Maximizing the Diversity of Exposure in a Social Network

Antonis Matakos, Cigdem Aslay, Esther Galbrun, and Aristides Gionis

Abstract—Social-media platforms have created new ways for citizens to stay informed and participate in public debates. However, to enable a healthy environment for information sharing, social deliberation, and opinion formation, citizens need to be exposed to sufficiently diverse viewpoints that challenge their assumptions, instead of being trapped inside filter bubbles. In this paper, we take a step in this direction and propose a novel approach to maximize the diversity of exposure in a social network. We formulate the problem in the context of information propagation, as a task of recommending a small number of news articles to selected users. In the proposed setting, we take into account content and user leanings, and the probability of further sharing an article. Our model allows to capture the balance between maximizing the spread of information and ensuring the exposure of users to diverse viewpoints.
The resulting problem can be cast as maximizing a monotone and submodular function, subject to a matroid constraint on the allocation of articles to users. It is a challenging generalization of the influence-maximization problem. Yet, we are able to devise scalable approximation algorithms by introducing a novel extension to the notion of random reverse-reachable sets. We experimentally demonstrate the efficiency and scalability of our algorithm on several real-world datasets.

1 Introduction

Over the past decade, the emergence of social-media platforms has changed society in unprecedented ways, completely altering the landscape of societal debates and creating radically new ways of collective action. In this networked public sphere, members of society have access to a public podium where they can participate in public debate and speak up about topics they deem to be of public concern. This emerging environment of participatory culture has made the diversity of citizens' views more relevant than ever before.

While having the potential to expose individuals to diverse opinions, social-media platforms typically resort to personalization algorithms that filter content based on social connections and previously expressed opinions, creating filter bubbles [1]. The resulting echo chambers tend to amplify and reinforce pre-existing opinions, catalyzing an environment that has a corrosive effect on the democratic debate.

In this paper we propose *a novel approach towards breaking filter bubbles*. We consider social-media discussions around a topic that are characterized by a number of viewpoints falling within a predefined spectrum of opinions. To accurately model the dynamics of social-media platforms, we assume that each viewpoint is represented by a number of items (articles, posts) propagating through the network via messages, re-shares, retweets, etc. Furthermore, we assume that each individual is associated with a *leaning* with respect to the issue, which impacts whether they will further disseminate any article they come across, depending on how it aligns with their leaning. We think that this is a realistic assumption, since, for example, an individual with a conservative leaning might be reluctant to share an article with a liberal leaning.

We refer to the diversity of the information that a user is exposed to as the user's "*diversity exposure level*". It depends on the viewpoint expressed in the articles the user consumes, referred to as *article leanings*, and the users' existing viewpoint on the matter, referred to as *user leanings*. We assume that the diversity exposure level of users can be increased through *content recommendations* made by the social-media platform. Considering that filter bubbles result from a lack of exposure to diverse viewpoints, our aim is to measure and maximize the total diversity exposure levels of all users in the network.

Our problem can be naturally defined in an *information-propagation setting* [2]: we ask to select a small number of seed users and the articles that should be recommended to them so as to maximize the total diversity of exposure in the network. Since the recommended articles are inserted into the timeline of the users, disrupting the organic flow of the content in the network, we also consider a limit on the number of articles that can be recommended to a user in this way.

An attractive aspect of our problem setting is that it consolidates many aspects of the functionality of real-life social networks. By incorporating article leanings, user leanings, and the probabilities of further sharing an article, we ask to find the recommendations that translate to a good spread and simultaneously maximize the diversity exposure level of the users. To better understand the interplay between spread and diversity, observe that assigning articles that match the users' predisposition is likely to result in a high spread but minimal increase of diversity, while recommending articles that are opposed to users' predispositions, will likely result in high diversity locally but hinder the spread of the articles. This trade-off is central to the diversity-maximization problem we consider.

We show that taking all the aforementioned components into account, the problem of maximizing the diversity of exposure

- *A. Matakos and C. Aslay are with the Department of Computer Science, Aalto University, Finland.*
  E-mail: firstname.lastname@aalto.fi
- *E. Galbrun is with the School of Computing, University of Eastern Finland.*
  E-mail: esther.galbrun@uef.fi
- *A. Gionis is with the Department of Computer Science, KTH Royal Institute of Technology, Sweden, and the Department of Computer Science, Aalto University, Finland.*
  E-mail: argioni@kth.se

This work was supported by Academy of Finland projects 286211, 313927, and 317085, and the EC H2020 RIA project "SoBigData" (654024)

in a social network can be cast as maximizing a monotone and submodular function subject to a matroid constraint on the allocation of articles to users. We show that this problem is NP-hard and is far more challenging than the classical influence-maximization problem. We introduce a non-trivial generalization of random reverse-reachable sets (RR-sets) [3], which we call random *reverse co-exposure* sets (RC-sets), for accurately estimating the diversity of exposure in a social network. We propose a scalable approximation algorithm, named *Two-phase Diversity Exposure Maximization* (TDEM), that leverages random RC-sets and an adaptive sample size determination procedure, ensuring quality guarantees on the returned solution with high probability.

Although our approach belongs to a large body of work on information propagation and breaking filter bubbles, there are significant differences and novelties. In particular:

- We are the first to address the problem of maximizing the diversity of exposure and breaking filter bubbles in an item-aware information propagation setting. We leverage several real-world aspects of social-media functionality, such as how users consume and share articles, while considering user-article dependent propagation probabilities.
- We formally define the problem of maximizing the diversity of exposure, prove its hardness, and develop a simple greedy algorithm.
- We then introduce the notion of random *reverse co-exposure sets* and devise a scalable instantiation of the greedy algorithm with provable guarantees.
- Our extensive experimentation on real-world datasets confirms that our algorithm is scalable and delivers high quality solutions, significantly outperforming several natural baselines.

A preliminary version of our work provided a first theoretical and experimental treatment of the problem under a simpler formulation [4]. Specifically, we previously defined the diversity exposure level of a user to be equal to the breadth of leanings spanned by the items the user is exposed to, in addition to the user's own leaning. In this paper, we extend our preliminary results in several directions. First, we propose a refined scheme to quantify the diversity exposure level of a user. The new diversity definition measures not only the range of leanings in a set of items but also their spread within this range. That is, our refined scores does not only look at the extremes of represented leanings but also at how well intermediate leanings are covered. Second, we show that the total diversity exposure function remains submodular and monotone and we extend our scalable approximation framework based on random reverse co-exposure sets to operate under this new score. Finally, we provide additional experiments on many real-world datasets.

2 Related work

Our work relates to the emerging line of research on breaking filter bubbles in social media. To the best of our knowledge, this is the first work that approaches this problem from the angle of maximizing the diversity of information exposure in an item-aware independent-cascade model.

Filter bubbles and echo chambers. Recently, there have been a number of studies on the effects of "echo chambers" [5], [6], where users are only exposed to information from like-minded individuals, and of "filter bubbles" [1], [5], where algorithms only present personalized content that agrees with the user's viewpoint. In particular, Garrett et al. [6] observed that news stories containing opinion-challenging informations spread less than other news.

In order to measure how strongly these phenomena manifest themselves on social media, a significant body of work has emerged that focuses on measures for characterizing polarization [7], [8], [9], [10], [11], [12].

In a similar vein to ours, previous works have studied the problem of diversifying exposure. This task presents various aspects, such as the questions of *who* to target, *what* viewpoints to promote, or *how* best to present possibly opposing viewpoints to users [13]. Recent approaches focus on targeting users so as to reduce the polarization of opinions and bridge opposing views [7], [12], [14], [15]. These works consider an *opinion-formation* model whereas our underlying model is an *influence-propagation* model. From this angle, the works by Garimella et al. [16] and Rawal and Khan [17] are closest to our work. They consider an influence propagation setting, where two conflicting campaigns propagate in the network and the goal is to maximize the number of users exposed to both campaigns. The granularity of our setting is finer, however, since we consider items with leanings lying across a spectrum rather than two opposing sides. Additionally, we consider the leanings of users, which affect the propagation probabilities. Since our goal is to identify assignments of items to users, we aim to identify both the users to target and the viewpoints to expose them to.

Influence maximization. Our problem is also related to the work on influence maximization. Kempe et al. [2] formalized the influence maximization problem and proposed two propagation models, the *independent-cascade model* and the *linear-threshold model*. These models were subsequently extended to handle the case of multiple competing campaigns in a network [18], [19], [20]. As other authors have suggested, we consider a *central authority* selecting the seed set [16], [21], [22], [23]. Our setting is related to social advertising [22], [23], which also considers item-aware propagation models, aiming to allocate ads so as to maximize the engagement of users. Key to our work is the idea of *reverse reachable sets* introduced by Borgs et al. [3], which provides scalable solutions for the influence maximization problem. Subsequent works [24], [25], [26], [27] introduced techniques to improve upon this idea even further. We extend these ideas to our setting, and obtain an algorithm that scales to very large datasets.

3 Problem definition

Notation. The input to the problem of maximizing the diversity of exposure consists of the following ingredients: (i) a directed social graph $G = (V, E)$, with $|V| = n$ nodes and $|E| = m$ edges, where nodes represent users and a directed edge (u, v) indicates that user v follows user u, thus, v can see and propagate posts by u; (ii) a set H of (news) items on a (possibly controversial) topic, with $|H| = h$; (iii) item-specific propagation probabilities p^i_{uv}, for all items $i \in H$ and edges $(u, v) \in E$, where p^i_{uv} represents the probability that item i will propagate from user u to user v; (iv) a *leaning function* $\ell : V \cup H \to [-1, 1]$ that quantifies the polarity of the viewpoints of items and users with respect to the considered issue or topic.

Cascade model. We assume that the propagation of an item $i \in H$ from user u to user v follows the *independent-cascade model* with

parameter p^i_{uv}, and is independent from the propagation of other items to v from its in-neighbors. Thus, once u becomes active on item i at time t, the probability p^i_{uv} that u succeeds in activating v with item i at time $t+1$ is independent of other items with which user u or other in-neighbors of v might succeeded to activate v at any time. We incorporate the different tendency of users to share items with leanings diverging from or similar to their own by allowing item-specific propagation probabilities for each edge. Hence, p^i_{uv} implicitly takes into account the leanings of users u and v and of item i. The leaning of a user reflects the user's viewpoint, which is considered to be stable. Therefore, we assume that the transmission probabilities remain fixed over time, and probabilities p^i_{uv} are constant values input to our cascade model. We consider the estimation of user leanings and transmission probabilities as orthogonal to our work.[1]

Quantifying diversity of exposure. We say that user v is *exposed* to item i if v is *activated* on item i by an in-neighbor that is itself exposed on item i, or if user v is a *seed node* for item i. Consider a user v that is exposed to a set $I \subseteq H$ of items. It follows that user v is exposed to a set of leanings $\{\ell(v)\} \cup \{\ell(i) : i \in I\}$. Intuitively, we want each user to be aware of a multitude of viewpoints, while also retaining a balanced perspective. To account for both factors, we define a *penalty* function that quantifies the lack of diversity of exposure.

Specifically, we want to penalize large gaps in the spread of leanings, which correspond to ranges of opinions not represented among items the user is exposed to. Therefore, the function is defined for each user by considering the set of distinct leanings he is exposed to, sorted by polarity, and taking the sum of squared distances between consecutive leanings, also accounting for the extreme values of leanings. We consider that each item contributes only once to the diversity of exposure of a user. Therefore seeing the same article multiple times should have no impact on the objective.

We let $L(v, I) = \langle \ell_1, \ldots, \ell_\eta \rangle$ denote the set $\{\ell(v)\} \cup \{\ell(i) : i \in I\} \cup \{-1, 1\}$ sorted by increasing values, i.e., such that $\ell_i \leq \ell_j$ for all $i < j$. This set contains user v's own leaning, the distinct leanings among the items in I that user v has been exposed to, as well as the two extreme leanings across the spectrum of opinions, $\ell_1 = -1$ and $\ell_\eta = 1$. Then, we define the penalty for node v, $g_v : 2^H \to [0, 4]$, as

$$g_v(I) = \sum_{j=1}^{\eta-1} (\ell_{j+1} - \ell_j)^2, \quad \ell \in L(v, I). \tag{1}$$

Given the penalty function $g_v(I)$ that quantifies the lack of diversity in the leanings of the items I that v is exposed to, we define the *level of diversity exposure* $f_v : 2^H \to [0, 1]$ of v as

$$f_v(I) = 1 - \frac{1}{4} g_v(I). \tag{2}$$

Notice that the range of the diversity exposure function f_v is $[0, 1]$, where a value of 1 corresponds to the maximum possible diversity of exposure.

To motivate the definition of our diversity function f_v we provide the following two lemmas, which illustrate some of its desirable properties.

1. Twitter offers a built-in feature, that users can choose to opt-in, to estimate their preferences with respect to various topics, which remains valid for a limited amount of time. See https://help.twitter.com/en/using-twitter/tailored-suggestions

Lemma 1. *For all $I \subseteq J \subseteq H$, we have $f_v(I) \leq f_v(J)$.*

Lemma 1, for which the proof is provided as part of Lemma 3, states that f_v is *monotone*, i.e., the diversity exposure level of v cannot decrease as the user is exposed to more items.

Next we formally show that, if we fix the number of items that user v will see, then the configuration in which f_v is maximized corresponds to the *desired* scenario where the user leaning of v and the leanings of the items v is exposed to are equally spaced across $[-1, 1]$.

Lemma 2. *Consider a set of items I, so that I has fixed cardinality κ. Then, the diversity function $f_v(I)$ is maximized if the leanings of the items in I are equidistantly positioned in the interval $[-1, 1]$.*

Proof. For the sake of simplicity, and without loss of generality, assume that neither the leaning of v nor the extreme leanings -1 and 1 are represented in I, so that $|L(v, I)| = \kappa + 3$. Let $r_j = \ell_{j+1} - \ell_j$, for $j = 1, \ldots, \kappa+2$. Notice that $\sum_{j=1}^{\kappa+2} r_j$ is a constant that depends only on how we model the range of the leanings, i.e., for $[-1, 1]$, we have $\sum_{j=1}^{\kappa+2} r_j = 2$. Remember that by definition $f_v(I)$ is maximized whenever $g_v(I)$ is minimized. Then, solving the equations resulting from $g'_v(I) = 0$ and $\sum_{j=1}^{\kappa+2} r_j = 2$, we see that $g_v(I)$ attends its minimum value when

$$r_1 = \ldots = r_{\kappa+2} = \frac{2}{\kappa + 2}\,.$$

□

Assignment to seed nodes. We consider selecting a set of users in V as the *seed nodes* and expose them to a subset of items from H. Let $\mathcal{E} = V \times H$ denote the set of all possible (user, item) pairs and let $A \subseteq \mathcal{E}$ denote an assignment such that the set $A_i = \{u \in V : (u, i) \in A\}$ contains the seed nodes selected for initial exposure to item i and the set $A_u = \{i \in H : (u, i) \in A\}$ contains the items assigned to seed node u. For each $v \in V$, we denote by $I_v(A)$ the set of items that v is exposed to when the propagation process initialized with assignment A converges. The *diversity of exposure score* $F(A)$ of an assignment A is then defined as the sum of diversity exposure levels of all the users resulting from the assignment A in G

$$F(A) = \sum_{v \in V} f_v\big(I_v(A)\big). \tag{3}$$

Note that the function $f_v(I_v) : 2^{\mathcal{E}} \to [0, 1]$ is a composition $f_v(I_v) = f_v \circ I_v$ of the functions $I_v : 2^{\mathcal{E}} \to 2^H$ and $f_v : 2^H \to [0, 1]$. We will later use this fact to show that $f_v(I_v)$ is a submodular function over $\mathcal{E}$.

Constraints on assignments. We assume that we are interested in assignments of size at most $k \in \mathbb{N}$. Moreover, taking into account the limited attention span of users, which can be user-specific [28], we also limit the number of items that a user can be seeded with.[2] We model this using an *attention bound constraint* $k_u \in \mathbb{N}$ for each user $u \in V$. We say that an assignment A is feasible if $|A| \leq k$ and $|A_u| \leq k_u$, for each seed node u.

Assumptions. We assume that there exist $e, e' \in V \cup H$ such that $\ell(e) \neq \ell(e')$. This weak assumption is simply a bare minimum

2. As in previous work [22], we do not assume any attention bound on the number of items that are not recommendations in our problem definition, i.e., items that appear in the news-feed of the users in the social network, as such items are part of the organic operation of the network.

requirement on the diversity of the leanings of the users and items, aligned with the motivation of the problem. We will use this assumption in the greedy approximation analysis to constrain the optimal values of expected diversity exposure score to $\mathbb{R}_+$.

We are now ready to formally define our problem.

Problem 1 (Diversity Exposure Maximization)**.** *Given a directed social graph $G = (V, E)$ with user leanings $\ell(v)$, for all $v \in V$, a set of items H with item leanings $\ell(i)$, for all $i \in H$, item-specific propagation probabilities p^i_{uv}, for all $(u, v) \in E$ and all $i \in H$, positive integers k_u as attention bound constraints for all $u \in V$, and a positive integer k, find a feasible assignment A that maximizes the expected diversity exposure score*

$$\begin{aligned} \underset{A \subseteq \mathcal{E}}{\text{maximize}} \quad & \mathbb{E}\left[F(A)\right] \\ \text{subject to} \quad & |A| \leq k, \\ & |A_u| \leq k_u, \text{ for all } u \in V. \end{aligned}$$

We use A^* to denote the optimal solution of Problem 1, and $\mathrm{OPT} = \mathbb{E}\left[F(A^*)\right]$ to denote its expected score in G.

4 Theoretical Analysis

4.1 Possible-world semantics

A *probabilistic graph* $\mathcal{G} = (V, E, p)$, comprises a vertex set V and an edge set E, where each edge e is associated with a probability $p_e \in p$. Given a probabilistic graph, a *possible world* is a *deterministic* graph obtained from $\mathcal{G}$ with edges sampled independently according to p. We now introduce the *possible-world model* for our problem, that can capture the co-exposure of nodes to items resulting from any given assignment.

We start by defining a *directed edge-colored multigraph* $\widetilde{G} = (V, \widetilde{E}, \widetilde{p})$ from $G = (V, E)$, by creating h copies of each directed edge $(u, v) \in E$. For each item $i \in H$ we create a parallel edge $(u, v)_i$ in $\widetilde{G}$, having distinct color and associated probability p^i_{uv}. We interpret $\widetilde{G}$ as a probability distribution over all subgraphs of $(V, \widetilde{E})$, i.e., we sample each edge $(u, v)_i \in \widetilde{E}$ independently at random with probability p^i_{uv}. The probability of a possible world $g \sqsubseteq \widetilde{G}$ is given by

$$\Pr[g] = \prod_{i \in H} \prod_{(u,v)_i \in g} p^i_{uv} \prod_{(u,v)_i \in \widetilde{E} \setminus g} (1 - p^i_{uv}). \tag{4}$$

Let $\mathrm{path}^i_g(u, v)$ denote an indicator variable that equals 1 if node $v \in V$ is reachable by node u via the colored edges of i in g, and 0 otherwise. We say that a pair (u, i) can *color-reach* node v if $\mathrm{path}^i_g(u, v) = 1$. For an assignment A and a node $v \in V$ let $I^g_v(A)$ be the set of items that v is exposed to, due to A, in network g. It can be written as

$$I^g_v(A) = \{i \in H \mid \text{ exists } (u, i) \in A \text{ and } \mathrm{path}^i_g(u, v) = 1\}.$$

The value of the objective $\mathbb{E}\left[F(A)\right]$ in Problem 1 is given by

$$\begin{aligned} \mathbb{E}\left[F(A)\right] &= \mathbb{E}\left[\sum_{v \in V} f_v(I^g_v(A))\right] \\ &= \sum_{g \sqsubseteq \widetilde{G}} \Pr[g] \sum_{v \in V} f_v(I^g_v(A)). \end{aligned} \tag{5}$$

4.2 Hardness and approximation

We will first show that the objective function of Problem 1 is monotone and submodular.

Lemma 3. *The function $\mathbb{E}\left[F(\cdot)\right]$ is monotone and submodular.*

Proof. To prove the lemma, we utilize the possible-world semantics. It is well known that a non-negative linear combination of submodular functions is also submodular. Therefore, to prove submodularity of $\mathbb{E}\left[F(\cdot)\right]$, it is sufficient to show that in any possible world $g \sqsubseteq \widetilde{G}$, $f_v : 2^{\mathcal{E}} \to [0, 1]$ is submodular. Similarly, to prove monotonicity of $\mathbb{E}\left[F(\cdot)\right]$, it suffices to show the monotonicity of $f_v(\cdot)$ in any possible world g.

Now, recall that we have $f_v(I^g_v(A)) = 1 - \frac{1}{4} g_v(I^g_v(A))$. We will show that $g_v(I^g_v(A))$ is supermodular and monotonically non-increasing in A which will directly imply the submodularity and monotonicity of $f_v(I^g_v(A))$.

First we show that $g_v(I^g_v(A))$ is monotonically non-increasing in A by showing that $g_v(I^g_v(A)) \geq g_v(I^g_v(A \cup e))$ for any $A \subseteq \mathcal{E}$ and $(w, x) \in \mathcal{E} \setminus A$.

First, consider the case $path^x_g(w, v) = 0$. Notice that in this case we have $g_v(I^g_v(A)) = g_v(I^g_v(A \cup \{(w, x)\}))$ as $I^g_v(A) = I^g_v(A \cup \{(w, x)\})$. Now, consider the case $path^x_g(w, v) = 1$. In this case, we have $I^g_v(A \cup \{(w, x)\}) = I^g_v(A) \cup \{x\}$. Let $i, j \in I^g_v(A)$ be such that $\ell(i)$ and $\ell(j)$ are the immediate predecessor and successor of $\ell(x)$ in $L(v, I^g_v(A \cup \{(w, x)\}))$ respectively, i.e., $\nexists y \in L(v, I^g_v(A \cup \{(w, x)\}))$ such that $\ell(i) \leq \ell(y) \leq \ell(x)$ or $\ell(x) \leq \ell(y) \leq \ell(j)$.

Then we have,

$$\begin{aligned} & g_v(I^g_v(A \cup \{(w, x)\})) - g_v(I^g_v(A)) \\ &\quad = (\ell(i) - \ell(x))^2 + (\ell(x) - \ell(j))^2 - (\ell(i) - \ell(j))^2 \\ &\quad = (\ell(i) - \ell(x))^2 + (\ell(x) - \ell(j))^2 \\ &\qquad - (\ell(i) - \ell(x) + \ell(x) - \ell(j))^2 \\ &\quad \leq 0. \end{aligned}$$

We have just shown that $g_v(I^g_v(A))$ is monotonically non-increasing in A.

We now show that $g_v(I^g_v(A))$ is supermodular in A. Let $g_v(I^g_v((w, x) \mid A))$ denote the marginal decrease in the penalty when (w, x) is added to the assignment A:

$$g_v(I^g_v((w, x) \mid A)) = g_v(I^g_v(A \cup \{(w, x)\})) - g_v(I^g_v(A)).$$

To show that $g_v(I^g_v(\cdot))$ is supermodular, we need to show that

$$g_v(I^g_v((w, x) \mid A)) \leq g_v(I^g_v((w, x) \mid B)),$$

for any $A \subseteq B \subseteq \mathcal{E}$ and $(w, x) \notin B$.

Let $B = A \cup \{(z, y)\}$ for some $(z, y) \in \mathcal{E} \setminus A$. First, notice that if $\mathrm{path}^x_g(w, v) = 0$ and $\mathrm{path}^y_g(z, v) = 0$, then the analysis is trivial, since, $I^g_v((w, x) \mid A) = I^g_v((w, x) \mid B) = I^g_v(A)$, resulting in $g_v(I^g_v((w, x) \mid A)) = g_v(I^g_v((w, x) \mid B)) = 0$. Next, we provide the analysis for the case $\mathrm{path}^x_g(w, v) = 1$ and $\mathrm{path}^y_g(z, v) = 1$, and omit the analysis of the other two cases in which either $\mathrm{path}^x_g(w, v) = 0$ or $\mathrm{path}^y_g(z, v) = 0$ as their analysis use similar arguments.

We now start the analysis for the case $\mathrm{path}^x_g(w, v) = 1$ and $\mathrm{path}^y_g(z, v) = 1$. To do so, we perform case-by-case analysis based on how $\ell(x)$ is compares to the leanings in $L(v, I^g_v(A))$ and $L(v, I^g_v(B))$.

Let $i, j \in I_v^g(A)$ be such that $\ell(i)$ and $\ell(j)$ are the immediate predecessor and successor of $\ell(x)$ in $L(v, I_v^g(A \cup \{(w,x)\}))$. Next, we consider the following two cases.

- Case 1: y is such that $\ell(y) < \ell(i)$ or $\ell(y) > \ell(j)$. Then we have $g_v(I_v^g((w,x) \mid B)) = g_v(I_v^g((w,x) \mid A))$.
- Case 2: y is such that $\ell(i) \leq \ell(y) \leq \ell(x)$. Then we have

$$\begin{aligned} &g_v(I_v^g((w,x) \mid A)) - g_v(I_v^g((w,x) \mid B)) \\ &\quad = (\ell(i) - \ell(x))^2 - (\ell(i) - \ell(j))^2 - (\ell(y) - \ell(x))^2 \\ &\qquad + (\ell(y) - \ell(j))^2 \\ &\quad = -2\ell(i)\ell(x) + 2\ell(i)\ell(j) + 2\ell(y)\ell(x) - 2\ell(y)\ell(j) \\ &\qquad + (\ell(x) - \ell(j))(2\ell(y) - 2\ell(i)) \leq 0 \end{aligned}$$

- Case 3: y is such that $\ell(x) \leq \ell(y) \leq \ell(j)$. This case is symmetric to Case 2, so we omit the proof for brevity.

□

Theorem 1. *Problem 1 is* **NP***-hard.*

Proof. We will show that Problem 1 contains the influence maximization problem as a restricted special case, which is shown to be **NP**-hard [2]. Consider the case where H consists of a single item, which we denote by i'. Let $\ell(i') = 0$ and $\ell(v) = 1$, for all $v \in V$. Notice that since $h = 1$, the multiplicity of each edge in $\widetilde{G}$ is 1. In any $g \sqsubseteq \widetilde{G}$, if a node v is exposed to A then $f_v(I_v^g(A)) = \frac{1}{2}$, while $f_v(I_v^g(A)) = 0$ if v is not exposed to A. For any assignment A, define the seed set $S = \{u \mid (u, i') \in A\}$. Let $C_g(S)$ denote the number of nodes reachable by S in g. Notice the equivalence $C_g(S) = 2\sum_{v \in V} f_v(I_v^g(A))$ in g. Now, let $S^* \subseteq V$ denote the optimal solution to the influence maximization problem with parameter k. The expected spread of S^* is given by $\mathbb{E}[C_g(S^*)]$. If S^* is the seed set that maximizes $\mathbb{E}[C_g(S)]$, then, $A^* = \{(u, i') \mid u \in S^*\}$ is the assignment that maximizes $\mathbb{E}\left[\sum_{v \in V} f_v(I_v^g(A))\right]$. Thus, solving the influence maximization problem and obtaining S^*, yields the optimal solution A^* for Problem 1. □

Given the monotonicity and submodularity of the objective function, a standard greedy algorithm can be used to solve Problem 1. The pseudocode is given in Algorithm 1. Let $\mathbb{E}[F((u,i) \mid A)] = \mathbb{E}[F(A \cup \{(u,i)\})] - \mathbb{E}[F(A)]$ denote the marginal increase in the expected diversity exposure score of an assignment A if (u,i) is added to A. Let A_G denote the greedy solution. At each iteration, the greedy algorithm chooses the feasible pair (u^*, i^*) that yields the maximum gain in the expected diversity exposure score among all the feasible pairs.[3] The algorithm terminates when $|A_G| = k$.

Before we analyze the approximation guarantee of the greedy algorithm, we remind the reader of the following notions.

Definition 1 (Matroid)**.** *A set system $(\mathcal{E}, \mathcal{F})$, defined over a finite ground set $\mathcal{E}$ and a family $\mathcal{F}$ of subsets of $\mathcal{E}$, is a matroid $\mathcal{M} = (\mathcal{E}, \mathcal{F})$ if*

(i) $\mathcal{F}$ is non-empty;

(ii) $\mathcal{F}$ is downward closed, i.e., $X \in \mathcal{F}$ and $Y \subseteq X$ implies $Y \in \mathcal{F}$; and

(iii) $\mathcal{F}$ satisfies the augmentation property, i.e., for all $X, Y \in \mathcal{F}$ with $|Y| > |X|$, there exists an element $e \in Y \setminus X$ such that $X \cup \{e\} \in \mathcal{F}$.

3. We say that a pair (u,i) is feasible if it can be added to the current assignment A_G without breaking the attention bound constraint k_u.

Algorithm 1: Greedy Algorithm

Input : $\widetilde{G} = (V, \widetilde{E}, \widetilde{p})$; size constraint k; attention bound constraint k_u for all $u \in V$; leanings $\ell(i)$ for all $i \in H$, and $\ell(u)$ for all $u \in V$

Output: Greedy solution A_G

```
A_G ← ∅
while |A_G| ≤ k do
    (u*, i*) ← arg max_(u,i) E[F((u,i) | A_G)],
                  subject to |{i : (u,i) ∈ A_G}| ≤ k_u
    A_G ← A_G ∪ {(u*, i*)}
return A_G
```

Definition 2 (Uniform Matroid)**.** *A matroid $\mathcal{M} = (\mathcal{E}, \mathcal{F})$ is a uniform matroid if $\mathcal{F} = \{X \subseteq \mathcal{E} : |X| \leq k\}$.*

Definition 3 (Partition Matroid)**.** *Let $\mathcal{E}_1, \cdots, \mathcal{E}_Z$ be a partition of the ground set $\mathcal{E}$ into Z non-empty disjoint subsets. Let d_z be an integer with $0 \leq d_z \leq |\mathcal{E}_z|$, for each $z = 1, \ldots, Z$. A matroid $\mathcal{M} = (\mathcal{E}, \mathcal{F})$ is a partition matroid if $\mathcal{F} = \{X \subseteq \mathcal{E} : |X \cap \mathcal{E}_z| \leq d_z$, for all $z = 1, \cdots, Z\}$. In other words, a partition matroid contains exactly the sets $X \subseteq \mathcal{E}$ that share at most d_z elements with each subset $\mathcal{E}_z$.*

Lemma 4. *Given the ground set $\mathcal{E} = V \times H$ of user $\times$ item assignments, an integer k, and integers k_u, for all $u \in V$, let $\mathcal{F} \subseteq 2^{\mathcal{E}}$ denote the set of feasible solutions to Problem 1. Then, $\mathcal{M} = (\mathcal{E}, \mathcal{F})$ is a matroid defined on $\mathcal{E}$.*

Proof. To prove this result, we will show that (i) the constraint $|A| \leq k$ corresponds to a uniform matroid defined on $\mathcal{E}$, which we denote by $\mathcal{M}_k$; (ii) the constraints $|\{i : (u,i) \in A\}| \leq k_u$, for all $u \in V$, correspond to a partition matroid defined on $\mathcal{E}$, which we denote by $\mathcal{M}_p$; (iii) the intersection $\mathcal{M} = \mathcal{M}_k \cap \mathcal{M}_p$ is also a matroid defined on $\mathcal{E}$.

Let $\mathcal{F}_k \subseteq 2^{\mathcal{E}}$ denote the set of assignments of size at most k, i.e., $\mathcal{F}_k = \{A \subseteq \mathcal{E} : |A| \leq k\}$. It is easy to see that $\mathcal{M}_k = (\mathcal{E}, \mathcal{F}_k)$ is a uniform matroid.

Let $\mathcal{F}_p \subseteq 2^{\mathcal{E}}$ denote the set of assignments that do not violate any user attention bound constraint, i.e., for all $A \in \mathcal{F}_p$, we have $|\{i : (u,i) \in A\}| \leq k_u$, for all $u \in V$. Define $\mathcal{E}_u = \{(u,i) : i \in H\}$, for all $u \in V$. The sets $\mathcal{E}_u$, with $u \in V$, form a partition of $\mathcal{E}$ into n disjoint sets. Notice that an assignment $\mathcal{A} \subseteq \mathcal{E}$ can belong in $\mathcal{F}_p$ if and only if

$$|\mathcal{A} \cap \mathcal{E}_u| \leq k_u, \text{ for all } u \in V.$$

Hence, $\mathcal{M}_p = (\mathcal{E}, \mathcal{F}_p)$ is a partition matroid.

Notice that the set $\mathcal{F}$ of feasible solutions to Problem 1 is given by $\mathcal{F} = \mathcal{F}_k \cap \mathcal{F}_p$. Hence, the set system $(\mathcal{E}, \mathcal{F})$ corresponds to the intersection of matroids $\mathcal{M}_k$ and $\mathcal{M}_p$ that are both defined on $\mathcal{E}$. Note that the intersection of two matroids is not necessarily a matroid in general. However, in this case we have the intersection of a matroid with a uniform matroid, which is known to always result in a matroid; this operation is known as the *truncation* of a matroid [29]. □

Theorem 2. *Algorithm 1 achieves an approximation guarantee of* $1/2$.

Proof. As we have shown in Lemma 4, the constraints of Problem 1 correspond to a matroid defined on the ground set $\mathcal{E}$. Moreover, we have shown in Lemma 3 that the objective function

of Problem 1 is monotone and submodular. Thus, Problem 1 corresponds to monotone submodular function maximization subject to a matroid constraint.

Therefore, the approximation guarantee of Algorithm 1 thus follows from the result of Fisher et al. [30] for submodular function maximization subject to a matroid constraint. □

5 Scalable approximation algorithms

The efficient implementation of the greedy algorithm (Algorithm 1) is a challenge as the operation on line 3 translates to a large number of expected spread computations: in each iteration, the greedy algorithm requires to compute the expected marginal gain $\mathbb{E}\left[F((u,i) \mid A_G)\right]$ for every feasible pair (u,i), which in turn requires to identify the set $I_v^g(A \cup \{(u,i)\})$ of items that every v is exposed to in each $g \sqsubseteq \widetilde{G}$, which is akin to computing the expected influence spread when $h = 1$.

Computing the expected influence spread of a given set of nodes under the independent-cascade model is $\#\mathbf{P}$-hard [31]. A common practice is to estimate the expected spread using Monte Carlo (MC) simulations [2]. However, accurate estimation requires a large number of MC simulations.

Hence, considerable effort has been devoted in the literature to developing scalable approximation algorithms. Recently, Borgs et al. [3] introduced the idea of sampling *reverse-reachable* sets (RR-sets), and proposed a quasi-linear time randomized algorithm. Tang et al. improved it to a near-linear time randomized algorithm, called *Two-phase Influence Maximization* (TIM) [27], and subsequently tightened the lower bound on the number of random RR-sets required to estimate influence with high probability [26].

Random RR-sets are critical for efficient estimation of the expected influence spread. However, they are designed for the standard influence-maximization problem, which is a special case of Problem 1. We introduce a non-trivial generalization of reverse-reachable sets, which we name *reverse co-exposure* sets (RC-sets), and devise estimators for accurate estimation of the expected diversity exposure score $\mathbb{E}\left[F(\cdot)\right]$.

5.1 Reverse co-exposure sets

Recall that we can interpret $\widetilde{G}$ as a probability distribution over all subgraphs of $(V, \widetilde{E})$, where each edge $(u,v)_i \in \widetilde{E}$ is realized with probability p_{uv}^i. Let $g \sim \widetilde{G}$ be a graph drawn from the random graph distribution $\widetilde{G}$. Notice that, over the randomness in g, the set $I_v^g(A)$ can be regarded as a Multinoulli random variable with 2^h outcomes, where each outcome corresponds to one of the subsets of H. Now, let $\widetilde{R}_{v,g} \subseteq \mathcal{E}$ denote the set of pairs in g that can color-reach v, i.e., $\widetilde{R}_{v,g} = \{(u,i) \in \mathcal{E} : \mathrm{path}_g^i(u,v) = 1\}$. Also let

$$I(A \cap \widetilde{R}_{v,g}) = \{i \in H : (u,i) \in A \cap \widetilde{R}_{v,g}\}.$$

The following lemma establishes the activation equivalence property that forms the foundations of random *reverse co-exposure* sets (RC-sets).

Lemma 5. *Let I be a subset of H. For any assignment A and for all $v \in V$, we have*

$$\Pr_{g\sim\widetilde{G}}\left(I_v^g(A) = I\right) \;=\; \Pr_{g\sim\widetilde{G}}(I(A \cap \widetilde{R}_{v,g}) = I).$$

Proof. Notice that in any possible world g, we have:

$$\begin{aligned} I_v^g(A) &= \{i \in H : \exists\,(u,i) \in A \text{ such that } \mathrm{path}_g^i(u,v) = 1\} \\ &= \{i \in H : (u,i) \in A \cap \widetilde{R}_{v,g}\} \\ &= I(A \cap \widetilde{R}_{v,g}). \end{aligned}$$

Hence we have

$$\begin{aligned} \Pr_{g\sim\widetilde{G}}\left(I_v^g(A) = I\right) &= \sum_{g \sqsubseteq \mathcal{G}} Pr[g]\; \mathbb{1}_{[I_g(A)=I]} \\ &= \sum_{g \sqsubseteq \mathcal{G}} Pr[g]\; \mathbb{1}_{[I(A\cap\widetilde{R}_{v,g})=I]} \\ &= \Pr_{g\sim\widetilde{G}}(I(A \cap \widetilde{R}_{v,g}) = I). \end{aligned}$$

□

Next we formally define the concept of random RC-sets.

Random RC-sets. Given a probabilistic multi-graph $\widetilde{G} = (V, \widetilde{E}, \widetilde{p})$ and a set H of items, a random RC-set $\widetilde{R}_{v,g}$ is generated as follows. First, we remove each edge $(u,v)_i$ from $\widetilde{G}$ with probability $1 - p_{uv}^i$, generating thus a possible world g. Next, we pick a *target* node v uniformly at random from V. Then, $\widetilde{R}_{v,g}$ consists of the pairs that can *color-reach* v, i.e., all pairs (u,i) for which $\mathrm{path}_g^i(u,v) = 1$.

Sampling a random RC-set $\widetilde{R}_{v,g}$ can be implemented efficiently by first choosing a target node $v \in V$ uniformly at random and then performing a breadth-first search (BFS) from v in $\widetilde{G}$. Notice that a random RC-set $R_{v,g}$ is subject to two levels of randomness: (i) randomness over $g \sim \widetilde{G}$, and (ii) randomness over the selection of target node $v \sim V$.

Lemma 6. *For any random* RC*-set $R_{v,g}$, let the random variable $w(A \cap \widetilde{R}_{v,g}) = f_v(I(A \cap \widetilde{R}_{v,g}))$ represent the diversity exposure weight of A on $\widetilde{R}_{v,g}$. Then, $\mathbb{E}\left[F(A)\right] = n \mathop{\mathbb{E}}_{v,g}\left[w(A \cap \widetilde{R}_{v,g})\right]$, where the expectation is taken over the randomness in $v \sim V$ and $g \sim \widetilde{G}$.*

Proof. First, notice that over the randomness in g, $f_v(I_v^g(A))$ is a function of a random variable $I_v^g(A)$, hence, by the LOTUS theorem [32], which defines expectation for functions of random variables, its expectation can be computed as

$$\mathop{\mathbb{E}}_{g}\left[f_v(I_v^g(A))\right] = \sum_{I \in 2^H} \Pr_g\left(I_v^g(A) = I\right)\; f_v(I). \tag{6}$$

Then, by Equation (6) and the activation equivalence property shown in Lemma 5, we have

$$\begin{aligned} \mathbb{E}\left[F(A)\right] &= \mathop{\mathbb{E}}_{g}\left[\sum_{v\in V} f_v(I_v^g(A))\right] \\ &= \sum_{v\in V} \mathop{\mathbb{E}}_{g}\left[f_v(I_v^g(A))\right] \\ &= \sum_{v\in V}\sum_{I\in 2^H} \Pr_g\left(I_v^g(A) = I\right)\; f_v(I) \\ &= n \sum_{I\in 2^H} \Pr_{v,g}(I(A \cap \widetilde{R}_{v,g}) = I)\, f_v(I) \\ &= n \mathop{\mathbb{E}}_{v,g}\left[f_v(I(A \cap \widetilde{R}_{v,g}))\right]. \end{aligned}$$

□

Lemma 6 shows that we can estimate $\mathbb{E}\left[F(A)\right]$ by estimating $n\,\mathbb{E}\left[f_v(I(A \cap \widetilde{R}_{v,g}))\right]$ on a set of random RC-sets. This

Algorithm 2: TDEM $(\widetilde{G}, k, l, \epsilon, \ell)$

$\widetilde{\mathcal{R}} \leftarrow \text{Sampling}(\widetilde{G}, k, \epsilon, \ell)$
$\tilde{A} \leftarrow \text{RC-Greedy}(\widetilde{\mathcal{R}}, k, l)$
return $\tilde{A}$

Algorithm 3: RC-Greedy$(\widetilde{\mathcal{R}}, k, l)$

$\tilde{A} \leftarrow \emptyset$
while $|\tilde{A}| \leq k$ **do**
  $(u^*, i^*) \leftarrow \arg\max_{(u,i)} \mathcal{W}_{\widetilde{\mathcal{R}}}((u,i) \mid \tilde{A})$, subject to $|\{i : (u,i) \in \tilde{A}\}| \leq k_u$
  $\tilde{A} \leftarrow \tilde{A} \cup \{(u^*, i^*)\}$
return $\tilde{A}$

suggests that if we have a sample $\widetilde{\mathcal{R}}$ of random RC-sets from which we can obtain, with high probability, accurate estimations of $\mathbb{E}[F(A)]$ for every assignment A such that $|A| \leq k$, then, we can accurately solve Problem 1 on the sample $\widetilde{\mathcal{R}}$ with high probability, as we show next.

Given a sample $\widetilde{\mathcal{R}}$ of random RC-sets, let

$$\mathcal{W}_{\widetilde{\mathcal{R}}}(A) = \frac{\sum_{\widetilde{R}_{v,g} \in \widetilde{\mathcal{R}}} w(A \cap \widetilde{R}_{v,g})}{|\widetilde{\mathcal{R}}|},$$

denote the diversity exposure weight of A on the sample. Notice that, as a direct consequence of Lemma 6, the quantity $n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(A)$ is an unbiased estimator of $\mathbb{E}[F(A)]$.

Moreover, let

$$\mathcal{W}_{\widetilde{\mathcal{R}}}((u,i) \mid A) = \mathcal{W}_{\widetilde{\mathcal{R}}}(A \cup \{(u,i)\}) - \mathcal{W}_{\widetilde{\mathcal{R}}}(A),$$

denote the marginal increase in the diversity exposure weight of A if the pair (u,i) is added to A.

5.2 Two-phase Diversity Exposure Maximization

We now present our Two-phase Diversity Exposure Maximization algorithm (TDEM), which provides an approximate solution to Problem 1. The pseudocode is shown in Algorithm 2. As it names suggests, TDEM operates in two phases: a *sampling* phase and a *greedy pair-selection* phase. In the sampling phase, a sample $\widetilde{\mathcal{R}}$ of random RC-sets is generated (details later). This sample is provided as input to RC-Greedy (Algorithm 3), which greedily selects feasible pairs (u,i) into $\tilde{A}$. The algorithm terminates when $|\tilde{A}| = k$ and it returns $\tilde{A}$ as a solution to Problem 1.

Theorem 3. *Assume that the algorithm* RC-Greedy *receives as input a sample $\widetilde{\mathcal{R}}$ of random* RC*-sets such that for any assignment A of size at most k it holds that*

$$\left|n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(A) - \mathbb{E}[F(A)]\right| < \frac{\epsilon}{2}\,\mathrm{OPT}, \tag{7}$$

with probability at least $1 - n^{-\ell}/\binom{nh}{k}$. Then, RC-Greedy *returns a $(\frac{1}{2} - \epsilon)$-approximate solution to Problem 1 with probability at least $1 - n^{-\ell}$. The running time of* RC-Greedy *is $\mathcal{O}(\sum_{\widetilde{R} \in \widetilde{\mathcal{R}}} |\widetilde{R}|)$, that is, linear in the total size of the* RC*-sets in the sample.*

Proof. First, notice that, $\mathcal{W}_{\widetilde{\mathcal{R}}}(\cdot)$ is a linear combination of submodular $f_v(\cdot)$'s, hence is also submodular. Moreover, the activation equivalence property depicted in in Lemma 5 shows that we can approximately solve Problem 1 by finding the assignment that maximizes $\mathcal{W}_{\widetilde{\mathcal{R}}}(\cdot)$ on a sample $\widetilde{\mathcal{R}}$ of RC-sets. Now, given that the size of $\widetilde{\mathcal{R}}$ is such that, the diversity exposure score of any assignment of size at most k is accurately estimated w.p. at least $1 - n^{-\ell}/\binom{nh}{k}$, it follows, via union bound, that w.p. at least $1 - 1/n^{\ell}$ we have:

$$\mathbb{E}\left[F(\tilde{A}^G)\right] \geq \mathbb{E}\left[F(A^G)\right] - \epsilon\,\mathrm{OPT} \tag{8}$$

where A^G is the real greedy solution and $\tilde{A}^G$ is the approximate greedy solution that TDEM returns. Note that, $n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A}^G) \geq n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(A^G)$ is the greedy solution obtained on $\widetilde{\mathcal{R}}$.

The correctness of Equation 8 follows from the following case analysis: (*i*) $\tilde{A}^G$ is the real greedy solution A^G to Problem 1; (*ii*) $\tilde{A}^G$ is an assignment with $\mathbb{E}\left[F(\tilde{A}^G)\right] > \mathbb{E}\left[F(A^G)\right]$; or (*iii*) $\tilde{A}^G$ is an assignment with $\mathbb{E}\left[F(\tilde{A}^G)\right] < \mathbb{E}\left[F(A^G)\right]$ such that its maximum possible accurate estimate (that satisfies Equation 7) is higher than the minimum possible accurate estimate of $\mathbb{E}\left[F(A^G)\right]$, hence is returned by RC-Greedy instead of A^G, i.e.,

$$\begin{aligned}\mathbb{E}\left[F(\tilde{A}^G)\right] + \frac{\epsilon}{2}\,\mathrm{OPT} &\geq n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A}^G)\\ &\geq n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(A^G)\\ &\geq \mathbb{E}\left[F(A^G)\right] - \frac{\epsilon}{2}\,\mathrm{OPT}.\end{aligned}$$

Obviously, the approximation guarantee does not deteriorate from $(1/2)$ for the first two cases. For case (*iii*) we have:

$$\begin{aligned}\mathbb{E}\left[F(\tilde{A}^G)\right] &\geq \mathbb{E}\left[F(A^G)\right] - \epsilon\,\mathrm{OPT}\\ &\geq (1/2)\,\mathrm{OPT} - \epsilon\,\mathrm{OPT}.\end{aligned}$$

Therefore the result follows.

Now we analyze the running time of RC-Greedy. First, we remind that the running of the greedy algorithm on RR sets, for approximately solving the influence maximization problem, follows from the running time of the maximum cover problem [27]. For the analysis of RC-Greedy, we use a similar reasoning and exploit a connection to the weighted version of the maximum coverage problem. However, we note that our problem does not correspond to the weighted maximum coverage problem since (i) we are interested in the weights of RC-sets even in the case when they have been already covered by a pair (u,i),[4] (ii) the weights of the ground set elements (which correspond to RC-sets) dynamically change based on the pairs that already covered them. However, these differences do not affect the running time analysis much. The constant time operation to check whether an RC-set is covered by a pair (u,i) is replaced by the operation of finding the next smaller and next larger labels compared to $l(i)$ from the labels of the items that have previously covered this RC-set. Using binary search, this can be done in logarithmic time.

Since this operation is independent of the seed node u, the number of "covered" checks performed on each RC-set is upper-bounded by the size of the RC-set, times a logarithmic factor as explained above. Hence, the total running time complexity of RC-Greedy is $\mathcal{O}(\sum_{\widetilde{R} \in \widetilde{\mathcal{R}}} |\widetilde{R}| \log(|\widetilde{R}|))$. □

Let θ^* be the minimum sample size such that Equation (7) holds for all assignments of size at most k. Notice that since the desired estimation accuracy is a function of OPT, the value

4. We say that a pair (u,i) covers an RC-set $\widetilde{R}$ if $(u,i) \in \widetilde{R}$

of θ^* also depends on OPT, which is unknown and in fact **NP**-hard to compute. To circumvent the problem we follow a similar approach to TIM [27] and IMM [26]: we estimate a lower bound on the value of the optimal solution, and use it for the determination of the sample size. We also generalize the *statistical test* employed by IMM [26] for estimating a lower bound when working with random RC-sets. Note that the results from influence maximization do not carry over to our case, therefore our extension of the technique is non-trivial.

5.3 Determining the sample size

Let $\widetilde{R}_1, \widetilde{R}_2, \ldots, \widetilde{R}_\theta$ be the sequence of random RC-sets generated in the sampling phase of TDEM. For a given assignment A, let w_j denote its weight on the RC-set $\widetilde{R}_j$. Notice that the choices of v and g during the creation of $\widetilde{R}_j$ are independent of $\widetilde{R}_1, \ldots, \widetilde{R}_{j-1}$. However, as we will see soon, the sampling phase of TDEM employs an adaptive procedure, in which the decision to generate $\widetilde{R}_j$ depends on the outcomes of $\widetilde{R}_1, \ldots, \widetilde{R}_{j-1}$. This creates dependencies between the RC-sets in the sample $\widetilde{R}$. Thus, we can only use concentration inequalities that allow dependencies in the sample. We first introduce the notions that are crucial in our analysis.

Definition 4 (Martingale)**.** *A sequence $X_1, X_2, \ldots$ of random variables is a martingale if and only if $\mathbb{E}[|X_j|] < +\infty$ and $\mathbb{E}[X_j \mid X_1, \ldots, X_{j-1}] = X_{j-1}$ for any j.*

We now establish the connections to martingales. Let $w = \mathbb{E}\left[F(A)\right]/n$. By Lemma 6 we have $\mathbb{E}[w_j] = w$, for all $j \in [1, \theta]$. Noting that the choice of v and g during the creation of $\widetilde{R}_j$ is independent of $\widetilde{R}_1, \ldots, \widetilde{R}_{j-1}$, we have

$$\mathbb{E}[w_j \mid w_1, \ldots, w_{j-1}] = \mathbb{E}[w_j] = w.$$

Let $M_j = \sum_{z=1}^{j}(w_z - w)$, so $\mathbb{E}[M_j] = 0$, and

$$\begin{aligned}\mathbb{E}[M_j \,|\, M_1, \ldots, M_{j-1}] &= \mathbb{E}[M_{j-1} + w_j - w \,|\, M_1, \ldots, M_{j-1}] \\ &= M_{j-1} - w + \mathbb{E}[w_j \,|\, M_1, \ldots, M_{j-1}] \\ &= M_{j-1} - w + \mathbb{E}[w_j] \\ &= M_{j-1},\end{aligned}$$

therefore, the sequence $M_1, \ldots, M_\theta$ is a martingale.

The following lemma from Chung and Lu [33] shows a concentration result for martingales, analogous to Chernoff bounds for independent random variables.

Lemma 7. *[Theorem 6.1, [33]] Let $X_1, X_2, \ldots$ be a martingale, such that $X_1 \leq a$, $\mathrm{Var}[X_1] \leq b_1$, $|X_z - X_{z-1}| \leq a$ for $z \in [2, j]$, and*

$$\mathrm{Var}[X_z \mid X_1, \ldots, X_{z-1}] \leq b_j, \text{ for } z \in [2, j],$$

where $\mathrm{Var}[\cdot]$ denotes the variance. Then, for any $\gamma > 0$

$$\Pr\left(X_j - \mathbb{E}[X_j] \geq \gamma\right) \leq \exp\left(-\frac{\gamma^2}{2(\sum_{z=1}^{j} b_z + a\gamma/3)}\right)$$

We now discuss how to use this concentration result for the martingale $M_1, \ldots, M_\theta$. Notice that since $w_j \in [0, 1]$ for all $j \in [1, \theta]$, we have $|M_1| = |w_1 - w| \leq 1$ and $|M_j - M_{j-1}| \leq 1$ for any $j \in [2, \theta]$. We also have $\mathrm{Var}[M_1] = \mathrm{Var}[w_1]$, and for any $j \in [2, \theta]$

$$\begin{aligned}&\mathrm{Var}[M_j \mid M_1, \ldots, M_{j-1}] \\ &\quad= \mathrm{Var}[M_{j-1} + w_j - w \mid M_1, \ldots, M_{j-1}] \\ &\quad= \mathrm{Var}[w_j \mid M_1, \ldots, M_{j-1}] \\ &\quad= \mathrm{Var}[w_j].\end{aligned}$$

Recall that $f_v(I_v^g(A))$ is a function of the Multinoulli random variable $I_v^g(A)$, hence, $w(A \cap \widetilde{R}_{v,g}) = f_v(I(A \cap \widetilde{R}_{v,g}))$. Based on the LOTUS theorem [32] again, we have

$$\mathbb{E}[f_v(I(A \cap \widetilde{R}_{v,g}))^2] = \sum_{I \in 2^H} \Pr_{v,g}(I(A \cap \widetilde{R}_{v,g}) = I)\,(f_v(I))^2.$$

Hence, we can bound the variance as follows

$$\begin{aligned}\mathrm{Var}[f_v(I(A \cap \widetilde{R}_{v,g}))] &= \mathbb{E}[f_v(I(A \cap \widetilde{R}_{v,g}))^2] - w^2 \\ &\leq \mathbb{E}[f_v(I(A \cap \widetilde{R}_{v,g}))^2] \\ &= \sum_{I \in 2^H} \Pr_{v,g}(I(A \cap \widetilde{R}_{v,g}) = I)\,(f_v(I))^2 \\ &\leq \sum_{I \in 2^H} \Pr_{v,g}(I(A \cap \widetilde{R}_{v,g}) = I)\, f_v(I) \\ &= w,\end{aligned}$$

where the last inequality follows from the fact that $f_v(\cdot)$ is bounded by 1. Therefore, $\mathrm{Var}[w_j] \leq w$ for all $j \in [1, \theta]$. Then, by using Lemma 7, for $M_\theta = \sum_{j=1}^{\theta}(w_j - w)$, with $\mathbb{E}[M_\theta] = 0$, $a = 1$, $b_j = w$, for $j = 2, \ldots, \theta$, and $\gamma = \delta\theta w$, we have the following corollary.

Corollary 1. *For any $\delta > 0$,*

$$\Pr\left(\sum_{j=1}^{\theta} w_j - \theta w \geq \delta\theta w\right) \leq \exp\left(-\frac{\delta^2}{\frac{2\delta}{3} + 2}\,\theta w\right).$$

Moreover, for the martingale $-M_1, \ldots, -M_\theta$, we similarly have $a = 1$ and $b_j = w$ for $j = 1, \ldots, \theta$. Note also that $\mathbb{E}[-M_\theta] = 0$. Hence, for $-M_\theta = \sum_{j=1}^{\theta}(w - w_j)$ and $\gamma = \delta\theta w$ we can obtain:

Corollary 2. *For any $\delta > 0$,*

$$\Pr\left(\sum_{j=1}^{\theta} w_j - \theta w \leq -\delta\theta w\right) \leq \exp\left(-\frac{\delta^2}{\frac{2\delta}{3} + 2}\,\theta w\right).$$

We will use these corollaries frequently. We are now ready to start our analysis. We first provide a lower bound on the sample size, which depends on OPT.

Lemma 8. *Let $\theta = |\widetilde{\mathcal{R}}|$ denote the size of the random* RC*-sets returned by the sampling phase of* TDEM*. Suppose that θ satisfies*

$$\theta \geq 2n(\epsilon + 6)\,\frac{\ln\binom{nh}{k} + \ell \ln n + \ln 2}{3\epsilon^2\,\mathrm{OPT}}. \tag{9}$$

Then, for any assignment A of size at most k, the following holds with probability at least $1 - n^{-\ell}/\binom{nh}{k}$

$$\left|n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(A) - \mathbb{E}\left[F(A)\right]\right| < \frac{\epsilon}{2}\,\mathrm{OPT}. \tag{10}$$

For better readability, we have included the proof of Lemmas 8, 9, and 10 in the supplementary material.

As stated in Theorem 3 the greedy pair selection phase of TDEM requires as input a sample $\widetilde{\mathcal{R}}$ of random RC-sets such that Equation (7) holds for all assignments of size at most k. As shown in Lemma 8, this requirement translates to the lower bound $|\widetilde{\mathcal{R}}| \geq \lambda/\mathrm{OPT}$, where

$$\lambda = 2n(\epsilon+6)\,\frac{\ln\binom{nh}{k} + \ell \ln n + \ln 2}{3\epsilon^2}. \tag{11}$$

Given that OPT is unknown and **NP**-hard to compute, our objective is to identify a lower bound on OPT, which is as tight as possible, so as to reduce the computational cost of generating the sample $\widetilde{R}$. To achieve this goal, we extend the technique introduced by IMM and we perform a statistical test $B(x)$, such that if $\mathrm{OPT} < x$ then $B(x) = \texttt{false}$ with high probability. Given that $\mathrm{OPT} \in (0, n]$ and using the value of the greedy solution as an indicator of the magnitude of OPT, we can identify a lower bound on OPT by running the test $B(x)$ on $\mathcal{O}(\log_2 n)$ values of x, i.e., $x = n/2, n/4, \ldots, 2$.

We now give details of our sampling algorithm, which first adaptively estimates a lower bound on the value of OPT by employing the statistical test, and then it keeps generating random RC-sets into $\widetilde{R}$ until $|\widetilde{R}| \geq \lambda/\mathrm{LB}$.

The sampling algorithm, pseudocode provided in Algorithm 4, first sets $\widetilde{\mathcal{R}} = \emptyset$ and initializes LB to a naïve lower bound — which we will explain soon. Then, it enters a `for`-loop with at most $\log_2 n$ iterations. In the i-th iteration, the algorithm computes $x = n/2^i$ and derives

$$\theta_i = \frac{(\frac{2\epsilon}{3}+2)\,\left(\ln\binom{nh}{k} + \ell\ln n + \ln\log_2 n\right)}{\epsilon^2}\,\frac{n}{x}.$$

Then the Algorithm inserts more random RC-sets into $\widetilde{\mathcal{R}}$ until $|\widetilde{\mathcal{R}}| \geq \theta_i$ and invokes RC-Greedy (Algorithm 3). If $\widetilde{\mathcal{R}}$ satisfies the following *stopping condition*

$$n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A}) \geq (1+\epsilon)\,x, \tag{12}$$

the algorithm sets the lower bound $\mathrm{LB} = \frac{n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A})}{1+\epsilon}$ and terminates the `for`-loop. If this is the case, then algorithm generates more random RC-sets into $\widetilde{R}$ until $|\widetilde{R}| \geq \frac{\lambda}{\mathrm{LB}}$ and returns $\widetilde{\mathcal{R}}$. Otherwise, the algorithm proceeds with the $(i+1)$-th iteration. If after $\mathcal{O}(\log_2 n)$ iterations the algorithm cannot set LB, then it uses the naïve lower bound and generates random RC-sets into $\widetilde{\mathcal{R}}$ until $|\widetilde{\mathcal{R}}| \geq \lambda/\mathrm{LB}_0$. The naïve bound LB_0 corresponds to the value of the minimum possible solution on the input instance for any positive integer k, hence, we set $\mathrm{LB}_0 = 1 - \frac{1}{4}\max_{(v,i)\in\mathcal{E}} g_v(\{i\})$.[5]

The following theorem gives the correctness of Algorithm 4.

Theorem 4. *With probability at least $1-n^{-\ell}$, Algorithm 4 returns a sample $\widetilde{\mathcal{R}}$ such that $|\widetilde{\mathcal{R}}| \geq \lambda/\mathrm{OPT}$.*

To prove Theorem 4, we first establish the following two lemmas, for which the proofs can be found in the supplementary material.

Lemma 9. *Assume that we invoke algorithm* RC-Greedy *on a sample $\widetilde{\mathcal{R}}$ of θ random* RC*-sets such that*

$$\theta \geq \frac{(\frac{2\epsilon}{3}+2)\,\left(\ln\binom{nh}{k} + \ell\ln n + \ln\log_2 n\right)}{\epsilon^2}\,\frac{n}{x}.$$

5. Notice that this is analogous to IMM's naive lower bound for the influence maximization problem that is equal to 1.

Algorithm 4: $\mathrm{Sampling}(\widetilde{G}, k, \epsilon, \ell)$

```
R̃ ← ∅ ;
LB ← LB_0 ;
for i = 1, ..., log_2 n − 1 do
    x ← n/2^i ;
    θ_i = ((2ε/3 + 2)(ln C(nh, k) + ℓ ln n + ln log_2 n)) / ε^2 · n/x ;
    while |R̃| ≤ θ_i do
        R̃ ← R̃ ∪ GenerateRC-Set;
    Ã_i ← RC-Greedy(R̃, k, l) ;
    if n W_R̃(Ã_i) ≥ (1 + ε) x, then
        LB ← n W_R̃(Ã) / (1+ε) ;
        break;
    θ ← λ/LB;
while |R̃| ≤ θ do
    R̃ ← R̃ ∪ GenerateRC-Set ;
return R̃
```

Let $\tilde{A}$ be the solution returned by the RC-Greedy*. If $n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A}) \geq (1+\epsilon)\,x$, then $\mathrm{OPT} \geq x$ with probability at least $1 - \frac{n^{-\ell}}{\log_2 2n}$.*

Lemma 10. *Assume x, ϵ, $\widetilde{\mathcal{R}}$, and $\tilde{A}$ are defined as in Lemma 9. If $\mathrm{OPT} \geq x$ then $n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A}) \leq (1+\epsilon)\,\mathrm{OPT}$ with probability at least $1 - \frac{n^{-\ell}}{\log_2 n}$.*

We are now ready prove Theorem 4.

Proof of Theorem 4. Let $i^* = \lceil \log_2 \frac{n}{\mathrm{OPT}} \rceil$. We will first show that the probability the stopping condition holds while $\mathrm{OPT} < x$ is at most $(i^*-1)/(n^\ell \log_2 n)$. Recall that the value of x is determined by $n/2^i$ at each iteration i. Therefore, for any $i < i^*$, we have $x = n/2^i < \mathrm{OPT}$. Hence, by Lemma 9 and the union bound over i^*-1 iterations, the probability that $\mathrm{OPT} < x$ and $n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A})/(1+\epsilon) \geq x$ is at most $(i^*-1)/(n^\ell \log_2 n)$. Moreover, it follows from Lemma 10 that the probability that $\mathrm{OPT} \geq x$ and $n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A}) > (1+\epsilon)\,\mathrm{OPT}$ is at most $1/(n^\ell \log_2 n)$. Hence, when the stopping condition holds, by union bound, the probability that $\mathrm{OPT} \geq x$ and $n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A}) \leq (1+\epsilon)\,\mathrm{OPT}$ is at least

$$1 - \left(\frac{i^*-1}{n^\ell \log_2 n} + \frac{1}{n^\ell \log_2 n}\right) \geq 1 - n^{-\ell}.$$

Then by Lemma 10 and the union bound, it follows that with probability at least $1-n^{-\ell}$, we have

$$\mathrm{OPT} \geq \frac{n\,\mathcal{W}_{\widetilde{\mathcal{R}}}(\tilde{A})}{1+\epsilon} \geq x.$$

Therefore, the algorithm sets $\mathrm{LB} \geq \mathrm{OPT}$ with probability at least $1-n^{-\ell}$ and returns a sample $\widetilde{\mathcal{R}}$ such that

$$|\widetilde{\mathcal{R}}| \geq \frac{\lambda}{\mathrm{LB}} \geq \frac{\lambda}{\mathrm{OPT}}$$

with probability at least $1-n^{-\ell}$. □

6 Experiments

In this section, we evaluate our proposed algorithm on a range of real-world datasets.

TABLE 1
Statistics of the datasets.

Dataset	n	m	$d(G)$	ℓ	ℓ^2	p^i_{uv}		
				avg	avg	min	avg	max
DBLP:BSch	167	634	3.80	−0.60	0.50	0.034	0.116	0.249
DBLP:CPap	144	800	5.56	−0.26	0.28	0.034	0.117	0.247
DBLP:PYu	342	1964	5.74	−0.52	0.42	0.034	0.118	0.249
TPair:X	140	1372	9.80	−0.03	0.34	0.034	0.112	0.249
TPair:Y	338	8436	24.96	−0.07	0.43	0.034	0.107	0.249
TPair:Z	577	24427	42.33	0.12	0.36	0.034	0.113	0.250
Tweet:S5	2719	7714	2.84	0.24	0.52	0.034	0.114	0.249
Tweet:S2	4379	27765	6.34	0.26	0.52	0.034	0.113	0.249
Tweet:M5	5183	50165	9.68	0.26	0.51	0.034	0.113	0.250
Twitt:Follow	5454	835725	153.23	0.27	0.52	0.034	0.116	0.250
G:Brexit	22745	48830	2.15	0.65	0.72	0.010	0.014	0.110
G:IPhone	36742	49248	1.34	0.87	0.90	0.010	0.053	1.000
G:US-elect	23816	844700	35.47	0.46	0.75	0.010	0.013	0.043
G:Abortion	279505	670501	2.40	0.02	0.80	0.010	0.011	0.110
G:Fracking	374403	1366909	3.65	0.55	0.61	0.010	0.011	0.110
G:ObamaC	334617	1511670	4.52	0.12	0.61	0.010	0.012	0.110
Twitt:XL	481523	52378856	108.78	0.07	0.39	4.2e-5	0.028	1.000

6.1 Datasets

In our experiments, we use five collections of networks, one based on data from the DBLP bibliographic database, and the other four collected from Twitter.

The first collection consists of the one-hop egonets of three well-known researchers: B. Schneiderman (DBLP:BSch), P. Yu (DBLP:PYu) and C. Papadimitriou (DBLP:CPap). Node leanings are derived from publication-venue information using the method proposed by Galbrun et al. [34].

Twitt:Follow is the Twitter follower network obtained by Lahoti et al. [35] and Twitt:XL is a larger variant of this same network. For node leanings we use rescaled ideology scores from Barberá et al. [36]. From the same harvest of tweets as Twitt:Follow, we construct two additional collections of networks. The first collection contains the networks TPair:X, TPair:Y, and TPair:Z. Each of these networks is obtained by selecting a pair of users who have opposite leanings but share neighbors, and extracting the neighborhood. The second collection contains the networks Tweet:S5, Tweet:S2, and Tweet:M5. Instead of follower-followee relationships, these networks capture actual exchanges of tweets between users, with increasing requirements on the strength of the exchanges.

The last collection consists of the six networks from the study by Garimella et al. [16]: G:Abortion, G:Brexit, G:Fracking, G:IPhone, G:ObamaC and G:US-elect, Each network represents a Twitter follower network focused around topics with two opposing sides. We obtain node leanings from the estimated probabilities of users to retweet content from either of the opposing sides.

Note that solving our problem for h items on a network with m edges requires maintaining in memory a multigraph of $h \times m$ edges, which is analogous to the requirement for solving the standard influence maximization problem on a graph with $h \times m$ edges. Hence, our largest configuration, Twitt:XL with 25 items, effectively yields a graph with $52.5\,\mathrm{M} \times 25 \approx 1.3\,\mathrm{B}$ edges and is comparable to IMM's largest dataset of 1.5 B edges [26].

For the largest configuration, Twitt:XL, following IMM [26] to retain comparability, we use the weighted-cascade model [2] that assigns $p^i_{u,v} = 1/|N^{in}(v)|$ for each item and edge. For the rest of the datasets, the propagation probabilities of items along the edges of the network depend on the leaning of the item being propagated, and on the leanings of the emitting and receiving users. Intuitively, the further away from the leaning of the users, the less likely an item is to be propagated.

We consider an exponential function to model how the propagation probability drops as the leaning of the item lies further away from that of the communicating nodes. More specifically, we use an exponential function with parameters β and γ:

$$\Phi_{\beta,\gamma}(u,v,i) = \beta \exp(-\gamma \max(|\ell(u)-\ell(i)|, |\ell(v)-\ell(i)|)/2) .$$

We set $\beta = 0.25$ for all collections except G for which we use the edge probabilities present in the network as values for β and add a 0.01 offset to all resulting values, in order to obtain reasonable propagation probabilities. We experiment with probabilities obtained with the exponential function, letting $\gamma = 2$. We compare the propagation probabilities resulting from this function to an exponential function with $\gamma = 4$ as well as a linear function. A heatmap of the resulting propagation probabilities can be found in the supplementary material.

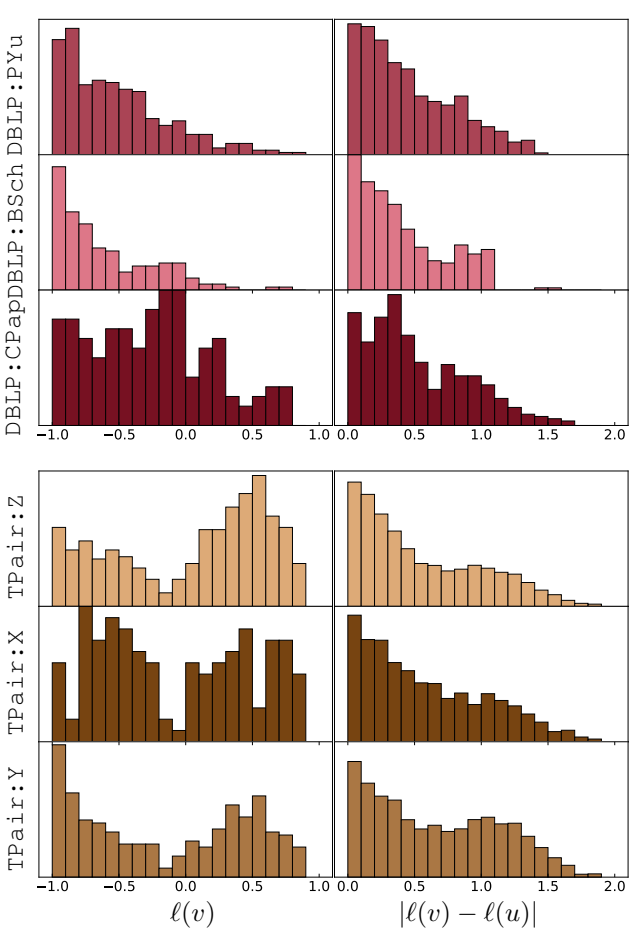
We use 25 items with leanings evenly spread over the interval $[-1, 1]$ as our pool of items in all the setups. For the smaller datasets, we look for assignments of size $k = 5$ with an attention bound $k_u = 1$, while for larger datasets we use $k = 50$ and $k_u = 5$. Following [26], we set $\epsilon = 0.2$ and $\ell = 1$ in all the experiments.

Table 1 shows the basic statistics of the datasets used in our experiments. For each dataset, we indicate the number of nodes (n), the number of edges (m), the density of the graph ($d(G) = m/n$), the average node leaning (ℓ), the squared node leaning (ℓ^2), as well as the minimum, average and maximum propagation probabilities, over all edges and items in the network (p^i_{uv}).

Figures 1–3 show histograms of node leanings and leaning differences across the edges of each network from the different collections.

6.2 Comparison baselines

To better understand the quality of the returned assignments, we compare the solution of our algorithm to item–user assignments obtained with simple yet intuitive baselines. Recall that the running time of TDEM is linear in the total number of generated



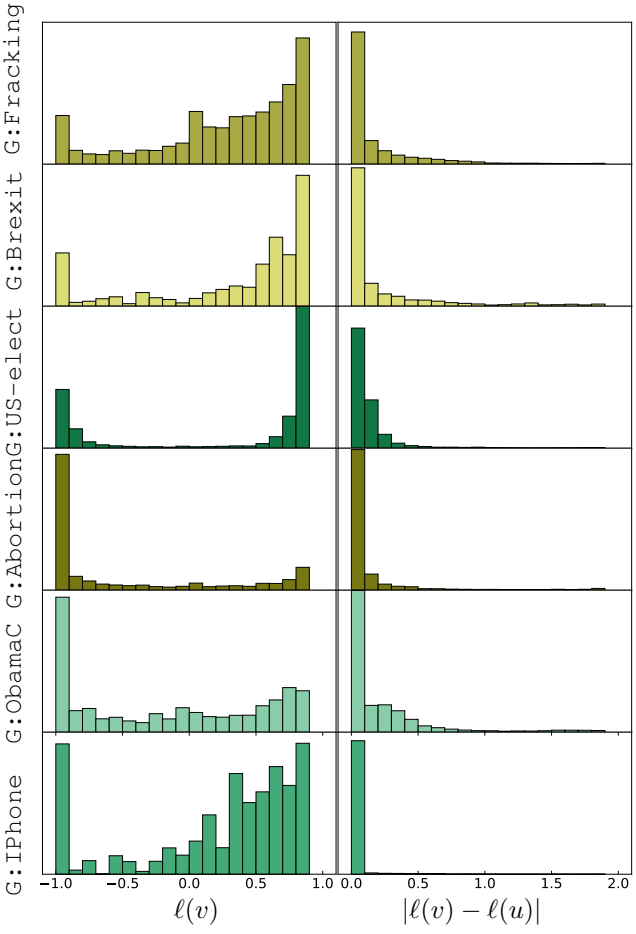


Fig. 1. Histograms of node leanings (left) and leaning differences across the edges (right) of DBLP, TPair and G networks.

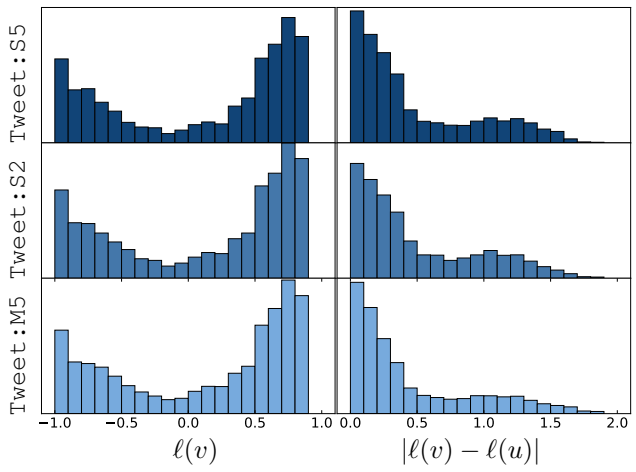


Fig. 2. Histograms of node leanings (left) and leaning differences across the edges (right) of Tweet networks.

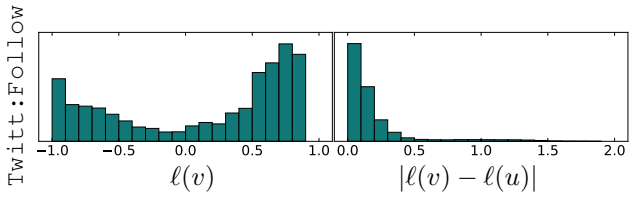


Fig. 3. Histograms of node leanings (left) and leaning differences across the edges (right) of Twitt networks.

RC-sets, which is very efficient. In order to not give it an unfair advantage against the comparison baselines, we store the RC-sets computed during the RC-set generation step of TDEM, and use them to also compute the baselines.[6]

The first baseline, MIN-VAR, selects at each iteration the highest-degree node v and greedily assigns to it items so as to minimize the variance among the leanings. Once k_v items have been assigned to v, MIN-VAR repeats the same process for the next highest degree node, until a total of k assignments are obtained. The second baseline, MAX-VAR, proceeds almost identically to MIN-VAR, but maximizes the variance instead of minimizing it. The third baseline, MYOPIC, selects the (next) highest degree node v at each iteration, like the other two baselines, but greedily assigns to it a set $A_v \in H$ of k_v items so as to maximize $f_v(A_v)$. We also considered a simple baseline that uses fully random assignments. However it performed very poorly, obtaining

6. Our implementation is publicly available: https://github.com/aslayci/TDEM_extension

exposure scores several orders of magnitude smaller than TDEM, so we decided to leave it out.

6.3 Results

Table 2 shows the diversity exposure scores achieved by the three baselines and by our algorithm, TDEM. For easier comparison, we report the average diversity exposure score of the individuals of the social network in each dataset. Recall that the smallest possible value is 0 and the maximum possible value is 1. Additionally, we report TDEM's memory consumption (in megabytes) and runtime (in seconds). The main computational bottleneck comes from the RC-generation step, which is also used by the baselines. Therefore, we do not report their memory consumption and runtime, since it differs only by a negligible amount to that of RC-Greedy, as the rest of the computations performed by the baselines are trivial. In summary, TDEM clearly outperforms the simple baselines in terms of the diversity exposure scores obtained. TDEM is able to identify non-trivial assignments that yield optimized diversity exposure in the network. That is, it finds a balance between exposure to diverse opinions yet selects items and nodes that do not have overly extreme leanings so as to not hinder propagation.

Observe that the runtime does not grow in proportion to the size of the network. Instead, it depends on the ability of items to propagate through the network, which depends, in turn, on the particular network structure, distribution of leanings, and propagation probabilities. Indeed, according to Theorem 4, the more limited the propagation of items, the more samples are needed to ensure adequate estimation of the spread. Thanks to the use of reverse exposure sets, we obtain a highly efficient algorithm, especially considering that we are dealing with h different influence spread problems, one for each item.

7 Conclusions

In this paper we present the first work tackling the problem of maximizing the diversity of exposure in an item-aware information propagation setting, taking a step towards breaking filter bubbles. Our problem formulation models many aspects of real-life social networks, resulting in a realistic model and a challenging computational problem. Despite the inherent difficulty of the problem, we are able to devise an algorithm that comes with an approximation guarantee, and is very scalable thanks to a novel extension of *reverse-reachable sets*. Through experiments on real-world datasets, we show that our method performs well and scales to large datasets.

Our work opens avenues for future work. One interesting problem is to improve the approximation guarantee of our algorithm by investigating further properties of the matroid formulation. Second, it would be interesting to experiment with different diversity functions, as well as to extend our approach to more complex propagation models such as, in particular, temporal variants of the independent-cascade model, with transmission probabilities that change over time.

References

[1] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*. Penguin UK, 2011.

[2] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2003, pp. 137–146.

[3] C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time." in *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2014, pp. 946–957.

[4] C. Aslay, A. Matakos, E. Galbrun, and A. Gionis, "Maximizing the diversity of exposure in a social network," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 863–868.

[5] E. Bakshy, S. Messing, and L. Adamic, "Exposure to ideologically diverse news and opinion on facebook," *Science*, 2015.

[6] R. K. Garrett, "Echo chambers online?: Politically motivated selective exposure among internet news users," *Journal of Computer-Mediated Communication*, vol. 14, no. 2, pp. 265–285, 2009.

[7] A. Matakos, E. Terzi, and P. Tsaparas, "Measuring and moderating opinion polarization in social networks," *Data Mining and Knowledge Discovery*, vol. 31, no. 5, pp. 1480–1505, 2017.

[8] L. Akoglu, "Quantifying political polarity based on bipartite opinion networks," 2014.

[9] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," 2011.

[10] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy in social media," in *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM)*, 2016.

[11] P. H. C. Guerra, W. M. Jr., C. Cardie, and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*, 2013.

[12] X. Chen, J. Lijffijt, and T. De Bie, "Quantifying and minimizing risk of conflict in social networks," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018, pp. 1197–1205.

[13] Q. V. Liao and W.-T. Fu, "Can you hear me now?: Mitigating the echo chamber effect by source position indicators," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 2014, pp. 184–196.

[14] C. Musco, C. Musco, and C. E. Tsourakakis, "Minimizing Polarization and Disagreement in Social Networks," *ArXiv*, 2017, 1712.09948.

[15] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Reducing controversy by connecting opposing views," in *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

[16] K. Garimella, A. Gionis, N. Parotsidis, and N. Tatti, "Balancing information exposure in social networks," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[17] K. Rawal and A. Khan, "Maximizing contrasting opinions in signed social networks," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1203–1210.

[18] S. Bharathi, D. Kempe, and M. Salek, "Competitive influence maximization in social networks," in *Proceedings of the 3rd International Conference on Internet and Network Economics (WINE)*, 2007, pp. 306–311.

[19] A. Borodin, Y. Filmus, and J. Oren, "Threshold models for competitive influence in social networks," in *Proceedings of the 6th International Conference on Internet and Network Economics (WINE)*, 2010, pp. 539–550.

[20] I. Valera and M. Gomez-Rodriguez, "Modeling adoption and usage of competing products," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2015, pp. 409–418.

[21] A. Borodin, M. Braverman, B. Lucier, and J. Oren, "Strategyproof mechanisms for competitive influence in networks," *Algorithmica*, vol. 78, no. 2, pp. 425–452, 2017.

[22] C. Aslay, W. Lu, F. Bonchi, A. Goyal, and L. V. S. Lakshmanan, "Viral marketing meets social advertising: Ad allocation with minimum regret," *Proceedings of the VLDB Endowment*, vol. 8, no. 7, pp. 814–825, 2015.

[23] C. Aslay, F. Bonchi, L. V. S. Lakshmanan, and W. Lu, "Revenue maximization in incentivized social advertising," *Proceedings of the VLDB Endowment*, vol. 10, no. 11, pp. 1238–1249, 2017.

[24] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation: Scaling up with guarantees," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM)*, 2014, pp. 629–638.

[25] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, 2016.

TABLE 2
Results summary: Diversity exposure scores.

Dataset (k, k_u)	$F(A)$ MYOPIC	MAX-VAR	MIN-VAR	TDEM	Mem. (MB)	RT (s)
DBLP:BSch $(5,1)$	0.034	0.014	0.042	0.050	457	2.16
DBLP:CPap $(5,1)$	0.077	0.019	0.070	0.111	276	1.8
DBLP:PYu $(5,1)$	0.098	0.018	0.129	0.167	285	2.68
TPair:X $(5,1)$	0.089	0.026	0.071	0.129	279	1.97
TPair:Y $(5,1)$	0.175	0.046	0.156	0.194	174	2.34
TPair:Z $(5,1)$	0.449	0.327	0.351	0.433	1 658	42.92
Tweet:S5 $(50,5)$	0.019	0.011	0.023	0.030	5 943	24.12
Tweet:S2 $(50,5)$	0.087	0.021	0.152	0.177	656	9.41
Tweet:M5 $(50,5)$	0.187	0.052	0.256	0.334	3 100	77.47
Twitt:Follow $(50,5)$	0.202	0.199	0.093	0.323	373	44.07
G:Brexit $(50,5)$	0.001	0.001	0.001	0.003	23 725	72.49
G:IPhone $(50,5)$	0.025	0.007	0.016	0.045	1 803	15.49
G:US-elect $(50,5)$	0.001	0.003	0.004	0.009	45 828	525.21
G:Abortion $(50,5)$	0.001	0.000	0.000	0.001	154 588	1 275.25
G:Fracking $(50,5)$	0.000	0.000	0.000	0.001	400 565	4 785.12
G:ObamaC $(50,5)$	0.000	0.000	0.000	0.001	360 449	3 936.16
Twitt:XL $(50,5)$	0.051	0.047	0.034	0.122	3 438	806.05

[26] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, 2015, pp. 1539–1554.

[27] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, 2014.

[28] S. Lin, Q. Hu, F. Wang, and P. S. Yu, "Steering information diffusion dynamically against user attention limitation," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2014.

[29] M. Cygan, F. V. Fomin, L. Kowalik, D. Lokshtanov, D. Marx, M. Pilipczuk, M. Pilipczuk, and S. Saurabh, *Parameterized Algorithms*, 1st ed., 2015.

[30] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, "An analysis of approximations for maximizing submodular set functions," in *Polyhedral combinatorics*, 1978, pp. 73–87.

[31] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2010, pp. 1029–1038.

[32] S. M. Ross, *Applied Probability Models with Optimization Applications*. Courier Corporation, 1970.

[33] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.

[34] E. Galbrun, B. Golshan, A. Gionis, and E. Terzi, "Finding low-tension communities," in *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2017.

[35] P. Lahoti, K. Garimella, and A. Gionis, "Joint non-negative matrix factorization for learning ideological leaning on twitter," in *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*, 2018.

[36] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau, "Tweeting from left to right: Is online political communication more than an echo chamber?" *Psychological Science*, vol. 26, no. 10, pp. 1531–1542, 2015.

Antonis Matakos is a PhD student in the Data Mining Group of the Computer Science Department at Aalto University. He obtained his MSc degree from the University of Ioannina, Greece. His research interests belong broadly to the area of algorithmic data mining, with specific focus on social network analysis, graph theory and web mining. His PhD Thesis focuses on proposing social media models and algorithms for social good.

Cigdem Aslay is a postdoctoral researcher in the Data Mining Group of the Computer Science Department at Aalto University. She received her PhD from the University of Pompeu Fabra in December 2016. During her PhD, she was hosted by Yahoo! Research as a research intern in the Web Mining Group. In 2017, she was a postdoctoral researcher in the Algorithmic Data Analytics Lab at ISI Foundation. Her work focuses on algorithmic methods for graph mining and social network analysis, with an emphasis on social influence propagation in online social networks.

Esther Galbrun is a researcher at the School of Computing, University of Eastern Finland, working mostly on data mining methods. She is also a research scientist at Inria, France, currently on leave. She received her PhD in 2013 from the University of Helsinki. In 2018, she was a postdoctoral researcher in the Data Mining Group of the Computer Science Department at Aalto University, where the work for this paper was done.

Aristides Gionis is a WASP professor in KTH Royal Institute of Technology, an adjunct professor in Aalto University, and a research fellow in ISI Foundation. Previously he was a senior research scientist in Yahoo! Research. He received his PhD from Stanford University in 2003. He is currently serving as an associate editor in DMKD, TKDD, and TWEB. His research interests include data mining, web mining, and social-network analysis.