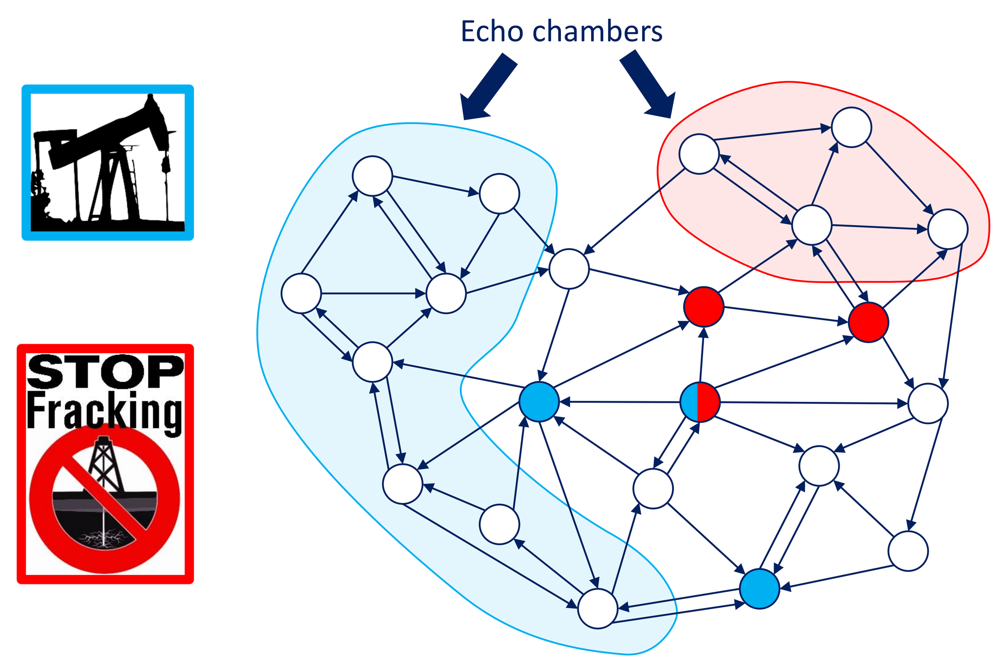
PROBLEM FORMULATION

Problem (BALANCE)

- Given a social network $G = (V, E)$, two sets I_1 and I_2 of initial seeds of the two campaigns, and a budget k.
- Let $r_i(S)$ be the random variable describing the set of exposed nodes to campaign i from the set of seeds S.
- Find two sets S_1 and S_2, where $|S_1| + |S_2| \leq k$ maximizing

$$\Phi(S_1, S_2) = \mathrm{E}[|V \setminus (r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2))|] .$$

- $\Phi(S_1, S_2)$ is the expected number of vertices that are either exposed by both campaigns or remain oblivious to both campaigns.



OUR RESULTS

Complexity:

- The BALANCE problem is **NP**-hard.
- Φ is *not* submodular.

Decomposition of the objective function:

Let $X = \{r_1(I_1) \cup r_2(I_2)\}$, $Y = V \setminus X$. The objective function can be written as:

$$\Phi(S_1, S_2) = \Omega(S_1, S_2) + \Psi(S_1, S_2)$$

$$\Omega(S_1, S_2) = \mathrm{E}[|X \setminus (r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2))|]$$

$$\Psi(S_1, S_2) = \mathrm{E}[|Y \setminus (r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2))|]$$

Proposition 1. *The function $\Omega(S_1, S_2)$ is monotone and submodular.*

Cover (greedy maximizing Ω):

- Initialize $S_1 \leftarrow S_2 \leftarrow \emptyset$
- While $|S_1| + |S_2| < k$
  - $s_1 \leftarrow \arg\max_s \Omega(S_1 \cup \{s\}, S_2)$
  - $s_2 \leftarrow \arg\max_s \Omega(S_1, S_2 \cup \{s\})$
  - If $\Omega(S_1 \cup \{s_1\}, S_2) \geq \Omega(S_1, S_2 \cup \{s_2\})$ then $S_1 \leftarrow S_1 \cup s_1$, else $S_2 \leftarrow S_2 \cup s_2$
- Return S_1, S_2

Proposition 2. *Let $\langle S_1^*, S_2^* \rangle$ be the optimal solution maximizing Φ. Let $\langle S_1, S_2 \rangle$ be the solution obtained via the Cover algorithm. Then*

$$\max\{\Phi(S_1, S_2), \Phi(\emptyset, \emptyset)\} \geq \frac{1 - 1/e}{2} \Phi(S_1^*, S_2^*).$$

Hedge (greedy considering adding to both campaigns at each iteration):

- Initialize $S_1 \leftarrow S_2 \leftarrow \emptyset$
- While $|S_1| + |S_2| < k$
  - $c \leftarrow \arg\max_c \Phi(S_1 \cup \{c\}, S_2 \cup \{c\})$
  - $s_1 \leftarrow \arg\max_s \Phi(S_1 \cup \{s\}, S_2)$
  - $s_2 \leftarrow \arg\max_s \Phi(S_1, S_2 \cup \{s\})$
  - add the best option among $\langle c, c \rangle$, $\langle \emptyset, s_1 \rangle$, $\langle s_2, \emptyset \rangle$ to $\langle S_1, S_2 \rangle$ while respecting the budget
- Return S_1, S_2

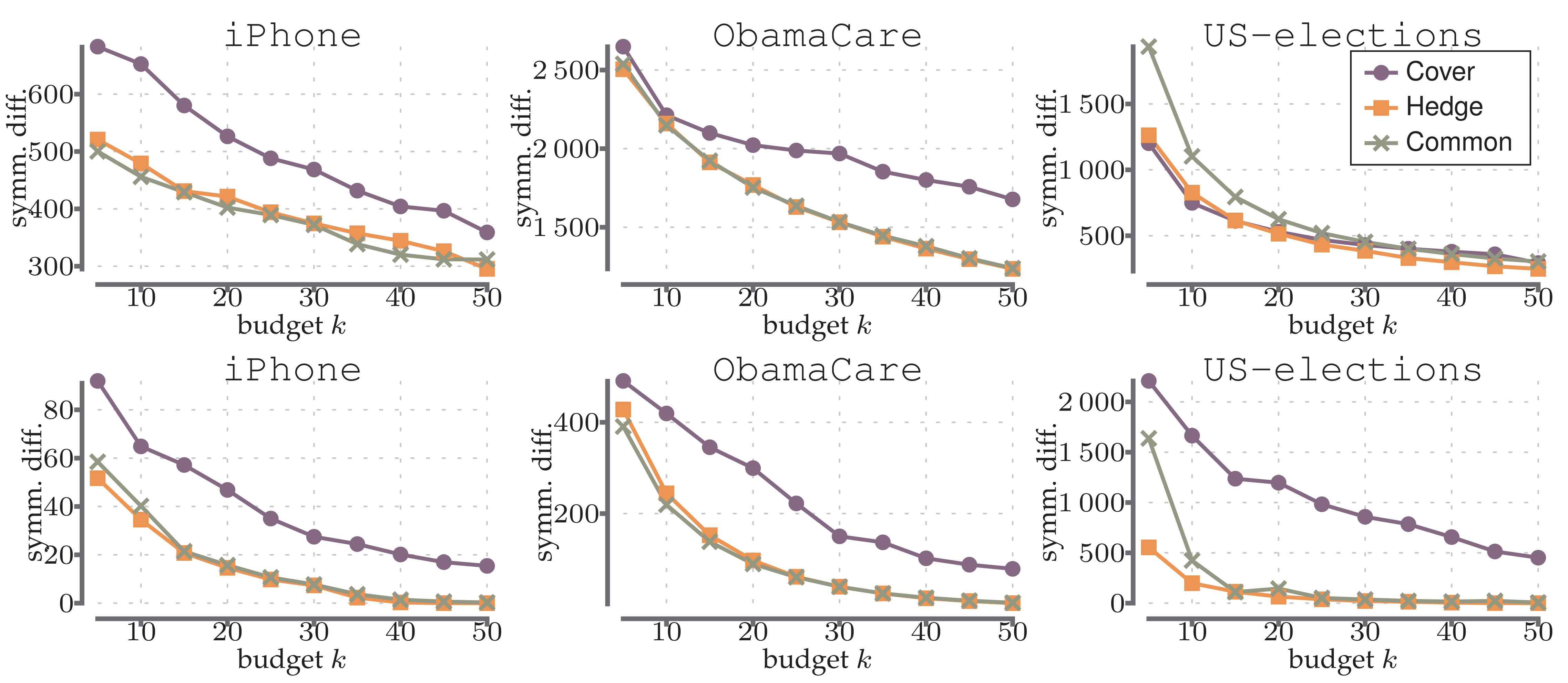
Proposition 3. *Algorithm Hedge achieves a $(1 - 1/e)/2$ approximation for the BALANCE problem in the correlated setting.*

Common (greedy that forces $\Psi = 0$): Greedy algorithm that only adds a common seeds to both campaigns, or adds to a campaign a seed of the opposing campaign. This forces $\Psi(S_1, S_2) = 0$.

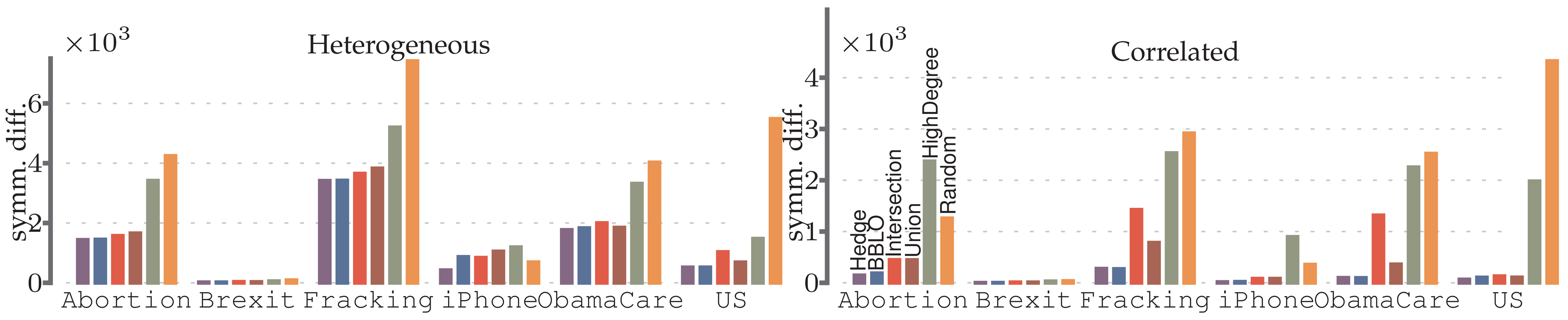
Proposition 4. *Algorithm Common achieves a $(1 - 1/e)/2$ approximation for the BALANCE problem in the correlated setting.*

EXPERIMENTS

Datasets: Experiments on real-world data collected from twitter. We use datasets from six topics with opposing viewpoints, covering politics, policy, and lifestyle.



Baselines: **BBLO** is an adaptation of the framework by Borodin et al. [1]. For the next two heuristics we proceed as follows. Each campaign $i = 1, 2$ selects S_i', $|S_i'| \gg k$, to optimize $r_i(S_i')$. **Union** sets S_1 and S_2 to be the $k/2$ first distinct nodes in $S_1' \cup S_2'$. **Intersection** sets S_1 and S_2 to be the $k/2$ first vertices in $S_1' \cap S_2'$.



REFERENCES

[1] A. Borodin, Y. Filmus, and J. Oren. Threshold models for competitive influence in social networks. In *WINE*, 2010.
[2] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Reducing controversy by connecting oppposing views. In *WSDM*, 2017.
[3] J. Gottfried and E. Shearer. News use across social media platforms 2016. *Pew Research Center*, 2016.