arXiv:1709.01491v2 [cs.SI] 4 Jan 2018

Balancing information exposure in social networks

Kiran Garimella
Aalto University & HIIT
Helsinki, Finland
kiran.garimella@aalto.fi

Aristides Gionis
Aalto University & HIIT
Helsinki, Finland
aristides.gionis@aalto.fi

Nikos Parotsidis
University of Rome Tor Vergata
Rome, Italy
nikos.parotsidis@uniroma2.it

Nikolaj Tatti
Aalto University & HIIT
Helsinki, Finland
nikolaj.tatti@aalto.fi

Abstract

Social media has brought a revolution on how people are consuming news. Beyond the undoubtedly large number of advantages brought by social-media platforms, a point of criticism has been the creation of *echo chambers* and *filter bubbles*, caused by *social homophily* and *algorithmic personalization*.

In this paper we address the problem of *balancing the information exposure* in a social network. We assume that two opposing campaigns (or viewpoints) are present in the network, and that network nodes have different preferences towards these campaigns. Our goal is to find two sets of nodes to employ in the respective campaigns, so that the overall information exposure for the two campaigns is *balanced*. We formally define the problem, characterize its hardness, develop approximation algorithms, and present experimental evaluation results.

Our model is inspired by the literature on *influence maximization*, but we offer significant novelties. First, *balance* of information exposure is modeled by a *symmetric difference* function, which is neither monotone nor submodular, and thus, not amenable to existing approaches. Second, while previous papers consider a setting with selfish agents and provide bounds on *best response* strategies (i.e., move of the last player), we consider a setting with a centralized agent and provide bounds for a global objective function.

1 Introduction

Social-media platforms have revolutionized many aspects of human culture, among others, the way people are exposed to information. A recent survey estimates that 62% of adults in the US get their news on social media [20]. Despite providing many desirable features, such as, searching, personalization, and recommendations, one point of criticism is that social media amplify the phenomenon of *echo chambers* and *filter bubbles*: users get less exposure to conflicting viewpoints and are isolated in their own informational bubble. This phenomenon is contributed to social homophily and algorithmic personalization, and is more acute for controversial topics [2, 12, 13, 16, 18].

In this paper we address the problem of reducing the filter-bubble effect by balancing information exposure among users. We consider social-media discussions around a topic that are characterized by two or more *conflicting viewpoints*. Let us refer to these viewpoints as *campaigns*. Our approach follows the popular paradigm of influence propagation [24]: we want to select a small number of seed users for each campaign so as to maximize the number of users who are *exposed to both campaigns*. In contrast to existing work on competitive viral marketing, we do not consider the problem of finding an optimal *selfish strategy* for each campaign separately. Instead we consider a

certalized agent responsible for balancing information exposure for the two campaings. Consider the following motivating examples.

Example 1: Prominent social-media companies, like Facebook and Twitter, have been called to act as arbiters so as to prevent ideological isolation and polarization in the society. The motivation for companies to assume this role could be for improving their public image or due to government policies.[1] Consider a controversial topic being discussed in social-media platform X, which has led to polarization and filter bubbles. Platform X has the ability to algorithmically detect such bubbles [16], identify the influential users on each side, and estimate the influence among users [14, 19]. As part of a new filter-bubble bursting service, platform X would like to disseminate two high-quality and thought-provoking dueling op-eds, articles, one for each side, that present the arguments of the other side in a fair manner. Assume that X is interested in following a viral-marketing approach. Which users should X target, for each of the two articles, so that people in the network are informed in the most balanced way?

Example 2: Government organization Y is initiating a programme to help assimilate foreigners who have newly arrived in the country. Part of the initiative focuses on bringing the communities of foreigners and locals closer in social media. Organization Y is interested in identifying individuals who can help spreading news of one community into the other.

From the technical standpoint, we consider the following problem setting: We assume that information is propagated in the network according to the *independent-cascade model* [24]. We assume that there are two opposing campaigns, and for each one there is a set of initial seed nodes, I_1 and I_2, which are not necessarily distinct. Furthermore, we assume that the users in the network are exposed to information about campaign i via diffusion from the set of seed nodes I_i. The diffusion in the network may occur with independent or correlated probabilities for the two campaigns; we consider both settings to which we are referring as *heterogeneous* or *correlated*.

The objective is to recruit two additional sets of seed nodes, S_1 and S_2, for the two campaigns, with $|S_1| + |S_2| \leq k$, for a given budget k, so as to maximize the expected number of balanced users, i.e., the users who are exposed to information from both campaigns (or from none!).

We formally define the problem of balancing information exposure and we show that it is $\mathbf{NP}$-hard. We develop different approximation algorithms for the different settings we consider, as well as heuristic variants of the proposed algorithm. We experimentally evaluate our methods, on several real-world (and realistic) datasets, collected from twitter, for different topics of interest.

Although our approach is inspired by the large body of work on information propagation, and resembles previous problem formulations for competitive viral marketing, there are significant differences and novelties. In particular:

- This is the first paper to address the problem of *balancing information exposure* and *breaking filter bubbles*, using the information-propagation methodology.
- The objective function that best suits our problem setting is related to the *size of the symmetric difference* of users exposed to the two campaigns. This is in contrast to previous settings that consider functions related to the *size of the coverage* of the campaigns.
- As a technical consequence of the previous point, our objective function is neither *monotone* nor *submodular* making our problem more challenging. Yet we are able to analyze the problem structure and provide algorithms with approximation guarantees.
- While most previous papers consider selfish agents, and provide bounds on *best-response* strategies (i.e., move of the last player), we consider a centralized setting and provide bounds for a global objective function.

We note that our datasets and implementation is publicly available.[2]

[1] For instance, Germany is now fining Facebook for the spread of fake news.

[2] `https://users.ics.aalto.fi/kiran/BalanceExposure/`

2 Related Work

Detecting and breaking filter bubbles. Several studies have observed that users in online social networks prefer to associate with like-minded individuals and consume agreeable content. This phenomenon leads to *filter bubbles*, *echo chambers* [36, 35], and to online polarization [1, 2, 5, 12, 16, 22, 30]. Once these filter bubbles are detected, the next step is to try to overcome them. One way to achieve this is by making recommendations to individuals of opposing viewpoints. This idea has been explored, in different ways, by a number of studies in the literature [17, 26, 27, 31, 39]. However, all previous studies address the problem of breaking filter bubbles by the means of *content recommendation*. To the best of our knowledge, this is the first paper that considers an *information diffusion* approach.

Information diffusion. Following a large body of work, we model diffusion using the *independent-cascade model* [24]. In the basic model a single item propagates in the network. An extension is when multiple items propagate simultaneously. All works that study optimization problems in the case of multiple items, consider that items *compete* for being adopted by users. In other words, every user adopts at most one of the existing items and participates in at most one cascade.

Myers and Leskovec [32] argue that spreading processes may either cooperate or compete. Competing contagions decrease each other's probability of diffusion, while cooperating ones help each other in being adopted. They propose a model that quantifies how different spreading cascades interact with each other. Carnes et al. [11] propose two models for competitive diffusion. Subsequently, several other models have been proposed [4, 7, 9, 15, 23, 25, 29, 38].

Most of the work on *competitive information diffusion* consider the problem of selecting the best k seeds for one campaign, for a given objective, in the presence of competing campaigns [6, 10, 34]. Bharathi et al. [6] show that, if all campaigns but one have fixed sets of seeds, the problem for selecting the seeds for the last player is submodular, and thus, obtain an approximation algorithm for the strategy of the last player. Game theoretic aspects of competitive cascades in social networks, including the investigation of conditions for the existence of Nash equilibrium, have also been studied [3, 21, 37].

The work that is most related to ours, in the sense of considering a *centralized authority*, is the one by Borodin et al. [8]. They study the problem where multiple campaigns wish to maximize their influence by selecting a set of seeds with bounded cardinality. They propose a centralized mechanism to allocate sets of seeds (possibly overlapping) to the campaigns so as to maximize the social welfare, defined as the sum of the individual's selfish objective functions. One can choose any objective functions as long as it is submodular and non-decreasing. Under this assumption they provide strategyproof (truthful) algorithms that offer guarantees on the social welfare. Their framework applies for several competitive influence models. In our case, the number of balanced users is not submodular, and so we do not have any approximation guarantees. Nevertheless, we can use this framework as a heuristic baseline, which we do in the experimental section.

3 Problem Definition

Preliminaries: We start with a directed graph $G = (V, E, p_1, p_2)$ representing a social network. We assume that there are two distinct campaigns that propagate through the network. Each edge $e = (u, v) \in E$ is assigned two probabilities, $p_1(e)$ and $p_2(e)$, representing the probability that a post from vertex u will propagate (e.g., it will be reposted) to vertex v in the respective campaigns.

Cascade model: We assume that information on the two campaigns propagates in the network following the independent-cascade model [24]. For instance, consider the propagation of the first campaign. The procedure for the second campaign is analogous. We assume that there exists a set of seeds I_1 from which the propagation process begins. These are vertices in the network that support the campaign and are active in generating content in favor of the campaign. Propagation in the independent-cascade model proceeds in rounds. At each round, there exists a set of active vertices A_1 (initially, $A_1 = I_1$), where each vertex $u \in A_1$ attempts to activate each vertex $v \notin A_1$, such that $(u, v) \in E$, with probability $p_1(u, v)$. If the propagation attempt from a vertex u to a vertex v is successful, we say that v propagates the first campaign. At the end of each round, A_1 is set to be the set of vertices that propagated the campaign during the current round.

Given a seed set S, we write $r_1(S)$ and $r_2(S)$ for the vertices that are reached from S using the aforementioned cascade process, for the respective campaign. Note that since this process is random, both $r_1(S)$ and $r_2(S)$ are random variables. Computing the expected number of active vertices is a **#P**-hard problem, however, we can approximate it within an arbitrary small factor ϵ, with high probability, via Monte-Carlo simulations. Due to this obstacle, all approximation algorithms that evaluate an objective function over diffusion processes immediately reduce their approximation by an additive ϵ. Throughout this work we avoid repeating this fact for the sake of simplicity of the notation.

Heterogeneous vs. correlated propagations: Our model needs also to specify how the propagation on the two campaigns interact with each other. We consider two settings: In the first setting, we assume that the campaign messages propagate independently of each other. Given an edge $e = (u, v)$, the vertex v is activated on the first campaign with probability $p_1(e)$, given that vertex u is activated on the first campaign. Similarly, v is activated on the second campaign with probability $p_2(e)$, given that u is activated on the second campaign. We refer to this setting as *heterogeneous*.[3] In the second setting we assume that $p_1(e) = p_2(e)$, for each edge e. We further assume that the *coin flips* for the propagation of the two campaigns are totally correlated. Namely, consider an edge $e = (u, v)$, where u is reached by either or both campaigns. Then with probability $p_1(e)$, *any* campaign that has reached u, will also reach v. We refer to this second setting as *correlated*.

Note that in both settings, a vertex may be active by *none*, *either*, or *both* campaigns. This is in contrast to most existing work in competitive viral marketing, where it is assumed that a vertex can be activated by *at most one* campaign. The intuition is that in our setting activation means merely passing a message or posting an article, and it does not imply full commitment to the campaign. We also note that the heterogeneous setting is more *realistic* than the correlated, however, we also study the correlated model as it is mathematically simpler.

Problem definition: We are now ready to state our problem for *balancing information exposure* (BALANCE). Given a directed graph, initial seed sets for both campaigns and a budget, we ask to find additional seeds that would balance the vertices. More formally:

Problem 3.1 (BALANCE)**.** *Let $G = (V, E, p_1, p_2)$ be a directed graph, and two sets I_1 and I_2 of initial seeds of the two campaigns. Assume that we are given a budget k. Find two sets S_1 and S_2, where $|S_1| + |S_2| \leq k$ maximizing*

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

The objective function $\Phi(S_1, S_2)$ is the expected number of vertices that are either reached by both campaigns or remain oblivious to both campaigns. Problem 3.1 is defined for both settings, *heterogeneous* and *correlated*. When we need to make explicit the underlying setting we refer to the respective problems by BALANCE-H and BALANCE-C. When referring to BALANCE-H, we denote the objective by Φ_H. Similarly, when referring to BALANCE-C, we write Φ_C. We drop the indices, when we are referring to both models simultaneously.

Computational complexity: As expected, the optimization problem BALANCE turns out to be **NP**-hard for both settings, heterogeneous and correlated. A straightforward way to prove it is by setting $I_2 = V$, so the problems reduce to standard influence maximization. However, we provide a stronger result. Note that instead of maximizing balanced vertices we can equivalently minimize the imbalanced vertices. However, this turns to be a more difficult problem.

Proposition 1. *Assume a graph $G = (V, E, p_1, p_2)$ with two sets I_1 and I_2 and a budget k. It is an **NP**-hard problem to decide whether there are sets S_1 and S_2 such that $|S_1| + |S_2| \leq k$ and $\mathrm{E}[|r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2)|] = 0$.*

Proof. To prove the hardness we will use SET COVER. Here, we are given a universe U and family of sets $C_1, \ldots, C_\ell$, and we are asked to select k sets covering the universe U.

To map this instance to our problem, we first define vertex set V to consist of 3 parts, V_1, V_2 and V_3. The first part corresponds to the universe U. The second part consists of k copies of ℓ vertices, ith vertex in jth copy corresponds to C_i. The third part consists of k vertices b_j. The edges

[3] Although *independent* is probably a better term than *heterogeneous*, we adopt the latter to avoid any confusion with the independent-cascade model.

are as follows: a vertex v in the jth copy, corresponding to a set C_i is connected to the vertices corresponding to the elements in C_i, furthermore v is connected to b_j. We set $p_1 = p_2 = 1$. The initial seeds are $I_1 = \emptyset$ and $I_2 = V_1 \cup V_3$. We set the budget to $2k$.

Assume that there is a k-cover, $C_{i_1}, \ldots, C_{i_k}$. We set

$$S_1 = S_2 = \left\{\text{vertex corresponding to } C_{i_j} \text{ in } j\text{th copy } \right\}.$$

It is easy to see that the imbalanced vertices in I_2 are exposed to the first campaign. Moreover, S_1 and S_2 do not introduce new imbalanced vertices. This makes the objective equals to 0.

Assume that there exists a solution S_1 and S_2 with a zero cost. We claim that $|S_1 \cap (V_1 \cup V_2)| \leq k$. To prove this, first note that $S_1 \cap V_2 = S_2 \cap V_2$, as otherwise vertices in V_2 are left unbalanced. Let $m = |S_1 \cap V_2|$. Since V_3 must be balanced and each vertex in V_2 has only one edge to a vertex in V_3, there at least k vertices in $|S_1 \cap \{V_2 \cup V_3\}|$, that is, we must have $|S_1 \cap V_3| \geq k - m$. Let us write $d_{ij} = |S_i \cap V_j|$. The budget constraints guarantee that

$$d_{11} + d_{12} + d_{22} + d_{13} \leq \sum_{ij} d_{ij} \leq 2k,$$

which can be rewritten as

$$d_{11} + d_{12} \leq 2k - d_{22} - d_{13} \leq 2k - m - (k - m) = k.$$

Construct C as follows: for each $S_1 \cap V_2$, select the set that correponds to the vertex, for each $S_1 \cap V_1$, select any set that contain this vertex (there is always at least one set, otherwise the problem is trivially false). Since V_1 must be balanced, C is a k-cover of U. □

This result holds for both models, even when $p_1 = p_2 = 1$. This result implies that the minimization version of the problem is **NP**-hard, and there is no algorithm with multiplicative approximation guarantee. It also implies that BALANCE-H and BALANCE-C are also **NP**-hard. However, we will see later that we can obtain approximation guarantees for these maximization problems.

4 Greedy algorithms yielding approximation guarantees

In this section we propose three greedy algorithms. The first algorithm yields an approximation guarantee of $(1 - 1/e)/2$ for both models. The remaining two algorithms yield a guarantee for the correlated model only.

Decomposing the objective: Recall that the objective function of the BALANCE problem is $\Phi(S_1, S_2)$. In order to show that this function admits an approximation guarantee, we decompose it into two components. To do that, assume that we are given initial seeds I_1 and I_2, and let us write $X = r_1(I_1) \cup r_2(I_2), Y = V \setminus X$. Here X are vertices reached by any initial seed in the two campaigns and Y are the vertices that are not reached at all. Note that X and Y are random variables. Since X and Y partition V, we can decompose the score $\Phi(S_1, S_2)$ as

$$\begin{aligned}
&\Phi(S_1, S_2) = \Omega(S_1, S_2) + \Psi(S_1, S_2), \quad \text{where} \\
&\Omega(S_1, S_2) = \mathrm{E}[|X \setminus (r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2))|], \\
&\Psi(S_1, S_2) = \mathrm{E}[|Y \setminus (r_1(I_1 \cup S_1) \triangle r_2(I_2 \cup S_2))|].
\end{aligned}$$

We first show that $\Omega(S_1, S_2)$ is monotone and submodular. It is well-known that for maximizing a function that has these two properties under a size constraint, the greedy algorithm computes an $(1 - \frac{1}{e})$ approximate solution [33].

Lemma 2. *$\Omega(S_1, S_2)$ is monotone and submodular.*

Before providing the proof, as a technicality, note that submodularity is usually defined for functions with one argument. Namely, given a universe of items U, we consider functions of the type $f : 2^U \to \mathbb{R}$. However, by taking $U = V \times \{1, 2\}$ we can equivalently write our objectives as functions with one argument, i.e., $\Phi, \Omega, \Psi : 2^U \to \mathbb{R}$.

Proof. The objective counts 3 types of vertices: (*i*) vertices covered by both initial seeds, (*ii*) additional vertices covered by I_1 and S_2, and (*iii*) additional vertices covered by I_2 and S_1. This allows us to decompose the objective as

$$\Omega(S_1, S_2) = \mathrm{E}[|A| + |B| + |C|], \quad \text{where}$$
$$A = r_1(I_1) \cap r_2(I_2), \quad B = (r_1(I_1) \setminus r_2(I_2)) \cap r_2(S_2), \quad C = (r_2(I_2) \setminus r_1(I_1)) \cap r_1(S_1).$$

Note that A does not depend on S_1 and S_2. B grows in size as we add more vertices to S_2, and C grows in size as we add more vertices to S_1. This proves that the objective is monotone.

To prove the submodularity, let us introduce some notation: given a set of edges F, we write $r(S; F)$ to be the set of vertices that can be reached from S via F. This allows us to define

$$A(F_1, F_2) = r(I_1; F_1) \cap r(I_2; F_2),$$
$$B(F_1, F_2) = (r(I_1; F_1) \setminus r(I_2; F_2)) \cap r(S_2; F_2),$$
$$C(F_1, F_2) = (r(I_2; F_2) \setminus r(I_1; F_1)) \cap r(S_1; F_1).$$

The score $\Omega(S_1, S_2)$ can be rewritten as

$$\sum_{F_1, F_2} p(F_1, F_2)(|A(F_1, F_2)| + |B(F_1, F_2)| + |C(F_1, F_2)|),$$

where $p(F_1, F_2)$ is the probability of F_1 being the realization of the edges for the first campaign and F_2 being the realization of the edges for the second campaign.

The first term $A(F_1, F_2)$ does not depend on S_1 or S_2. The second term is submodular as a function of S_2 and does not depend of S_1. The third term is submodular as a function of S_1 and does not depend of S_2. Since any linear combination of submodular function weighted by positive coefficients is also submodular, this completes the proof. □

We are ready to discuss our algorithms.

Algorithm 0: ignore Ψ. Our first algorithm is very simple: instead of maximizing Φ, we maximize Ω, i.e., we ignore any vertices that are made imbalanced during the process. Since Ω is submodular and monotone we can use the greedy algorithm. If we then compare the obtained result with the empty solution, we get the promised approximation guarantee. We refer to this algorithm as Cover.

Proposition 3. *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 greedy algorithm maximizing Ω. Then*

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

Proof. Write $c = 1 - 1/e$. Let $\langle S_1', S_2' \rangle$ be the optimal solution maximizing Ω. Lemma 2 shows that $\Omega(S_1, S_2) \geq c\Omega(S_1', S_2')$.

Note that $\Psi(\emptyset, \emptyset) \geq \Psi(S_1^*, S_2^*)$ as the first term is the average of vertices not affected by the initial seeds. Thus,

$$\begin{aligned}
\Phi(S_1^*, S_2^*) &= \Omega(S_1^*, S_2^*) + \Psi(S_1^*, S_2^*) \leq \Omega(S_1', S_2') + \Psi(S_1^*, S_2^*) \\
&\leq \Omega(S_1', S_2') + \Psi(\emptyset, \emptyset) \leq \Omega(S_1, S_2)/c + \Psi(\emptyset, \emptyset) \\
&\leq \Omega(S_1, S_2)/c + \Psi(\emptyset, \emptyset)/c \\
&\leq (2/c) \max\{\Omega(S_1, S_2), \Psi(\emptyset, \emptyset)\} \\
&\leq (2/c) \max\{\Phi(S_1, S_2), \Phi(\emptyset, \emptyset)\},
\end{aligned}$$

which completes the proof. □

Algorithm 1: force common seeds. Ignoring the Ψ term may prove costly as it is possible to introduce a lot of new imbalanced vertices. The idea behind the second algorithm is to force $\Psi = 0$. We do this by either adding the same seeds to both campaigns, or adding a seed that is covered by an opposing campaign. This algorithm has guarantees only in the correlated setting with even budget k but in practice we can use the algorithm also for the heterogeneous setting. We refer to this algorithm as Common and the pseudo-code is given in Algorithm 1.

We first show in the following lemma that adding common seeds may halve the score, in the worst case. Then, we use this lemma to prove the approximation guarantee

Algorithm 1: Common, greedy algorithm that only adds common seeds

$S_1 \leftarrow S_2 \leftarrow \emptyset$;
while $|S_1| + |S_2| \leq k$ **do**
  $c \leftarrow \arg\max_c \Phi(S_1 \cup \{c\}, S_2 \cup \{c\})$;
  $s_1 \leftarrow \arg\max_{s \in I_1} \Phi(S_1, S_2 \cup \{s\})$;
  $s_2 \leftarrow \arg\max_{s \in I_2} \Phi(S_1 \cup \{s\}, S_2)$;
  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.

Lemma 4. *Let $\langle S_1, S_2\rangle$ be a solution to* BALANCE-C*, with an even budget k. There exists a solution $\langle S_1', S_2'\rangle$ with $S_1' = S_2'$ such that $\Phi_C(S_1', S_2') \geq \Phi_C(S_1, S_2)/2$.*

Proof. As we are dealing with the correlated setting, we can write $r(S) = r_1(S) = r_2(S)$. Our first step is to decompose $\omega = \Phi_C(S_1, S_2)$ into several components. To do so, we partition the vertices based on their reachability from the initial seeds,

$$A = r(I_1) \cap r(I_2), \qquad B = r(I_1) \setminus r(I_2),$$
$$C = r(I_2) \setminus r(I_1), \qquad D = V \setminus (r(I_1) \cup r(I_2)).$$

Note that these are all random variables. If $S_1 = S_2 = \emptyset$, then $\Phi_C(S_1, S_2) = \mathrm{E}_C[|A| + |D|]$. More generally, S_1 may balance some vertices in C, and S_2 may balance some vertices in B. We may also introduce new imbalanced vertices in D. To take this into account we define

$$B' = B \cap r(S_2), \qquad C' = C \cap r(S_1),$$
$$D' = D \setminus (r(S_1) \triangle r(S_2)).$$

We can express the cost of $\Phi_C(S_1, S_2)$ as

$$\omega = \Phi_C(S_1, S_2) = \mathrm{E}_C[|A| + |B'| + |C'| + |D'|]\,.$$

Split $S_1 \cup S_2$ in two equal-size sets, T and Q, and define

$$\omega_1 = \Phi_C(T, T), \quad \omega_2 = \Phi_C(Q, Q).$$

We claim that $\omega \leq \omega_1 + \omega_2$. This proves the proposition, since $\omega_1 + \omega_2 \leq 2\max\{\omega_1, \omega_2\}$.

To prove the claim let us first split T and Q,

$$T_1 = T \cap S_1,\ T_2 = T \cap S_2,\ Q_1 = Q \cap S_1,\ Q_2 = Q \cap S_2.$$

Our next step is to decompose ω_1 and ω_2, similar to ω. To do that, we define

$$B_1 = B \cap r(T_2), \qquad B_2 = B \cap r(Q_2),$$
$$C_1 = C \cap r(T_1), \qquad C_2 = C \cap r(Q_1).$$

Note that, the pair $\langle T, T\rangle$ does not introduce new imbalanced nodes. This leads to

$$\omega_1 = \Phi_C(T, T) = \mathrm{E}_C[|A| + |B_1| + |C_1| + |D|]\,,$$

and similarly,

$$\omega_2 = \Phi_C(Q, Q) = \mathrm{E}_C[|A| + |B_2| + |C_2| + |D|]\,.$$

To prove $\omega \leq \omega_1 + \omega_2$, note that $|D'| \leq |D|$. In addition,

$$\begin{aligned} |B'| &= |B \cap (r(T_2) \cup r(Q_2))| \\ &\leq |B \cap r(T_2)| + |B \cap r(Q_2)| = |B_1| + |B_2| \end{aligned}$$

and

$$\begin{aligned} |C'| &= |C \cap (r(T_1) \cup r(Q_1))| \\ &\leq |C \cap r(T_1)| + |C \cap r(Q_1)| = |C_1| + |C_2|. \end{aligned}$$

Combining these inequalities proves the proposition. □

It is easy to see that the greedy algorithm satisfies the conditions of the following proposition.

Proposition 5. *Assume an iterative algorithm where at each iteration, we add one or two vertices to our solution until our constraints are met. Let S_1^i, S_2^i be the sets after the i-th iteration, $S_1^0 = S_2^0 = \emptyset$. Let $\eta_i = \Phi_C(S_1^i, S_2^i)$ be the cost after the i-th iteration. Assume that $\eta_i \geq \eta_{i-1}$. Assume further that for $i = 1, \ldots, k/2$ it holds that $\eta_i \geq \Phi_C(S_1^{i-1} \cup \{c\}, S_2^{i-1} \cup \{c\})$. Then the algorithm yields $(1 - 1/e)/2$ approximation.*

To prove the proposition, we need the following technical lemma, which is a twist of a standard technique for proving the approximation ratio of the greedy algorithm on submodular functions.

Lemma 6. *Assume a universe U. Let $f : 2^U \to \mathbb{R}$ be a positive function. Let $T \subseteq U$ be a set with k elements. Let $C_0 \subseteq \cdots \subseteq C_k$ be a sequence of subsets of U. Assume that $f(C_i) \geq \max_{t \in T} f(C_{i-1} \cup \{t\})$.*

Assume further that for each $i = 1, \ldots, k$, we can decompose f as $f = g_i + h_i$ such that

1. *g_i is submodular and monotonically increasing function,*
2. *$h_i(W) = h_i(C_{i-1})$, for any $W \subseteq T \cup C_{i-1}$.*

Then $f(C_k) \geq (1 - 1/e) f(T)$.

Proof. The assumptions of the propositions imply

$$\begin{aligned}
f(T) &= g_i(T) + h_i(T) \\
&= g_i(T) + h_i(C_{i-1}) \\
&\leq g_i(C_{i-1}) + h_i(C_{i-1}) + \sum_{t \in T} g_i(C_{i-1} \cup \{t\}) - g_i(C_{i-1}) \\
&= f(C_{i-1}) + \sum_{t \in T} h_i(C_{i-1}) + g_i(C_{i-1} \cup \{t\}) - g_i(C_{i-1}) - h_i(C_{i-1}) \\
&= f(C_{i-1}) + \sum_{t \in T} h_i(C_{i-1} \cup \{t\}) + g_i(C_{i-1} \cup \{t\}) - g_i(C_{i-1}) - h_i(C_{i-1}) \\
&= f(C_{i-1}) + \sum_{t \in T} f(C_{i-1} \cup \{t\}) - f(C_{i-1}) \\
&\leq f(C_{i-1}) + k(f(C_i) - f(C_{i-1})),
\end{aligned}$$

where the first inequality is due to the submodularity of g_i, and is a standard trick to prove the approximation ratio for the greedy algorithm.

We can rewrite the above inequality as

$$k f(T) + (1 - k) f(T) = f(T) \leq f(C_{i-1}) + k(f(C_i) - f(C_{i-1})).$$

Rearranging the terms leads to

$$\frac{k-1}{k}(f(C_{i-1}) - f(T)) \leq f(C_i) - f(T) \quad .$$

Applying induction over i, yields

$$f(C_k) - f(T) \geq \left(\frac{k-1}{k}\right)^k (f(C_0) - f(T)) \geq \frac{1}{e}(f(C_0) - f(T)) \geq -f(T)/e,$$

leading to $f(C_k) \geq (1 - 1/e) f(T)$. □

We can now prove the main claim. Note that since we are using the correlated model, we have $r_1 = r_2$. For notational simplicity, we will write $r = r_1 = r_2$.

Proof of Proposition 5. Let OPT be the cost of the optimal solution. Let D be the solution maximizing $\Phi_C(D, D)$ with $|D| \leq k/2$. Lemma 4 guarantees that $OPT/2 \leq \Phi_C(D, D)$.

In order to apply Lemma 6, we first define the universe U as

$$U = \{\langle u, v \rangle \mid u, v \in V\} \cup \{\langle v, \emptyset \rangle \mid v \in V\} \cup \{\langle \emptyset, v \rangle \mid v \in V\}.$$

The sets are defined as

$$C_i = \left\{\langle v, \emptyset\rangle \mid v \in S_1^i\right\} \cup \left\{\langle \emptyset, v\rangle \mid v \in S_2^i\right\}.$$

Given a set $C \subseteq U$, let us define $\pi_1(C) = \{v \mid \langle v, u\rangle \in C, v \neq \emptyset\}$ to be the union of the first entries in C. Similarly, define $\pi_2(C) = \{v \mid \langle u, v\rangle \in C, v \neq \emptyset\}$.

We can now define f as $f(C) = \Phi_C(\pi_1(C), \pi_2(C))$. To decompose f, let us first write

$$X_i = r(I_1 \cup \pi_1(C_{i-1})) \cup r(I_2 \cup \pi_2(C_{i-1})) = r(I_1 \cup S_1^{i-1}) \cup r(I_2 \cup S_2^{i-1}), \quad Y_i = V \setminus X_i.$$

and set

$$g_i(C) = \mathrm{E}[|X_i \setminus (r(I_1 \cup \pi_1(C)) \triangle r(I_2 \cup \pi_2(C)))|],$$
$$h_i(C) = \mathrm{E}[|Y_i \setminus (r(I_1 \cup \pi_1(C)) \triangle r(I_2 \cup \pi_2(C)))|].$$

Finally, we set $T = \{\langle d, d\rangle \mid d \in D\}$.

First note that $f = g_i + h_i$ since $X_i \cap Y_i = \emptyset$. The proof of Lemma 2 shows that g_i is monotonically increasing and submodular.

Let $C \subseteq C_{i-1} \cup T$. If there is a vertex v in $r(I_1 \cup \pi_1(C))$ but not in X_i, then this means v was influenced by $d \in D$. Since $d \in \pi_2(C)$, we have $v \in r(I_2 \cup \pi_2(C))$. That is,

$$r(I_1 \cup \pi_1(C)) \setminus X_i = r(I_2 \cup \pi_2(C)) \setminus X_i.$$

Since Y_i and X_i are disjoint, this gives us

$$\begin{aligned} h_i(C) &= \mathrm{E}[|Y_i \setminus (r(I_1 \cup \pi_1(C)) \triangle r(I_2 \cup \pi_2(C)))|] \\ &= \mathrm{E}[|Y_i \setminus ((r(I_1 \cup \pi_1(C)) \setminus X_i) \triangle (r(I_2 \cup \pi_2(C)) \setminus X_i))|] \\ &= \mathrm{E}[|Y_i|]. \end{aligned}$$

That is, $h_i(C)$ is constant for any $C \subseteq C_{i-1} \cup T$. Thus, $h_i(C) = h_i(C_{i-1})$.

Finally, the assumpion of the proposition guarantees that $f(C_i) \geq f(C_{i-1} \cup \{t\})$, for $t \in T$.

Thus, these definitions meet all the prerequisites of Lemma 6, guaranteeing that

$$(1 - 1/e)\Phi_C(D, D) \leq \Phi_C(S_1^{k/2}, S_2^{k/2}) \leq \Phi_C(S_1^k, S_2^k).$$

Since $OPT/2 \leq \Phi_C(D, D)$, the result follows. □

Algorithm 2: common seeds as baseline. Not allowing new imbalanced vertices may prove to be too restrictive. We can relax this condition by allowing new imbalanced vertices as long as the gain is at least as good as adding a common seed. We refer to this algorithm as Hedge and the pseudo-code is given in Algorithm 2. The approximation guarantee for this algorithm—in the correlated setting and with even budget—follows immediately from Proposition 5 as it also satisfies the conditions.

Algorithm 2: Hedge, greedy algorithm, where each step is as good as adding the best common seed

```
S_1 ← S_2 ← ∅;
while |S_1| + |S_2| ≤ k do
    c ← arg max_c Φ(S_1 ∪ {c}, S_2 ∪ {c});
    s_1 ← arg max_s Φ(S_1, S_2 ∪ {s});
    s_2 ← arg max_s Φ(S_1 ∪ {s}, S_2);
    add the best option among ⟨c, c⟩, ⟨∅, s_1⟩, ⟨s_2, ∅⟩, ⟨s_2, s_1⟩, to ⟨S_1, S_2⟩ while respecting the
      budget.
```

5 Experimental evaluation

In this section, we evaluate the effectiveness of our algorithms on real-world datasets. We focus on (i) analyzing the quality of the seeds picked by our algorithms in comparison to other heuristic

Table 1: Dataset statistics. The column $|C|$ refers to the average number of edges in a randomly generated cascade in the correlated case, while $|C_1|$ and $|C_2|$ refer to average number of edges generated in a cascade of the campaigns 1 and 2, respectively, in the heterogeneous case.

Dataset	# Nodes	# Edges	$\lvert C\rvert$	$\lvert C_1\rvert$	$\lvert C_2\rvert$
`Abortion`	279 505	671 144	2 105	326	1 801
`Brexit`	22 745	48 830	476	113	390
`Fracking`	374 403	1 377 085	4 156	1 595	3 103
`iPhone`	36 742	49 248	4 776	339	4 478
`ObamaCare`	334 617	1 511 670	6 614	2 404	4 527
`US-elections`	80 544	921 368	4 697	3 097	12 044

approaches and baselines; (ii) analyzing the efficiency and the scalability of our algorithms; and (iii) providing anecdotal examples of the obtained results.

For all experiments we report averages over 1 000 random simulations of the cascade process. As argued by Kempe et al. [24], a random cascade can be generated in advance by sampling each edge e with probability $p_1(e)$ or $p_2(e)$, depending on the model and the campaign. In all experiments we set k to range between 5 and 50 with a step of 5.

Datasets: To evaluate the effectiveness of our algorithms, we run experiments on real-world data collected from twitter. Let $G = (V, E)$ be the twitter follower graph. A directed edge $(u, v) \in E$ indicates that user v follows u; note that the edge direction indicates the "information flow" from a user to their followers. We define a cascade $G_X = (X, E_X)$ as a graph over the set of users $X \subseteq V$ who have retweeted at least one hashtag related to a topic (e.g., US elections). An edge $(u, v) \in E_X \subseteq E$ indicates that v retweeted u.

We use datasets from six topics with opposing viewpoints, covering politics (`US-elections`, `Brexit`, `ObamaCare`), policy (`Abortion`, `Fracking`), and lifestyle (`iPhone`, focusing on iPhone vs. Samsung). All datasets are collected by filtering the twitter streaming API (1% random sample of all tweets) for a set of keywords used in previous work [28]. For each dataset, we identify two sides (indicating the two view-points) on the retweet graph, which has been shown to capture best the two opposing sides of a controversy [16]. Details on the statistics of the dataset are shown in Table 1.

After building the graphs, we need to estimate the diffusion probabilities for the heterogeneous and correlated models. Note that the estimation of the diffusion probabilities is orthogonal to our contribution in this paper. For the sake of concreteness we have used the approach described below. One could use a different, more advanced, method; our methods are still applicable.

Let $q_1(v)$ and $q_2(v)$ be an *a priori* probability of a user v retweeting sides 1 and 2, respectively. These are measured from the data by looking at how often a user retweets content from users and keywords that are discriminative of each side. For example, for `US-elections`, the discriminative users and keywords for side Hillary would be @hillaryclinton and #imwither, and for Trump, @realdonaldtrump and #makeamericagreatagain. Additional details on data collection, and user/keyword sets for side identification are given in Table 2 in Appendix A.

The probability that user v retweets user u (cascade probability) is then defined as

$$p_i(u, v) = \alpha\, q_i(v) + (1 - \alpha) \left(\frac{R(u, v) + 1}{R(v) + 2} \right), \quad i = 1, 2,$$

where $R(u, v)$ is the number of times v has retweeted u, and $R(v)$ is the total number of retweets of user v. The cascade probabilities p_i capture the fact that users retweet content if they see it from their friends (term $\frac{R(u,v)+1}{R(v)+2}$) or based on their own biases (term $q_i(v)$). The additive terms in the numerator and denominator provide an additive smoothing by Laplace's rule of succession.

We set the value of α to 0.8 for the heterogeneous setting. For the correlated setting, α is set to zero.

Baselines. We use 5 different baselines. The first baseline, BBLO, is an adaptation of the framework by Borodin et al. [8]. This framework requires an objective function as an input, and here we use our objective function Φ. The framework works as follows: The two campaigns are given

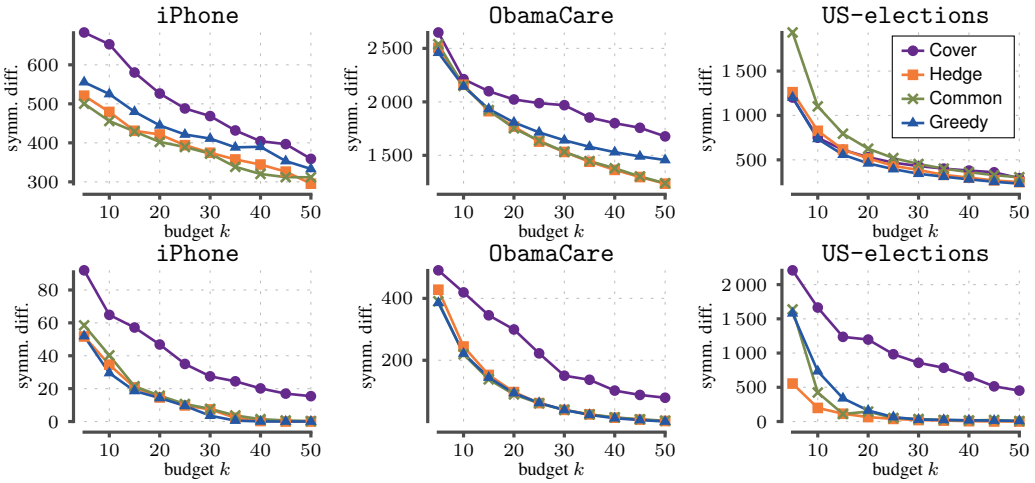

Figure 1: Expected symmetric difference $n - \Phi_C$ as a function of the budget k. Top row, heterogeneous model, bottom row: Correlated model. Low values are better.

a budget $k/2$ on the number of seeds that they can select. At each round, we select a vertex v optimizing $\Phi(S_1 \cup \{v\}, S_2)$, the vertex v is added to S_1, and then we select a vertex w optimizing $\Phi(S_1, S_2 \cup \{w\})$, which is added to S_2. We should stress that the theoretical guarantees by [8] do not apply because our objective is not submodular. Thus, BBLO is a heuristic.

The next two heuristics add a set of common seeds to both campaigns. We run a greedy algorithm for campaign $i = 1, 2$ to select the set S'_i with the $\ell \gg k$ vertices P_i that optimizes the function $r_i(S'_i \cup I_i)$. We consider two heuristics: Union selects S_1 and S_2 to be equal to the $k/2$ first distinct vertices in $S'_1 \cup S'_2$ while Intersection selects S_1 and S_2 to be equal to $k/2$ first vertices in $S'_1 \cap S'_2$. Here the vertices are ordered based on their discovery time. In both cases we (arbitrarily) break ties in favor of campaign 1.

Finally, HighDegree selects the vertices with the largest number of followers and assigns them alternately to the two cascades; and Random assigns $k/2$ random seeds to each campaign.

In addition to the baselines, we also consider a simple greedy algorithm Greedy. The difference between Cover and Greedy is that, in each iteration, Cover adds the seed that maximizes Ω, while Greedy adds the seed that maximizes Φ. We can only show an approximation guarantee for Cover but Greedy is a more intuitive approach as it considers directly the number of balanced vertices and we use it as a heuristic.

Comparison of the algorithms. We start by evaluating the quality of the sets of seeds computed by our algorithms, i.e., the number of equally-informed vertices.

Heterogeneous setting. We consider first the case of heterogeneous networks. The results for the selected datasets are shown in Figure 1. Full results are shown in Appendix A. Instead of plotting Φ, we plot the number of the remaining unbalanced vertices, $n - \Phi$, as it makes the results easier to distinguish; i.e., an optimal solution achieves the value 0.

The first observation is that the approximation algorithm Cover performs, in general, worse than the other two heuristics. This is due to the fact that Cover does not optimize directly the objective function. Hedge performs better than Greedy, in general, since it examines additional choices to select. The only deviation from this picture is for the `US-elections` dataset, where the Greedy outperforms Hedge by a small factor. This may due to the fact that while Hedge has more options, it allocates seeds in batches of two.

The algorithms seem to follow a diminishing-returns behavior, on most cases. This behavior appears despite the fact that the optimization function is not submodular.

Correlated setting. Next we consider correlated networks. We experiment with the three approximation algorithms Cover, Common, Hedge, and the heuristic Greedy. The results are shown in

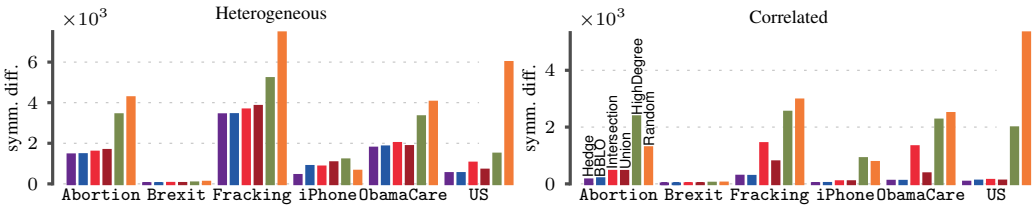

Figure 2: Expected symmetric difference $n - \Phi$ of Hedge and the baselines. $k = 20$. Low values are better.

Figure 1. Cover performs again the worst since it is the only method that introduces new unbalanced vertices without caring about their cardinality. Its variant, Greedy, performs much better in practice even though it does not provide an approximation guarantee. The algorithms Common, Greedy, and Hedge perform very similar to each other without a clear winner.

Comparison with baselines. Our next step is to compare against the baselines. For simplicity, we focus on $k = 20$; the overall conclucions hold for other budgets. The results for Hedge versus the five baselines are shown in Figure 2.

From the results we see that BBLO is the best competitor: its scores are the closest to Hedge, and it receives slightly better scores in 3 out of 12 cases. The competitiveness is not surprising because we specifically set the objective function in BBLO to be $\Phi(S_1, S_2)$. The Intersection and Union also perform well but are always worse than Hedge. Random is unpredictable but always worse than Hedge. In the case of heterogeneous networks, Hedge selects seeds that leave less unbalanced vertices, by a factor of two on average, compared to the seeds selected by the HighDegree method. For correlated networks, our method outperforms the two baselines by an order of magnitude.

Running time. We proceed to evaluate the efficiency and the scalability of our algorithms. The running times, in seconds, of our algorithms, for all datasets and for $k = 20$, are shown in Figure 5 in Appendix A as a function of network size. We observe that all algorithms have comparable running times and good scalability.

Use case with `Fracking`. We present a qualitative case-study analysis for the seeds selected by our algorithm. We highlight the `Fracking` dataset, even though we applied similar analysis to the other datasets as well (the results are given in Figure 6 in Appendix A). Recall that for each dataset we identify two sides with opposing views, and a set of initial seeds for each side (I_1 and I_2). We consider the users in the initial seeds I_1 (side supporting fracking), and summarize the text of all their Twitter profile descriptions in a word cloud. The result, as can be seen in Figure 6 in Appendix A, contains words that are used to emphasize the benefits of fracking (energy, oil, gas, etc.). We then draw a similar word cloud for the users identified by the Hedge algorithm as seed nodes in the sets S_1 and S_2 ($k = 50$). The result, shown in Figure 6 in Appendix A, contains a more balanced set of words, which includes many words used to underline the environmental dangers of fracking.

6 Conclusion

We presented the first study of the problem of balancing information exposure in social networks using techniques from the area of information diffusion. Our approach has several novel aspects. In particular, we formulate our problem by seeking to optimize a *symmetric difference* function, which is neither monotone nor submodular, and thus, not amenable to existing approaches. Additionally, while previous studies consider a setting with selfish agents and provide bounds on best-response strategies (i.e., move of the last player), we consider a centralized setting and provide bounds for a global objective function.

Our work provides several directions for future work. One interesting problem is to improve the approximation guarantee for the problem we define. Second, we would like to extend the problem definition for more than two campaigns and design approximation algorithms for that case.

Acknowledgments

Work partially done while Nikos Parotsidis was visiting Aalto University. This work has been supported by the Academy of Finland project "Nestor" (286211) and the EC H2020 RIA project "SoBigData" (654024).

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *LinkKDD*, pages 36–43, 2005.

[2] L. Akoglu. Quantifying political polarity based on bipartite opinion networks. In *ICWSM*, 2014.

[3] N. Alon, M. Feldman, A. D. Procaccia, and M. Tennenholtz. A note on competitive diffusion through social networks. *IPL*, 110(6):221–225, 2010.

[4] K. R. Apt and E. Markakis. Diffusion in social networks with competing products. In *SAGT*, pages 212–223, 2011.

[5] A. Beutel, B. A. Prakash, R. Rosenfeld, and C. Faloutsos. Interacting viruses in networks: can both survive? In *KDD*, pages 426–434, 2012.

[6] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *WINE*, 2007.

[7] A. Borodin, Y. Filmus, and J. Oren. Threshold models for competitive influence in social networks. In *WINE*, 2010.

[8] A. Borodin, M. Braverman, B. Lucier, and J. Oren. Strategyproof mechanisms for competitive influence in networks. *Algorithmica*, 78(2):425–452, 2017.

[9] M. Broecheler, P. Shakarian, and V. Subrahmanian. A scalable framework for modeling competitive diffusion in social networks. In *ICSC*, pages 295–302, 2010.

[10] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW*, pages 665–674, 2011.

[11] T. Carnes, C. Nagarajan, S. M. Wild, and A. Van Zuylen. Maximizing influence in a competitive social network: a follower's perspective. In *EC*, 2007.

[12] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political Polarization on Twitter. In *ICWSM*, 2011.

[13] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. Echo chambers in the age of misinformation. *arXiv:1509.00189*, 2015.

[14] N. Du, L. Song, M. G. Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In *Advances in neural information processing systems*, pages 3147–3155, 2013.

[15] P. Dubey, R. Garg, and B. De Meyer. Competing for customers in a social network: The quasi-linear case. In *WINE*, 2006.

[16] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *WSDM*, pages 33–42, 2016.

[17] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Reducing controversy by connecting oppposing views. In *WSDM*, 2017.

[18] R. K. Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users1. *JCMC*, 14(2):265–285, 2009.

[19] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[20] J. Gottfried and E. Shearer. News use across social media platforms 2016. *Pew Research Center*, 2016.

[21] S. Goyal, H. Heidari, and M. Kearns. Competitive contagion in networks. *Games and Economic Behavior*, 2014.

[22] P. H. C. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *ICWSM*, 2013.

[23] R. Jie, J. Qiao, G. Xu, and Y. Meng. A study on the interaction between two rumors in homogeneous complex networks under symmetric conditions. *Physica A*, 454:129–142, 2016.

[24] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[25] J. Kostka, Y. A. Oswald, and R. Wattenhofer. Word of mouth: Rumor dissemination in social networks. In *SIROCCO*, pages 185–196, 2008.

[26] Q. V. Liao and W.-T. Fu. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *CSCW*, pages 184–196, 2014.

[27] Q. V. Liao and W.-T. Fu. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *CHI*, pages 2745–2754, 2014.

[28] H. Lu, J. Caverlee, and W. Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *CIKM*, pages 213–222, 2015.

[29] W. Lu, W. Chen, and L. V. Lakshmanan. From competition to complementarity: comparative influence diffusion and maximization. *PVLDB*, 9(2):60–71, 2015.

[30] A. Morales, J. Borondo, J. Losada, and R. Benito. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos*, 25(3), 2015.

[31] S. A. Munson, S. Y. Lee, and P. Resnick. Encouraging reading of diverse political viewpoints with a browser widget. In *ICWSM*, 2013.

[32] S. A. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *ICDM*, pages 539–548, 2012.

[33] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions – I. *Mathematical Programming*, 14(1):265–294, 1978.

[34] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz. Containment of misinformation spread in online social networks. In *Web Science*, pages 213–222, 2012.

[35] D. Nikolov, D. F. Oliveira, A. Flammini, and F. Menczer. Measuring online social bubbles. *PeerJ Computer Science*, 1:e38, 2015.

[36] E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[37] V. Tzoumas, C. Amanatidis, and E. Markakis. A game-theoretic analysis of a competitive diffusion process over social networks. In *WINE*, 2012.

[38] I. Valera and M. Gomez-Rodriguez. Modeling adoption of competing products and conventions in social media. In *ICDM*, 2015.

[39] V. Vydiswaran, C. Zhai, D. Roth, and P. Pirolli. Overcoming bias to learn about controversial topics. *JAIST*, 2015.

A Additional tables and figures related to the experimental evaluation

Table 2: Dataset descriptions, as well as tags and rewteets that were used to collect the data.

USelections: Tweets containing hashtags and keywords identifying the USElections, such as #uselections, #trump2016, #hillary2016, etc. Collected using Twitter 1% sample for 2 weeks in September 2016	
Pro-Hillary RT @hillaryclinton, #hillary2016, #clintonkaine2016, #imwithher	*Pro-Trump* RT @realdonaldtrump, #makeamericagreatagain, #trumppence16, #trump2016
Brexit: Tweets containing hashtags #brexit, #voteremain, #voteleave, #eureferendum for all of June 2016, from the 1% Twitter sample.	
Pro-Remain #voteremain, #strongerin, #remain, #remaineu, #votein	*Pro-Leave* #voteleave, #strongerout, #leaveeu, #takecontrol, #leave, #voteout
Abortion: Tweets containing hashtags #abortion, #prolife, #prochoice, #anti-abortion, #pro-abortion, #plannedparenthood from Oct 2011 to Aug 2016.	
Pro-Choice RT @thinkprogress, RT @komenforthecure, RT @mentalabortions, #waronwomen, #nbprochoice, #prochoice, #standwithpp, #reprorights	*Pro-Life* RT @stevenertelt, RT @lifenewshq, #praytoendabortion, #prolifeyouth, #prolife, #defundplannedparenthood, #defundpp, #unbornlivesmatter
Obamacare: Tweets containing hashtags #obamacare, and #aca from Oct 2011 to Aug 2016.	
Pro-Obamacare RT @barackobama, RT @lolgop, RT @charlespgarcia, RT @defendobamacare, RT @thinkprogress, #obamacares, #enoughalready, #uniteblue	*Anti-Obamacare* RT @sentedcruz, RT @realdonaldtrump, RT @mittromney, RT @breitbartnews, RT @tedcruz, #defundobamacare, #makedclisten, #fullrepeal, #dontfundit
Fracking: Tweets containing hashtags and keywords #fracking, 'hydraulic fracturing', 'shale', 'horizontal drilling', from Oct 2011 to Aug 2016.	
Pro-Fracking RT @shalemarkets, RT @energyindepth, RT @shalefacts, #fracknation, #frackingez, #oilandgas, #greatgasgala, #shalegas	*Anti-Fracking* RT @greenpeaceuk, RT @greenpeace, RT @ecowatch, #environment, #banfracking, #keepitintheground, #dontfrack, #globalfrackdown, #stopthefrackattack
iPhone vs. Samsung: Tweets containing hashtags #iphone, and #samsung from April (release of Samsung Galaxy S7), and September 2015 (release of iPhone 7).	
Pro-iPhone #iphone	*Pro-Samsung* #samsung

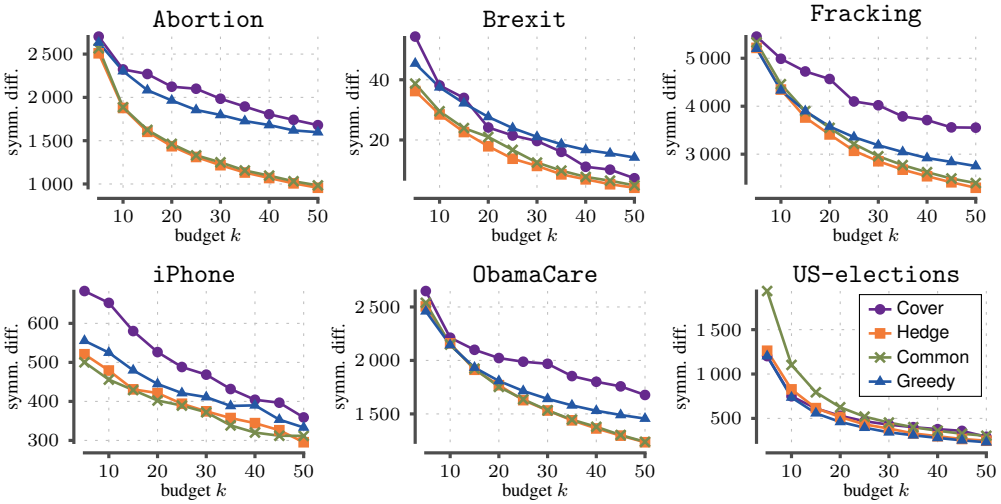

Figure 3: Expected symmetric difference $n - \Phi_H$ as a function of the budget k. Heterogeneous model. Low values are better.

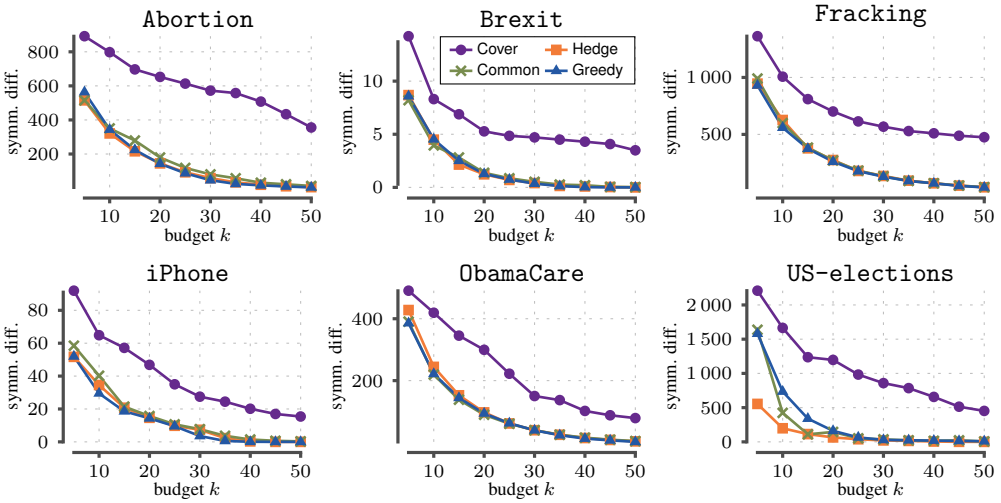

Figure 4: Expected symmetric difference $n - \Phi_C$ as a function of the budget k. Correlated model. Low values are better.

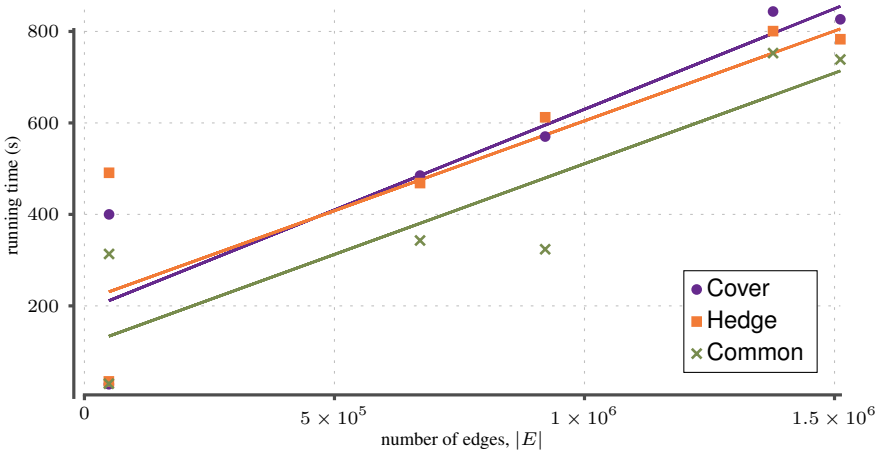

Figure 5: Running time as a function of number of edges. Correlated model with $k = 20$.

Side 1 | Side 2 | Hedge

Pro-Choice | *Pro-Life*

Pro-Remain | *Pro-Leave*

Pro-Fracking | *Anti-Fracking*

Pro-iPhone | *Pro-Samsung*

Pro-Obamacare | *Anti-Obamacare*

Pro-Hillary | *Pro-Trump*

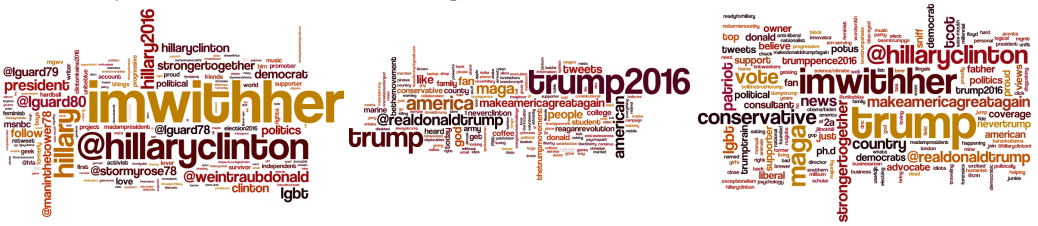

Figure 6: Word clouds of the profiles for the initial seeds, and profiles selected by Hedge.