Balanced Influence Maximization in the Presence of Homophily

Md Sanzeed Anwar
MIT
sanzeed@mit.edu

Martin Saveski
MIT
msaveski@mit.edu

Deb Roy
MIT
dkroy@media.mit.edu

ABSTRACT

The goal of influence maximization is to select a set of seed users that will optimally diffuse information through a network. In this paper, we study how applying traditional influence maximization algorithms affects the balance between different audience categories (e.g., gender breakdown) who will eventually be exposed to a message. More specifically, we investigate how structural homophily (i.e., the tendency to connect to similar others) and influence diffusion homophily (i.e., the tendency to be influenced by similar others) affect the balance among the activated nodes. We find that even under mild levels of homophily, the balance among the exposed nodes is significantly worse than the balance among the overall population, resulting in a significant disadvantage for one group. To address this challenge, we propose an algorithm that jointly maximizes the influence and balance among nodes while still preserving the attractive theoretical guarantees of the traditional influence maximization algorithms. We run a series of experiments on multiple synthetic and four real-world datasets to demonstrate the effectiveness of the proposed algorithm in improving the balance between different categories of exposed nodes.

CCS CONCEPTS

• **Information systems → Social advertising**; **Social networks**; • **Human-centered computing → Social network analysis**.

KEYWORDS

Influence Maximization; Balance; Fairness; Homophily.

ACM Reference Format:
Md Sanzeed Anwar, Martin Saveski, and Deb Roy. 2021. Balanced Influence Maximization in the Presence of Homophily. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441787

1 INTRODUCTION

The goal of influence maximization is to optimize the diffusion of content in a network by choosing an initial set of users who serve as seeds for spreading the information through the network. As suggested by the nature of this problem, influence maximization can play a vital role in addressing practical issues such as viral marketing. For example, if a corporate entity posts a job advertisement on a social network and wants it to reach as many users as possible, it needs to promote the advertisement to an optimal set of users so that the information spreads as much as possible through the network. To determine the optimal set of users, influence maximization algorithms need to be applied.

While these traditional algorithms are effective in amplifying the outreach of information pushed into the network, they solely focus on the number of users reached and do not consider specific properties that drive the connectivity and the person-to-person spread of information in the network. One such property is *homophily*, the phenomenon that a user is more likely to connect to or be influenced by other like-minded users. As a result, the existing discrepancies between various categories of users in a social network can be amplified by these algorithms, often putting certain groups of users at a significant disadvantage. For example, in the job advertisement case, by failing to account for the homophily in the structure of the network and the diffusion of information, the advertisement may reach a significantly smaller fraction of female users, hurting their interest.

The area of influence maximization has been very active over the last two decades. Previous work has investigated many different aspects of influence maximization, including how to scale the traditional influence maximization algorithms [23, 34, 36], how different spreading processes interact [9, 26, 29, 35], and how to seed in the presence of competing campaigns [8, 16, 17], to name a few. However, less research has been done on incorporating well-established social phenomena, such as homophily, in the influence models that largely determine the output of the influence maximization algorithms. Also, while there has been a great interest in how the predictions on machine learning algorithms affect different groups [5, 15, 18], there has been less focus on how influence maximization algorithms affect different audience segments and how to address any disparities.

In this paper, we aim to close this gap in the literature by making the following contributions[1]:

- We set up a simulation framework that allows us to systematically investigate the impact of structural and diffusion homophily on the categorical balance of the nodes reached by seeds selected using influence maximization algorithms (Section 3).
- We develop an influence maximization algorithm that jointly maximizes the spread of information and achieving categorical balance, and we demonstrate its effectiveness using both simulations and four real-world datasets (Section 4).

In the rest of this paper, we gradually demonstrate the impact of homophily on the categorical balance among the users who are eventually exposed to the information. First, we show that balance is not an issue in the absence of homophily (Section 3.1). Then, we

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441787

[1] The code and the data needed to reproduce our results can be found at the following URL: https://github.com/sanzeed/balanced_influence_maximization

test how homophily in network structure (Section 3.2), and later, homophily in both network structure and influence impact balance (Section 3.3), finding that it can have significant negative effects. Finally, we propose an algorithm that simultaneously maximizes influence and balance, and demonstrate that it performs better than existing approaches (Section 4).

2 BACKGROUND

We start by reviewing key notions in the influence maximization literature that we build upon in the rest of the paper. We highlight recent advances most closely related to our work in Section 5.

Influence Maximization. Traditionally, the problem of maximizing influence has been tackled in the context of marketing. Early data mining techniques aimed only at the intrinsic value of the customers, i.e., only considered the individual gain from a customer. Domingos and Richardson [14] were the first to account for the network value of the customers, i.e., the additional value of the customer's influence on other people.

Kempe et al. [21] formulated this question as a standard optimization problem, expressing it as: $\max_{S \subseteq V} f(S)$ s.t. $|S| = k$, for some parameter k, where $f(S)$ denotes the size of the active set for seed set S. Given the network of connections among the users, the goal is to find a set S of k seeds that results in the largest influence set, $f(S)$. In our experiments, we use a more optimized version of the algorithm by Kempe et al. named CELF [24] in order to reduce the running times.

Network Generation Models. One of the key properties of social networks is that they are scale-free, i.e., their degree distribution follows a power law: $p_k \sim k^{-\gamma}$, where p_k is the probability that a randomly sampled node has a degree k, and γ is a constant. Intuitively, this property implies that there will be fewer nodes with a higher degree.

Barabási and Albert [4] propose a model that generates undirected, scale-free networks, where the probability of two users connecting is directly proportional to their degrees. Bollobás et al. [6] extend this model to directed networks, where both the in and out-degree distributions follow a power law. Of interest to us are also models that incorporate homophily into the network generation process. Almeida et al. [12] propose a model for homophilic, scale-free networks but focus only on undirected networks.

In this work, we are interested in the impact of homophily on influence maximization in directed networks and build on ideas from [6] and [12] to generate homophilic directed scale-free networks and systematically vary the level of homophily in the network.

Influence Diffusion Models. In addition to specifying how the nodes (users) in a network are connected, we also need to specify how influence spreads from one node to another. Each node has two possible states: active and inactive. According to the *Independent Cascade* model, given an initial set of active nodes, the diffusion proceeds discretely: at each step, an active node u gets a single opportunity to activate its neighbor v with probability $p_{u,v}$ [21]. As the name suggests, the probability that an edge is activated is independent of whether any other edge is activated.

Homophily. In traditional network generation and influence models, users' probability to connect to or be influenced by another user only depends on the number of their connections. However, in the real world, people associate with and trust others very selectively. One of the fundamental properties that sets social networks apart from other networks is homophily, i.e., the tendency of "like to associate with like" [11, 22, 28, 32]. There are two main consequences of homophily that are relevant in the context of influence maximization: (i) that people who share the same attributes are more likely to be connected, and (ii) that, when influenced by their connections, people are more likely to be influenced by others with the same attributes [10, 13, 27].

Balance. In the context of influence maximization, the concept of balance may have multiple meanings. For example, we can define balance as the categorical balance among nodes in either the seed set or the active set (i.e., the nodes that are eventually reached). In this paper, we focus on the categorical balance in the active set, as it directly represents the effectiveness of an algorithm in reaching users from different categories fairly. In particular, we define the ideal categorical balance as the case in which the categorical ratio between the nodes in the network is preserved in the active set.

To simplify the exposition of the results, in the rest of the paper, we only consider the two-category case, i.e., we assume that any given node is either a majority node or a minority node. However, the simulations and the algorithm we propose in Section 4 can be easily extended to attributes or sets of attributes with multiple categories.

3 INFLUENCE MAXIMIZATION WITH TRADITIONAL ALGORITHMS

We start by investigating the impact of homophily on balance when applying traditional influence maximization algorithms. We propose a network generation model for homophilic directed scale-free networks and a homophilic influence diffusion model, which allow us to systematically vary the levels of structural and diffusion homophily. We measure the balance under three scenarios: (i) no homophily (ii) structural homophily, and (iii) structural and diffusion homophily. In each scenario, we:

1. Specify a pair of network generation model and influence diffusion model,
2. Generate a set of networks using a specific network generation model,
3. Assign each node to either majority or minority,
4. Run the traditional influence maximization algorithm [21],
5. Analyze the difference between the expected and the observed majority in the active set.

We consider three parameters in generating the networks: (i) n, the number of nodes; (ii) p_M, the fraction of nodes in the majority group; and (iii) h, the structural homophily index, i.e., the likelihood of a connection within vs. across groups [19].

We also consider two parameters in specifying the influence diffusion model: (i) b_p, the base probability of a node successfully influencing a neighbor node; and (ii) h_p, the diffusion homophily index, i.e., the likelihood of influence within vs. across groups. Finally, we use k to denote the number of seeds.

Recall that we denote by $f(S)$ the size of the active set for seed set S. We also define $M(S)$ and $m(S)$ to be the majority, and the minority in the active set for seed set S; then, $f(S) = M(S) + m(S)$. To measure the balance achieved by the algorithm, we consider the parameter

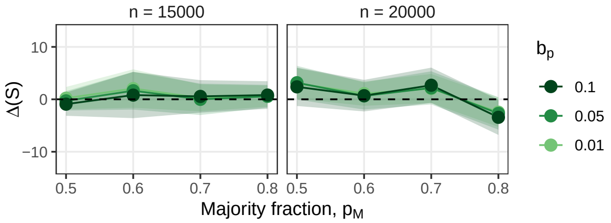

Figure 1: Seeding of non-homophilic networks using the traditional influence maximization algorithm. Mean $\Delta(S)$ (difference in size between the observed and the target majorities in the active set) for different values of p_M (fraction of nodes in the majority group) and b_p (base influence probability of the diffusion model). In the absence of homophily, the traditional influence maximization algorithm naturally achieves a categorical balance among the nodes in the active set. The error bands represent 95% CIs.

p_t, the fraction of majority we want in the active set of nodes, and we compare two quantities: (i) the observed size of the majority in the active set, given by $M(S)$, and (ii) the target size of the majority in the active set, given by $p_t f(S)$. According to our definition of ideal categorical balance, we want to preserve the majority vs. minority ratio in the active set, i.e., we want $p_t = p_M$. We achieve ideal categorical balance when the two quantities we compare are equal, i.e., when $\Delta(S) = M(S) - p_t f(S) = M(S) - p_M f(S) = 0$.

3.1 No Homophily

We start by analyzing the balance of the activated nodes in the absence of homophily.

Network Generation Model. We generate a series of directed scale-free networks using the model proposed by Bollobás et al. [6].

Let $G(t)$ denote the network at time t with exactly t edges and $n(t)$ nodes. We start with an initial network $G(t_0) = G_0$ at time t_0, and non-negative real numbers $\alpha, \beta, \gamma, \delta_{in}, \delta_{out}$ s.t. $\alpha + \beta + \gamma = 1$. For any node u, we denote its in-degree and out-degree by $d_{in}(u)$ and $d_{out}(u)$, respectively. For $t \geq t_0$, we build $G(t+1)$ from $G(t)$ by adding an edge to $G(t)$ at timestep $t+1$ as follows:

- With probability α, we add a new node v and an edge from v to an existing node w. We choose w from all existing nodes with probability proportional to $d_{in}(w) + \delta_{in}$,
- With probability β, we add an edge from an existing node v to an existing node w. We choose v from all existing nodes with probability proportional to $d_{out}(v) + \delta_{out}$, and we choose w from all existing nodes with probability proportional to $d_{in}(w) + \delta_{in}$,
- With probability γ, we add a new node w and an edge to w from an existing node v. We choose v from all existing nodes with probability proportional to $d_{out}(v) + \delta_{out}$.

In our experiments, we set $\alpha = \beta = \gamma = \frac{1}{3}$ in order to make the three edge addition scenarios equally likely. We also set $\delta_{in} = \delta_{out} = 1$ in order to prevent zero division while calculating the probability distribution. In order to categorize the nodes of the network into two distinct categories, we assign each node in the network to the majority category with probability p_M.

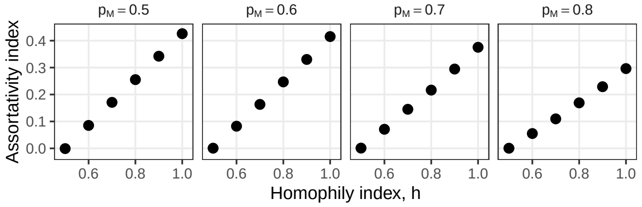

Figure 2: Assortativity of the networks generated using the homophilic network generation model introduced in Section 3.2 for different values of $h \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, the structural homophily index (a model parameter), and p_M, the fraction of nodes in the majority group. The assortativity of the networks has a clear linear relation to the homophily index h, demonstrating that the networks generated by the model are indeed homophilic.

Influence Diffusion Model. We use the traditional (non-homophilic) Independent Cascade model where the probability of any node successfully influencing a neighbor node is a constant, i.e., the base probability, b_p.

Setup. We vary the number of nodes in the network: $n \in \{15k, 20k\}$, the fraction of nodes in the majority group: $p_M \in \{0.5, 0.6, 0.7, 0.8\}$, and the base influence probability: $b_p \in \{0.01, 0.05, 0.1\}$. Since we are interested in the outcomes in the absence of any kind of homophily, we do not need to consider the structural homophily index (h) and the diffusion homophily index (h_p). For each pair (n, p_M), we generated 20 networks and ran an optimized version [24] of the influence maximization algorithm by Kempe et al. [21] to choose $k = 200$ seeds.

Results and Observations. Figure 1 shows the difference between the observed and the target majorities ($\Delta(S)$) as a function of the fraction of majority nodes in the network (p_M). We find that, for all values of p_M, $\Delta(S)$ is close to zero. This suggests that the traditional influence maximization algorithm naturally selects seed nodes that reach a balanced set of nodes when neither the network nor the influence diffusion is homophilic.

3.2 Network Homophily

Next, we analyze the balance of the activated nodes in the presence of network homophily, i.e., when similar nodes are more likely to be connected.

Network Generation Model. We build upon the network generation models by Bollobás et al. [6] and Karimi et al. [19] to generate homophilic, scale-free networks. Using the same notation as before, for any pair of nodes v and w, we define:

$$h(v,w) = \begin{cases} h & \text{if } v \text{ and } w \text{ are from the same category,} \\ 1-h & \text{otherwise.} \end{cases}$$

We add a single directed edge in each timestep. We build $G(t+1)$ from $G(t)$ by adding an edge to $G(t)$ at timestep $t+1$ as follows:

- With probability α, we add a new node v and an edge from v to an existing node w. We assign v to the majority category with probability p_M, and to the minority category otherwise. We choose w from all existing nodes with probability proportional to $h(v,w)d_{in}(w) + \delta_{in}$,

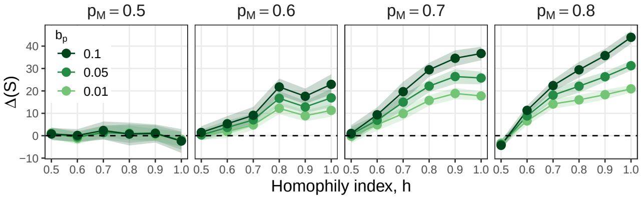

Figure 3: Seeding of homophilic networks using the traditional influence maximization algorithm. Mean $\Delta(S)$ (difference in size between the observed and the target majorities in the active set) for different values of p_M (the fraction on nodes in the majority group) and b_p (the base influence probability). $\Delta(S)$ rises as we approach stronger homophily ($h > 0.5$). The more extreme p_M, the steeper the rise.

- With probability β, we add an edge from an existing node v to an existing node w. We choose v from all existing nodes with probability proportional to $d_{out}(v) + \delta_{out}$, and we choose w from all existing nodes with probability proportional to $h(v, w)d_{in}(w) + \delta_{in}$,
- With probability γ, we add a new node w and an edge to w from an existing node v. We assign w to the majority category with probability p_M, and to the minority category otherwise. We choose v from all existing nodes with probability proportional to $h(v, w)d_{out}(v) + \delta_{out}$.

Similar to the previous section, we set $\alpha = \beta = \gamma = \frac{1}{3}$ in order to make the three edge addition scenarios equally likely, as well as $\delta_{in} = \delta_{out} = 1$ in order to prevent zero division while calculating the probability distribution.

This model is particularly appealing as it generates networks that resemble real-world social networks, i.e., have a small diameter, power-law in- and out-degree distributions, and allows us to vary the level of homophily in the network by changing the model parameters.

To verify that the networks generated using this model have the expected levels of homophily, we compute the assortativity index [31] (a well-established measure of network homophily) of networks generated using different values of h. We find that indeed networks generated with larger values of the structural homophily index, h, have higher assortativity (Figure 2), demonstrating that the model achieves the desired effect.

Influence Diffusion Model. We use the same simple, non-homophilic diffusion model for influence described in the previous section, where the probability of any node successfully influencing a neighbor node is b_p.

Setup. We fix the number of nodes in the network to $n = 20k$ and vary the faction of nodes in the majority group: $p_M \in \{0.5, 0.6, 0.7, 0.8\}$, the structural homophily index: $h \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, and the base influence probability: $b_p \in \{0.01, 0.05, 0.1\}$. Since the diffusion model is non-homophilic, we did not need to consider the diffusion homophily index, h_p. As before, for each set of values of the parameters, we generate 20 networks and choose $k = 200$ seeds using an optimized version [24] of the influence maximization algorithm by Kempe et al. [21].

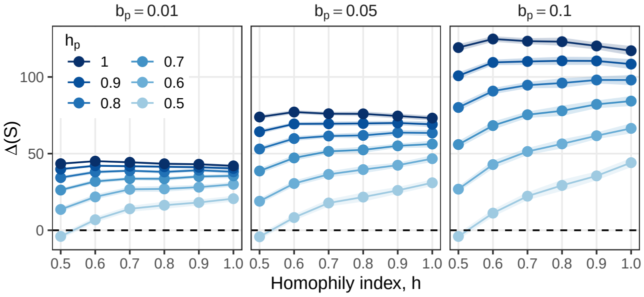

Figure 4: Seeding of homophilic networks under homophilic influence using the traditional influence maximization algorithm. Mean $\Delta(S)$ (difference in size between the observed and the target majorities in the active set) for different values of h (the structural homophily index), h_p (the diffusion homophily index), and b_p, (the base influence probability). $\Delta(S)$ is much higher when both the network and the diffusion model are homophilic.

Results and Observations. Figure 3 shows the difference between the observed and the target majorities ($\Delta(S)$) as a function of the structural homophily index (h). We find that when there is an equal number of nodes in the two groups ($p_M = 0.5$), the traditional influence maximization algorithm naturally achieves balance, regardless of the homophily level in the network (h). However, as soon as there are more nodes in one group (i.e., $p_M > 0.5$) and some homophily in the network formation (i.e., $h > 0.5$), the algorithm starts to favor the majority group, selecting seeds that reach more nodes in the majority group. This suggests that even when just the network structure becomes mildly homophilic, the traditional influence maximization algorithm fails to achieve balance.

3.3 Network and Diffusion Homophily

Finally, we analyze the balance of the activated nodes in the presence of network and diffusion homophily, i.e., when similar nodes are more likely to both connect and influence each other.

Network Generation Model. We use the same model as in the previous section to generate scale-free, homophilic networks.

Influence Diffusion Model. We now add homophily in the influence diffusion. First, we define:

$$h_p(v, w) = \begin{cases} h_p & \text{if } v \text{ and } w \text{ are from the same category,} \\ 1 - h_p & \text{otherwise.} \end{cases}$$

Using this definition, for a given base probability b_p, we define our homophilic influence diffusion model such that the edge probability assigned to an edge (v, w) is proportional to $h_p(v, w)$. The motivation behind this choice is to make the results comparable to the experiments in which we used non-homophilic diffusion.

Experiments. We fix the number of nodes in the network to $n = 20k$ and the fraction of nodes in the majority group to $p_M = 0.8$. We vary the network homophily by setting different values of the structural homophily index, $h \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and the diffusion homophily by using different values of the diffusion homophily index: $h_p \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. We also vary the base influence probability of our influence diffusion model:

$b_p \in \{0.01, 0.05, 0.1\}$. We average the results over 20 networks, in each case choosing $k = 200$ seeds using an optimized version [24] of the influence maximization algorithm by Kempe et al. [21].

Results and Observations. Figure 4 shows the difference between the observed and the target majorities ($\Delta(S)$) for different values of the structural (h) and diffusion (h_p) homophily indices. In all cases, we observe that the difference between the observed and the target majorities increases as the homophily index h increases from 0.5 to 1.0, and the network becomes more and more homophilic. The difference further increases for the same values of the structural homophily (h) as we increase the diffusion homophily (h_p) from 0.5 to 1.0. Finally, the larger the probability for any single edge to be activated (b_p), the larger the difference between the observed and the target majorities ($\Delta(S)$) for the corresponding values of h and h_p. These observations suggest that the higher is the structural and the diffusion homophily, the more severe is the imbalance of the nodes reached by the seeds selected using the traditional influence maximization algorithm.

4 BALANCED INFLUENCE MAXIMIZATION

4.1 Algorithm

So far, we have seen that even under mild homophily, the traditional influence maximization algorithm tends to select seeds that reach an imbalanced set of active nodes, leading to a systematic disadvantage for one group. To address this issue, in this section, we propose a new algorithm for selecting seed nodes that jointly maximizes the number and the balance of the users reached by the seed nodes.

We start by revisiting some basic terminology. Given a network $G(V, E)$ and seed set S, recall that we define $f(S)$ to be the size of the active set that results from seeding the nodes in S. We also define $M(S)$ and $m(S)$ to be the majority and the minority in the active set, respectively. Then, $f(S) = M(S) + m(S)$. To simplify the exposition, we focus on the case where the nodes belong to one of two categories, but the algorithm can be readily extended to cases where the nodes belong to more than two categories.

We aim to use a greedy hill-climbing approach inspired by Kempe et al. [21] using the following objective function:

$$F(S) = \underbrace{(1-\lambda)\cdot\frac{f(S)}{|V|}}_{\text{Influence}} + \underbrace{\lambda\cdot\frac{\sqrt{\gamma M(S)}+\sqrt{(1-\gamma)m(S)}}{\sqrt{\gamma p_M n}+\sqrt{(1-\gamma)(1-p_M)n}}}_{\text{Balance}} \tag{1}$$

The first component of $F(S)$ maximizes $f(S)$, the size of the active set, and the second component maximizes the categorical balance in the active set. The denominators in both components are normalizing constants that make the semantics of the hyperparameter λ consistent across different networks. The numerator of the balance component includes two terms, $\sqrt{\gamma M(S)}$ and $\sqrt{(1-\gamma)p_M m(S)}$. The intuition behind this balance component is that once many majority nodes are added to the active set, adding more majority nodes will lead to diminishing gains, thanks to the square root function (i.e., $\sqrt{\gamma M(S)}$), and the algorithm will favor minority nodes. The hyperparameter γ controls how much the balance component favors the majority vs. the minority group, and the hyperparameter λ controls the trade-off between balance and influence. For $\lambda = 0$ the objective function is the same as the algorithm by Kempe et al. [21].

```
Algorithm 1: Balanced Influence Maximization
  Input: $G(V, E),\ k,\ \lambda$, Diffusion model
  Output: $S$
  $S \leftarrow \emptyset$
  while $|S| < k$ do
      $u \leftarrow \arg\max_{v \in V} \Big\{F(S \cup \{v\}) - F(S)\Big\}$
      $S \leftarrow S \cup \{u\}$
  return $S$
```

Using the terminology defined above, we propose Algorithm 1 to achieve categorical balance in influence maximization. Next, we prove that Algorithm 1 provides a $(1 - \frac{1}{e})$-guarantee in approximating $\max_{S \subseteq V} F(S)$. To do so, we first prove the following theorems.

Theorem 4.1. *(Non-negativity) Given a network $G(V, E)$ and a set of realizations $\mathcal{R}$, for any $S \subseteq V$, $F(S) \geq 0$.*

Proof. This result follows from the non-negativity of $f(S)$, $M(S)$ and $m(S)$. □

Theorem 4.2. *(Monotonicity) Given a network $G(V, E)$, F is monotone, i.e., for any $S \subseteq T \subseteq V$, $F(S) \leq F(T)$.*

Proof. Note that seeding additional nodes cannot decrease the size of the active set. Therefore, f is monotone. Similarly, seeding additional nodes cannot decrease the size of the majority or the minority in the set of active nodes. Therefore, $M(S)$ and $m(S)$ are monotone as well. Then, from the definition, F is monotone. □

Theorem 4.3. *(Submodularity) Given a network $G(V, E)$, F is submodular, i.e., for any $S \subseteq T \subseteq V$ and any $u \in V$, $F(S \cup \{u\}) - F(S) \geq F(T \cup \{u\}) - F(T)$.*

Proof. Abusing notation, let $f(S)$ also denote the active set for seed set S. Consider an arbitrary $v \in V$ such that $v \in f(T \cup \{u\}) - f(T)$. Then $v \in f(T \cup \{u\})$ but $v \notin f(T)$, i.e., v is activated by u and not by the nodes in T. Since $S \subseteq T$, v cannot be activated by the nodes in S either. So, $v \notin f(S)$. However, v is activated by u, and so $v \in f(S \cup \{u\})$. Then, $v \in f(S \cup \{u\}) - f(S)$, and therefore, $f(S \cup \{u\}) - f(S) \supseteq f(T \cup \{u\}) - f(T)$. In other words, $f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$, and f is submodular.

Following the same logic, we can prove that M and m are submodular. Then, we can use the following theorem [25] to show that $\sqrt{M}$ and $\sqrt{m}$ are also submodular:

Theorem 4.4. *Given functions $h : 2^V \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, the composition $H = g \circ h : 2^V \to \mathbb{R}$ (i.e., $H(S) = g(h(S))$) is non-decreasing submodular, if g is non-decreasing concave and h is non-decreasing submodular.*

Finally, since F is a non-negative linear combination of f, $\sqrt{M}$ and $\sqrt{m}$, F must be submodular as well, as desired. □

Since F is non-negative, monotone and submodular, the following result by Nemhauser, Wolsey, and Fisher [30] applies to F:

Theorem 4.5. *For a non-negative, monotone submodular function h, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let S^* be a set that maximizes the*

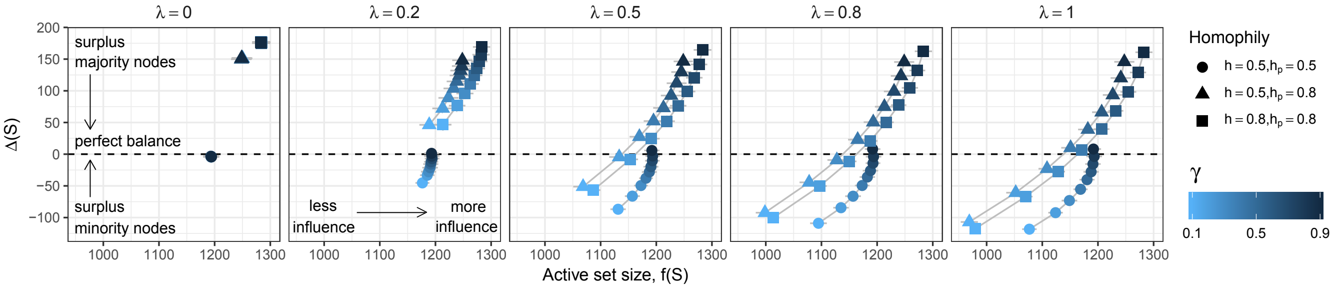

Figure 5: Seeding using balanced influence maximization (Algorithm 1) under structural and diffusion homophily. The effect of the hyperparameters λ (higher values give more importance on balance) and γ (higher values give more weight to the majority group) on the trade-off between influence (x axis) and balance (y axis) for different values of the structural (h) and diffusion (h_p) homophily indices. The points represent the means over ten runs, and the error bars represent 95% confidence intervals.

value of overall k-element sets. Then $h(S) \geq (1 - \frac{1}{e})h(S^*)$; in other words, S provides a $(1 - \frac{1}{e})$-approximation.

This theorem implies that our algorithm achieves a $(1 - \frac{1}{e})$-guarantee in approximating the maximum value of our objective.

Computing the objective. As influence diffusion is a random process, we cannot directly compute $F(S)$. Therefore, the greedy hill-climbing algorithm utilizes $\mathbb{E}[F(S)]$. To approximate $\mathbb{E}[F(S)]$, we define the following: a *realization* of $G(V, E)$ under a specified influence diffusion model is a subgraph $G'(V, E')$, such that for any $e \in E$, the probability that $e \in E'$ is the same as the diffusion probability assigned to e under the influence diffusion model. For a specific realization r of $G(V, E)$, let $f_r(S)$ be the value of $f(S)$ conditioned on r. We define $M_r(S)$ and $m_r(S)$ in a similar manner. Then, for a sufficiently large set $\mathcal{R}$ of realizations,

$$
\begin{aligned}
&\mathbb{E}[F(S)] \approx \mathbb{E}_{\mathcal{R}}[F(S)] \\
&= (1-\lambda) \cdot \frac{\mathbb{E}_{\mathcal{R}}[f(S)]}{|V|} + \lambda \cdot \frac{\mathbb{E}_{\mathcal{R}}\left[\sqrt{\gamma M(S)} + \sqrt{(1-\gamma)m(S)}\right]}{\sqrt{\gamma p_M n} + \sqrt{(1-\gamma)(1-p_M)n}} \\
&= (1-\lambda) \cdot \frac{\frac{1}{|\mathcal{R}|}\sum\limits_{r \in \mathcal{R}} f_r(S)}{|V|} + \lambda \cdot \frac{\frac{1}{|\mathcal{R}|}\sum\limits_{r \in \mathcal{R}} \sqrt{\gamma M_r(S)} + \sqrt{(1-\gamma)m_r(S)}}{\sqrt{\gamma p_M n} + \sqrt{(1-\gamma)(1-p_M)n}}.
\end{aligned}
$$

4.2 Simulation Experiments

Next, we use our simulation framework to test the performance of Algorithm 1 on synthetic networks.

Setup. Similar to our previous experiments, we generate a set of homophilic networks (using the model described in Section 3.2), assign each node to one of two categories (majority and minority), assume a homophilic diffusion model (as proposed in Section 3.3), and choose a set of seeds using Algorithm 1.

We fix the number of nodes in this network to $n = 20k$ and fraction of nodes in the majority to $p_M = 0.8$. We also fix the base influence probability to $b_p = 0.2$ and the number of realizations to $|\mathcal{R}| = 1k$. We experiment with three pairs of values for the structural homophily h and the diffusion homophily h_p: ($h = 0.5, h_p = 0.5$), ($h = 0.5, h_p = 0.8$), and ($h = 0.8, h_p = 0.8$)[2]. For each pair, we run Algorithm 1 to choose $k = 200$ seeds. We use different setting of the hyperparameters $\lambda \in \{0.0, 0.2, 0.5, 0.8, 1.0\}$ and $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and test their effect on the size and the balance of the active set. To account for the intrinsic randomness of the network generation and influence models, we repeat each experiment 10 times and report the means and confidence intervals (Figure 5).

Results and Observations. We find that our algorithm exhibits a similar pattern in all three homophily settings (h and h_p). When $\lambda = 0$, our algorithm focuses solely on maximizing influence, and consequently varying γ affects neither $f(S)$, the size of the active set, nor $\Delta(S)$, the difference between the observed and the target majorities. As we increase λ, we begin to observe the trade-off between $f(S)$ and $\Delta(S)$ for different γ; namely, increasing γ increases $f(S)$, but also increases $\Delta(S)$. When $\lambda = 1$, our algorithm focuses on balance, and this trade-off is at its maximum. We note that due to the nature of the balancing component of our objective function (Equation 1), even when $\lambda = 1$, the algorithm still implicitly aims to maximize the size of the active set of each group.

In the absence of both structural and diffusion homophily ($h = 0.5, h_p = 0.5$), we can achieve balance by simply focusing on influence, i.e., setting $\lambda = 0$, aligning with our observations in Section 3.1. In fact, increasing λ and using more extreme values of $\gamma \in \{0.1, 0.2, 0.3\}$ can hurt both the influence and the balance. However, in the presence of structural homophily ($h = 0.5$, $h_p = 0.8$) or both structural and diffusion homophily ($h = 0.8$, $h_p = 0.8$), focusing solely on maximizing influence yields poor balance, i.e., high $\Delta(S)$. In these scenarios, higher values of λ achieve better categorical balance in exchange for a decrease in influence.

Setting λ to a high value (e.g., $\lambda = 1$) and varying γ (the weight assigned to the majority group) allows us to sample a wide range of seeding choices that have different influence vs. balance trade-offs. This is especially important in cases where we are interested in adopting a more extreme definition of balance, e.g., an equal number of majority and minority nodes in the active set.

4.3 Experiments in Real-World Networks

Next, we test the balanced influence maximization algorithm on four real-world networks and compare its performance to the algorithm proposed by Stoica et al. [33].

[2] We also analyzed ($h = 0.8$, $h_p = 0.5$) and found a very similar pattern to ($h = 0.5$, $h_p = 0.8$). To avoid visual clutter in Figure 5, we exclude those results.

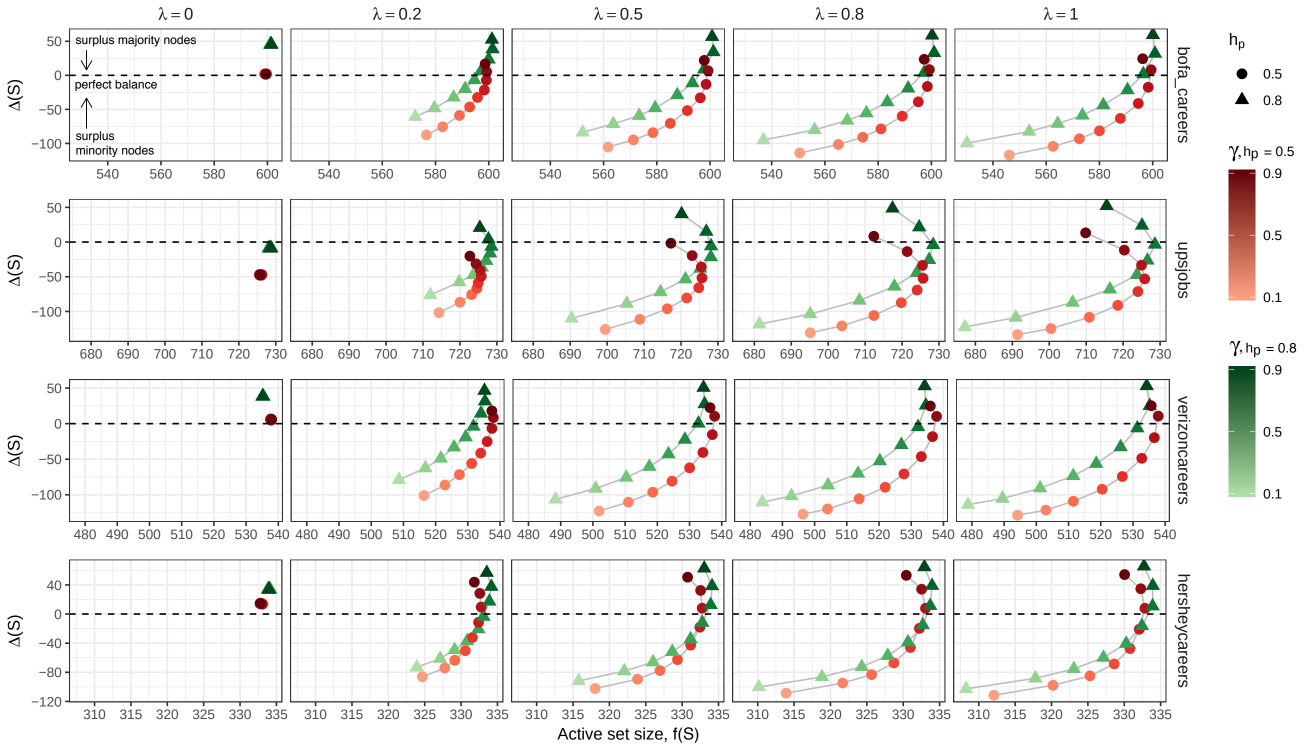

Figure 6: Results of seeding four real-world networks using Algorithm 1. The plots illustrate how the hyperparameters λ (higher values give more importance on balance) and γ (higher values give more weight to the majority group) affect the trade-off between influence (x axis) and balance (y axis), both in the absence ($h_p = 0.5$) and presence ($h_p = 0.8$) of diffusion homophily. The points represent the means over ten runs; the 95% confidence intervals are too small to be visible.

Setup. We consider the careers Twitter accounts of four major companies: Bank of America (@bofa_careers), UPS (@upsjobs), Verizon (@verizoncareers), and Hershey's (@hersheycareers), where they often post job announcements. Using the Twitter API, we fetch the followers of each account and the followers of the followers to construct the network of connections among the followers of each account. We also fetch the followers' names and use `genderize.io` to determine their gender[3]. `genderize.io` uses an extensive database of first and last names and their gender associations from many countries/languages and has been shown to have high accuracy [20]. The networks vary in size (n), the fraction of users in the majority group (p_M), and structural homophily (here we report assortativity, A, which maps very closely to the h parameter in our simulations, Figure 2):

@bofa_careers: $n = 13{,}688$, $p_M = 0.77$ (male), $A = 0.08$,
@upsjobs: $n = 13{,}851$, $p_M = 0.69$ (male), $A = 0.33$,
@verizoncareers: $n = 9{,}226$, $p_M = 0.77$ (male), $A = 0.04$,
@hersheycareers: $n = 3{,}726$, $p_M = 0.68$ (male), $A = 0.05$.

We choose $k = 200$ seeds using Algorithm 1, assuming the homophilic influence model with $b_p = 0.01$, varying the level of diffusion homophily, $h_p \in \{0.5, 0.8\}$ and measure the effect of varying the hyperparameters $\lambda \in \{0.0, 0.2, 0.5, 0.8, 1.0\}$ and $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. To account for the randomness in the diffusion simulations, we repeat each experiment 10 times and report the means and 95% confidence intervals.

Baseline. In addition to the traditional influence maximization algorithm, we also compare our algorithm's performance with the algorithm introduced by Stoica et al. [33]. They propose choosing seed nodes based on node degree, but instead of simply selecting the highest degree nodes from the full population, they select the highest degree nodes for each group individually. The main idea behind the algorithm is that imposing balance on the seed set will lead to balance in the active set. To make fair compassion with our algorithm, we add an additional parameter to their algorithm, k_{offset}, which allows us to vary the proportion of seeds in the majority group. In particular, we select as seeds $k_M = \lfloor p_M k \rfloor + k_{offset}$ majority nodes with highest degree and $k_m = k - k_M$ minority nodes with highest degree. We vary $k_{offset} \in \{-50, -40, -30, -20, -10, 0, 10, 20, 30, 40, 50\}$, repeat each experiment 10 times, and report the means and confidence intervals.

Results and Observations. First, we analyze our algorithm's behavior for different values of λ and γ (Figure 6). We find that in the absence of diffusion homophily ($h_p = 0.5$), the hyperparameters that achieve the best balance also achieve the greatest influence. This is perhaps because the networks have very low structural homophily. One exception is the @upsjobs network, which has a

[3] We tokenize each name, remove punctuation and tokens with less than three characters, and query for the gender of each token. We discard cases where none or an equal number of tokens are associated with a male or a female name, which includes accounts by organizations. We ignore private Twitter accounts.

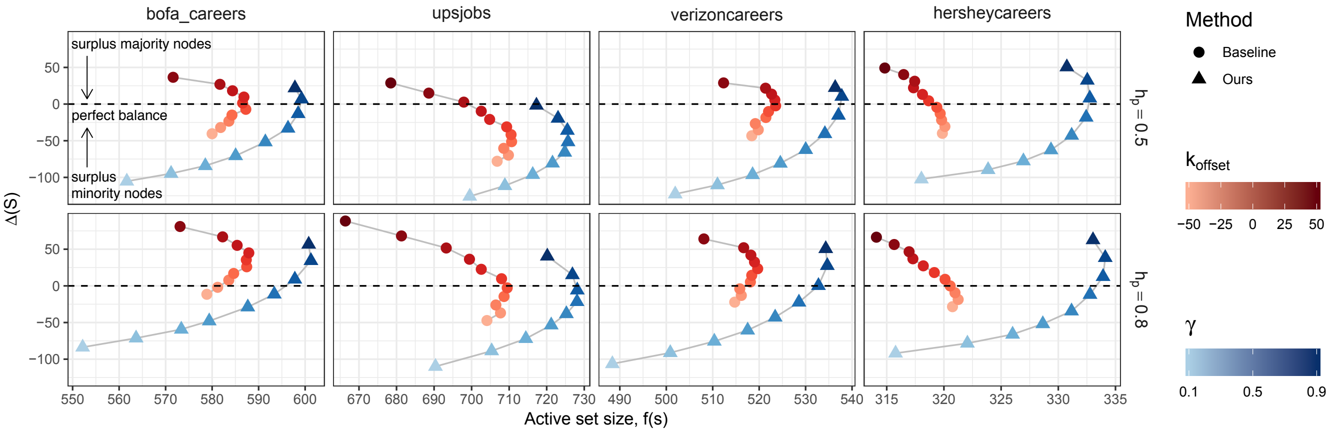

Figure 7: Comparison between the performance of the baseline algorithm and our algorithm with $\lambda = 0.5$ in seeding four real-world networks. The plots illustrate the trade-off between influence (x axis) and balance (y axis) achieved by the two algorithms, both in the absence ($h_p = 0.5$) and presence ($h_p = 0.8$) of diffusion homophily. Both algorithms achieve the desired balance with the right hyperparameter settings, but our algorithm consistently achieves higher influence. The points represent the means over ten runs; the 95% confidence intervals are too small to be visible.

higher level of homophily (assortativity A=0.33), and where we need to make a small sacrifice in the active set size to achieve the desired balance. In the presence of diffusion homophily ($h_p = 0.8$), the hyperparameter settings that achieve the desired balance do not lead to the largest influence, except for the @upsjobs network. However, in these cases, the algorithm successfully trades-off influence with balance, achieving the desired balance by only slightly reducing the size of the active set. For instance, in the @bofa_careers network, the algorithm is able to decrease the surplus in majority nodes from 52 to 6, achieving nearly perfect balance, while reducing the size of the active set by only 4 nodes (601 to 597).

Next, we compare the performance of our algorithm with the performance of the baseline algorithm (Figure 7). We consider only the results of our algorithm for $\lambda = 0.5$, assigning equal importance to influence and balance. We observe that both algorithms can achieve the desired balance in the active set with the right hyperparameter settings. However, we find that our algorithm is consistently able to achieve a higher influence, selecting seeds that reach a significantly larger number of nodes. This pattern holds across all four networks, both in the absence ($h_p = 0.5$) and the presence ($h_p = 0.8$) of diffusion homophily.

4.4 Hyperparameter Selection

The influence and the balance achieved by our algorithm depend on many factors, including the network topology, the proportion of majority vs. minority nodes, and the levels of structural and diffusion homophily. Therefore, we cannot expect any single hyperparameter setting to achieve balance on every network. However, based on our experiments, we recommend starting the exploration of the hyperparameter space by setting $\lambda = 0.5$ and $\gamma = 1 - p_t$, where p_t is the desired proportion of majority nodes in the active set. Setting $\lambda = 0.5$ gives equal importance to influence and balance, and setting $\gamma = 1 - p_t$ gives more importance to the minority group, counteracting the natural advantage of the majority group.

5 RELATED WORK

We build on existing work that has studied the performance of traditional seeding algorithms under various network models, analyzed the relationship between homophily and diffusion, and investigated the question of balance in influence maximization.

Aral and Dhilon [2] and Aral et al. [3] demonstrate the importance of using empirically motivated influence models. They show that traditional influence models that do not model any empirical properties of information diffusion, such as the independent cascade model, can significantly underestimate influence propagation. In this paper, we considered an influence model that takes into account the homophily among the users and studied how different levels of homophily affect the balance of the influenced users.

Several previous studies have modeled the relationship between homophily and information diffusion [27], measured the gains in accuracy of predicting diffusion when considering homophily [13], and tested the effects of homophily in the adoption of behaviors [10]. In this work, we demonstrate that homophily can lead to an imbalance among the influenced individuals when applying influence maximization algorithms and propose an algorithm that mitigates it.

Bredereck et al. [7] consider the problem of assembling a group of individuals that both score high on a certain quality measure and are diverse as a group. Their method can be adopted for influence maximization, where “quality” is defined as a certain measure of the individual’s position in the network, e.g., their page-rank. However, such an approach would not account for the fact that in influence maximization, the “quality” of a user changes depending on which other users are also selected to be in the group. For example, selecting two high page-rank nodes that influence the same users is suboptimal, although each of the two nodes is a good choice individually. That is precisely the issue that our algorithm addresses.

Stoica et al. [33] propose the algorithm we used as a baseline in Section 4.3, but also investigate the trade-off between balance and influence as a function of the seed set size. They find that when the size of the seed set is sufficiently large, imposing balance in the seed set also leads to more influence. The intuition is that after selecting the high degree majority nodes as seeds, promoting balance helps reach other parts of the network that are not connected to the high degree majority nodes.

In this paper, we focused on minimizing the disparities in information exposure of users belonging to different categories when applying influence maximization algorithms to a single campaign. Garimella et al. [16] and Yu et al. [37] consider another scenario where the goal is to reach a balanced set of users across several campaigns simultaneously running in the network.

Finally, similar to the approach we took in this paper, Ali et al. [1] consider group fairness in influence maximization. While we address the challenges of applying influence maximization algorithms in the presence of homophily, they focus on time-critical influence maximization, i.e., applications where it is only beneficial to influence users before a deadline.

6 CONCLUSION

In this paper, we studied how homophily in network formation and influence diffusion affects the categorical balance of the nodes reached by seeds selected to maximize influence. We found that applying traditional influence maximization algorithms leads to a significant imbalance in outreach even in the presence of mild network or diffusion homophily. To address this issue, we proposed a new influence maximization algorithm that jointly maximizes influence and balance, and has strong performance guarantees. Through experiments in synthetic and real-world networks, we show that it effectively trades-off between influence and balance, and outperforms existing algorithms for balanced influence maximization.

Our work opens new directions for future work, including how to measure and mitigate imbalance in terms of continuous node attributes and how to adopt recent advances in the design of influence maximization algorithms to improve the scalability of our algorithm.

REFERENCES

[1] Junaid Ali, Mahmoudreza Babaei, Abhijnan Chakraborty, Baharan Mirzasoleiman, Krishna Gummadi, and Adish Singla. 2019. On the fairness of time-critical influence maximization in social networks. In *Proceedings of the Human-Centric Machine Learning Workshop at NeurIPS*.

[2] Sinan Aral and Paramveer Dhillon. 2018. Social influence maximization under empirical influence models. *Nature Human Behaviour* (2018).

[3] Sinan Aral, Lev Muchnik, and Arun Sundararajan. 2013. Engineering social contagions: Optimal network seeding in the presence of homophily. *Network Science* (2013).

[4] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* (1999).

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NIPS Tutorial* (2017).

[6] Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. 2003. Directed scale-free graphs. In *Proceedings of the Annual ACM-SIAM symposium on Discrete algorithms*.

[7] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. 2018. Multiwinner elections with diversity constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[8] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the International Conference on World Wide Web*.

[9] Tim Carnes, Chandrashekhar Nagarajan, Stefan M Wild, and Anke Van Zuylen. 2007. Maximizing influence in a competitive social network: a follower's perspective. In *Proceedings of the International Conference on Electronic commerce*.

[10] Damon Centola. 2011. An experimental study of homophily in the adoption of health behavior. *Science* (2011).

[11] Sergio Currarini, Matthew O Jackson, and Paolo Pin. 2010. Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences* (2010).

[12] Maurício de Almeida, Gabriel Mendes, Madras Viswanathan, and Luciano da Silva. 2013. Scale-free homophilic network. *The European Physical Journal B* (2013).

[13] Munmun De Choudhury, Hari Sundaram, Ajita John, Doree Duncan Seligmann, and Aisling Kelliher. 2010. "Birds of a feather": Does user homophily impact information diffusion in social media? *arXiv preprint arXiv:1006.1702* (2010).

[14] Pedro Domingos and Matt Richardson. 2001. Mining the network value of customers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science*. ACM.

[16] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. 2017. Balancing information exposure in social networks. In *Proceedings of the International Conference on Neural Information Processing Systems*.

[17] Sanjeev Goyal, Hoda Heidari, and Michael Kearns. 2014. Competitive contagion in networks. *Games and Economic Behavior* (2014).

[18] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the International Conference on Neural Information Processing Systems*.

[19] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2017. Visibility of minorities in social networks. *arXiv preprint arXiv:1702.00150* (2017).

[20] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the International Conference on World Wide Web (Companion)*.

[21] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[22] Gueorgi Kossinets and Duncan J Watts. 2009. Origins of homophily in an evolving social network. *Amer. J. Sociology* (2009).

[23] Ravi Kumar, Benjamin Moseley, Sergei Vassilvitskii, and Andrea Vattani. 2015. Fast greedy algorithms in mapreduce and streaming. *ACM Transactions on Parallel Computing* (2015).

[24] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Van-Briesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[25] Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

[26] Wei Lu, Wei Chen, and Laks VS Lakshmanan. 2015. From competition to complementarity: comparative influence diffusion and maximization. *Proceedings of the VLDB Endowment* (2015).

[27] Minh-Duc Luu and Ee-Peng Lim. 2015. Latent factors meet homophily in diffusion modelling. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

[28] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* (2001).

[29] Seth A Myers and Jure Leskovec. 2012. Clash of the contagions: Cooperation and competition in information diffusion. In *Proceedings of the IEEE International Conference on Data Mining*.

[30] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming* (1978).

[31] Mark EJ Newman. 2002. Assortative mixing in networks. *Physical Review Letters* (2002).

[32] Mark EJ Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E* (2003).

[33] Ana-Andreea Stoica, Jessy Xinyi Han, and Augustin Chaintreau. 2020. Seeding network influence in biased networks and the benefits of diversity. In *Proceedings of the Web Conference*.

[34] Youze Tang, Yanchen Shi, and Xiaokui Xiao. 2015. Influence maximization in near-linear time: A martingale approach. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

[35] Isabel Valera, Manuel Gomez-Rodriguez, and Krishna Gummadi. 2015. Modeling adoption of competing products and conventions in social media. In *Proceedings of the IEEE International Conference on Data Mining*.

[36] Chi Wang, Wei Chen, and Yajun Wang. 2012. Scalable influence maximization for independent cascade model in large-scale social networks. *Data Mining and Knowledge Discovery* (2012).

[37] Ying Yu, Jinglan Jia, Deying Li, and Yuqing Zhu. 2017. Fair multi-influence maximization in competitive social networks. In *Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications*.