

# Towards Consensus: Reducing Polarization by Perturbing Social Networks

Miklos Z. Rácz  and Daniel E. Rigobon 

**Abstract**—This article studies how a centralized planner can modify the structure of a social or information network to reduce polarization. First, polarization is found to be highly dependent on degree and structural properties of the network – including the well-known isoperimetric number (i.e., Cheeger constant). We then formulate the planner’s problem under full information, and motivate disagreement-seeking and coordinate descent heuristics. A novel setting for the planner in which the population’s innate opinions are adversarially chosen is introduced, and shown to be equivalent to maximization of the Laplacian’s spectral gap. We prove bounds for the effectiveness of a strategy that adds edges between vertices on opposite sides of the cut induced by the spectral gap’s eigenvector. Finally, these strategies are evaluated on six real-world and synthetic networks. In several networks, we find that polarization can be significantly reduced through the addition of a small number of edges.

**Index Terms**—Network consensus and synchronization, opinion dynamics, social influence and recommendations, social network design and architecture.

## I. INTRODUCTION

IN RECENT years there has been a substantial increase in sociopolitical polarization – it is clear that our society does not agree on issues in politics, science, healthcare, and beyond. Counter-intuitively, this has been accompanied by the growth of social media platforms; individuals are connecting with others and sharing information more than ever before. How is it that “bringing the world closer together”<sup>1</sup> resulted in our opinions drifting further apart?

This phenomenon is a byproduct of the structure of our social networks; a greater number of connections does not necessarily reflect a closeness to consensus. It is possible for the proliferation of social media to reduce one’s exposure to other opinions, and thereby entrench them in a community of like-minded users. This feature is known as an “echo chamber,” and has been found to emerge through the incentives of recommender systems rewiring the network [1]. Furthermore, confirmation bias and structural similarity have been found to contribute to increases in polarization as the structure of the network evolves [2], [3].

Manuscript received 20 December 2022; revised 0 2023; accepted 26 March 2023. Date of publication 4 April 2023; date of current version 25 October 2023. This work was supported in part by the NSF under Grant DMS 1811724 and in part by the Princeton SEAS Innovation Award. Recommended for acceptance by Dr. Jiming Chen. (Corresponding author: Daniel E. Rigobon.)

The authors are with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: mracz@princeton.edu; drigobon@princeton.edu).

Digital Object Identifier 10.1109/TNSE.2023.3262970

<sup>1</sup>The original mission statement of Facebook.

Therefore, *how* the population is connected – as opposed to how connected the population *is* – may be most important to the emergence of polarization.

In this article, we seek an understanding of how a network planner can reduce polarization by changing the structure of a population’s social or information network. To that end, we present a model of *budgeted network perturbation*, where the planner is given a small budget with which to modify the structure of a given network. We study the planner’s problem in two different settings, and evaluate simple heuristics on both real-world and synthetic networks.

There has been a significant research effort towards reducing polarization in networks [4], [5], [6], [7], [8]. In contrast to both [7] and [8], we hold fixed the population’s opinions – while allowing the network structure to be modified. This article differs from [5] and [6] in both our use of a distinct measure of polarization and incorporating opinion dynamics. Finally, we improve upon the closely related work of [4] through more detailed theoretical analysis, consideration of weighted networks, and use of larger datasets.

A very similar article to our own is recent work by [9], where the authors present a variation of the problem studied in [10]. Both these studies aim to minimize the sum of polarization and disagreement by changing the network structure, but [9] impose a budget that ensures only a small number of edges can be changed. These authors use a similar budget constraint to our own, but their polarization-disagreement index varies greatly with the edge density of the graph. Although it is convenient for analysis and computation, their index is inadequate for capturing the dynamics of polarization alone. Nonetheless, we believe the formulation in this article and [9] to be practical. The network structure is not assumed to be completely malleable, but small changes are permitted. For instance, while social media platforms such as Facebook or Twitter cannot dictate who an individual chooses to ‘friend’ or ‘follow,’ these platforms can curate an individual’s feed to change one’s relative exposure levels to certain content. This process perturbs the structure of external influence on an individual, so that it differs from their endogenously created network of ‘friends’ or ‘follows’. If, instead, any of these platforms suddenly decided to completely rewire their social networks, users may be upset.

It is then natural to consider the questions in this article: how does the network planner decide to allocate their budget? How much of an impact can be made? How large of a budget is needed to achieve a significant reduction in polarization?

We begin by first establishing a relationship between structural properties of a social network and its level of polarization. We find that both the degree profiles and the strength of information bottlenecks – quantified by the well-known Cheeger constant in spectral graph theory – are closely tied to polarization. This result naturally captures the intuition and dangers of echo chambers in real-world networks.

Next, we focus on the formulation and analysis of two settings for network optimization. In the first, the planner has full information of the population's opinions. We provide theoretical motivation for two heuristics: coordinate descent and a stepwise disagreement-seeking algorithm. The former is standard in optimization, while the latter is the antithesis of confirmation bias. Existing research has shown that addition of edges between like-minded individuals contributes to increasing polarization [2]. Moreover, according to [11] it is 'costly' for individuals to be connected to others who disagree with them, and recommender systems can be designed to minimize disagreement [1]. Therefore, the incentives of both individuals and social media platforms may naturally lead to polarization growing over time. In contrast, we show that a simple disagreement-seeking approach taken by the planner leads to substantial reductions in polarization. This result is closely tied to our choice of the opinion dynamics model. In this article, interactions between individuals are always *attractive* – bringing opinions closer together. In reality, this is not the case (see [12], [13]). This simplification, however, will facilitate theoretical results – which we believe can be leveraged for partial understanding of polarization-reduction strategies in a richer class of models.

This article also presents a novel setting for the network planner, wherein the population's opinions are chosen adversarially. In several articles from the literature (see, for instance [7], [8], [14], [15]), an adversary is able to change the individuals' opinions – seeking to maximize polarization. The setting we study represents a planner whose network design must be *robust* to the adversary's disruption. We show that this setting for the planner's problem is intimately related to maximizing the spectral gap of the graph's Laplacian, which is a well-studied problem [16], [17], [18]. We provide theoretical guarantees for a heuristic that connects vertices on opposite sides of the cut corresponding to the spectral gap.

We then evaluate several natural heuristics on real-world and synthetic networks. There are significant reductions in polarization for networks with strong initial community structures. Furthermore, we study how the spectral gap and homophily are affected by the planner's modifications. We find that the largest reductions in polarization are accompanied by reductions in homophily. In many cases, however, one of our heuristics effectively reduces polarization with little effect on homophily. We also observe that two heuristics lead to vertices with extreme opinions becoming more central in the graph structure. In many of the networks studied, a small budget yields substantial reductions in polarization.

The article is organized as follows. Section II reviews recent and related work. Section III introduces relevant notation, definitions, and preliminaries. Section IV provides theoretical ground for three heuristics, which are described and evaluated

on several networks in Section V. Finally, Section VI concludes and discusses directions for future work.

## II. RELEVANT LITERATURE

The articles most similar to our own are recent studies by [4], [15], [1], and [9]. [15] motivates our adversarial disruption of the population's opinions, while both [4] and [9] aim to modify a social network's structure by adding a small number of edges. [1] impose a constraint on the edge weight modified – but not the number of edges. In particular, they focus on changing a large number of edges by a small amount, whereas we seek to do the converse. Our work differs from [4] through greater emphasis on theory and generalization to weighted graphs. The objective function in [9] fundamentally differs from our own, and represents a different research problem.

In addition, this article is broadly tied to the literature on opinion dynamics, perturbation of network structures, and influencing polarization. Relevant studies in each of these areas are discussed in the following.

1) *Opinion Dynamics*: The study of consensus-forming begins with the seminal work of [19], where under weak conditions on the social network, the opinions eventually converge to a perfect consensus. This model was expanded by [20] (and more recently by [21]), so that the long-term opinions are heterogeneous. Because of this feature and its simplicity, the Friedkin-Johnsen (FJ) model has appeared in several recent studies on opinion polarization and disagreement – see for instance, [1], [4], [7], [9], [10], [14], [15]. In this article, we will also use the FJ model. Not only is it standard in the literature, but it is mathematically convenient for analysis. There are also rich areas of work which justify and extend the FJ model. For instance, [11] show that the expressed opinions of this model correspond to the Nash equilibrium of a network game.

There are a few notable extensions to the FJ model, in which individuals have more complex behavior. For example, a recent survey by [22] presents several generalizations and (relevantly) assesses if polarization can occur in each. A fundamental feature of the FJ model is that individuals are always drawn toward the opinions of their neighbors – but experimental evidence of this feature is inconclusive and contextual [12], [13]. Motivated by this observation, new models have been developed in which individuals have bounded confidence [23] or even experience repulsion [8], [24]. It is also possible to incorporate geometric structures into the dynamics, such as recent work by [25] and [26]. Finally, we note that there are several related studies within the controls literature, which focus on consensus dynamics on a network, for instance, when agents have antagonistic dynamics [27], or are stubborn [28] – see [29] for a more complete survey.

2) *Optimizing Network Structures*: This article formulates an optimization problem over network structures, aiming to reduce a particular definition of polarization. There are several related works in the literature. For example, [10] allows unconstrained rewiring of the social network to reduce the *polarization-disagreement index*, which is defined as the sum of polarization and disagreement. A recent article of [9] optimizes

the same index via addition of a limited number of edges. This index is analytically and computationally convenient because of its monotonicity and convexity, but it is highly sensitive to the edge density of the graph.<sup>2</sup> We instead focus exclusively on minimizing polarization, which is shown to be neither convex nor monotone in Section IV-B. However, this article restricts edge modifications similarly to [9].

A more closely related work by [4] presents several definitions of ‘conflict’ in social networks, and studies how they can be minimized through iterative perturbations to the graph. One such measure of conflict equals polarization. We expand on the authors’ work by providing a detailed theoretical analysis of edge perturbations on polarization, generalizing the analysis to weighted graphs, and conducting simulations on larger real-world and synthetic networks.

The aforementioned articles share with ours a definition of polarization. However, it is possible to optimize for other notions of ‘cohesiveness’ or ‘consensus’. For instance, [5] and [6] both present measures of polarization based on random walks, and propose algorithms for reducing it via edge addition. The greatest similarity between their work and ours lies in the use of a greedy, stepwise approach to an optimization problem. However, the authors’ definitions of polarization do not directly incorporate opinion dynamics.<sup>3</sup> Moreover, in [6], nodes represent webpages, not individuals.

Another definition of cohesiveness, which does not depend on any node opinions or labels, is the spectral gap of a graph. The spectral gap controls the synchronizability of dynamical systems and mixing times of Markov chains [16], and therefore its maximization is of great interest. For instance, [18] seek to increase the spectral gap by removing nodes. Unlike these authors, we focus on changes to a graph’s edges. More relevantly, [17] study how the algebraic connectivity (i.e., spectral gap) can be increased by adding edges. The authors present two strategies for doing so, one of which is derived from the eigenvector corresponding to the spectral gap. In this article, we show that the adversarial setting of the planner’s problem is closely related to their work, and provide bounds on polarization using this eigenvector-based strategy.

3) *Natural Network Dynamics*: A different branch of research aims to understand how polarization is shaped by *rewiring dynamics* in the network. For instance, a recent article by [2] studies how individuals’ local rewiring rules can lead to higher polarization. The authors conclude that confirmation bias and friend-of-friend behavior are critical for this result. However, their theoretical results focus on the polarization-disagreement index. Moreover, we derive an improved upper bound for polarization in Section IV-A. A similar article by [3] shows that allowing individuals’ rewiring decisions based on structural similarity leads to polarization, although the authors use a distinct model of opinion dynamics.

<sup>2</sup>The polarization-disagreement index consists of adding polarization, which is on the order of  $n$  (the number of vertices), and disagreement, which is of order  $m$  (the number of edges). Therefore this index is dominated by disagreement for dense graphs (specifically, if  $m \gg n$ ).

<sup>3</sup>We note that the Friedkin-Johnson model has a random walk interpretation of the long-term opinions, see [30].

It is also possible to study the dynamics driven by a network administrator. [1] present a setting in which a network administrator rewrites the network over time by providing ‘recommendations’ to users based on minimizing disagreement. They show that without a regularization term in the optimization problem, the administrator greatly increases polarization. The authors’ result contrasts with one of the main findings of this article, namely that connecting disagreeing individuals is effective for reducing polarization.

4) *Optimizing Opinion Profiles*: While less relevant to this article, a complimentary line of work assumes that the network structure remains fixed, but the innate opinions are subject to change. For instance, [30] establish NP-Hardness of an opinion maximization problem, in which an administrator takes over a small set of individuals and sets their opinions to the largest possible value. Articles by [7] (resp. [31]) seek to minimize polarization (resp. maximize diversity, i.e., disagreement) by choosing a small subset of individuals to have neutral opinions. Finally, the work of [8] aims to minimize polarization in an extension of the FJ model by slightly shifting individuals’ opinions.

These studies have generally taken the perspective of a benevolent network planner. It is also possible to consider the perspective of an adversary, who takes over a small number of individuals and seeks to maximize polarization or disagreement [14]. A more powerful adversary in [15] chooses the opinions of the entire population to the same end. In particular, [15] present a problem of *defending* the network from this adversary by making some opinions more resistant to change. In this article, we will consider a similar setting, but where the network is defended by altering its structure instead. Nonetheless, the adversary faced is modeled on their work.

### III. MODEL

An undirected graph  $\mathcal{G}(V, E, W)$  is defined by a set of vertices  $V$  given by  $[n] := \{1, \dots, n\}$ , a set of edges  $E \subset V \times V$  consisting of unordered pairs of vertices, and symmetric edge weight matrix  $W \in [0, \bar{w}]^{n \times n}$ .  $W$  is assumed to satisfy  $w_{ij} > 0$  if and only if  $(i, j) \in E$ , and  $\bar{w} < \infty$  indicates the maximum possible edge weight. For a graph  $\mathcal{G}$ , its degree matrix  $D$  is diagonal, and satisfies  $D_{ii} = d_i$ , where  $d_i = \sum_j w_{ij}$  is the (weighted) degree of vertex  $i$ . Let  $L = D - W$  denote the combinatorial graph Laplacian, and  $\mathcal{L} = D^{-1/2} L D^{-1/2}$  denote the normalized Laplacian. Finally,  $N(i) := \{j \in [n] : (i, j) \in E\}$  denotes the neighbors of vertex  $i$ .

Vertices are given *innate opinions*  $\mathbf{s} \in [0, 1]^n$ , which represent a continuum between two extreme positions on an issue. For instance, an individual who is totally in favor of strict firearm laws may have an opinion of 0, whereas one extremely against any such regulations would have an opinion of 1. The population’s opinions evolve over time, beginning from the innate opinions  $\mathbf{s}$ . The evolution of opinions follows the dynamics of [20] (see below), and the opinions converge to a fixed point – denoted  $\mathbf{z}$  and called the *expressed opinions* of the population. In this article, we are interested in modifications to the underlying graph  $\mathcal{G}$ , and therefore take the innate opinions  $\mathbf{s}$  to be fixed.



Consequently, we write  $\mathbf{z}$  and  $\mathbf{z}'$  for the expressed opinions corresponding to the social networks  $\mathcal{G}$  and  $\mathcal{G}'$ , respectively. Occasionally, to emphasize the underlying graph  $\mathcal{G}$ , we will write  $\mathbf{z}_{\mathcal{G}}$ .

#### A. Opinion Dynamics

In the seminal model of [19], the population's expressed opinions converge to a perfect consensus under weak conditions. A notable extension of the DeGroot model is by [20], whose model preserves long-term heterogeneity of opinions. In particular,  $\mathbf{z} = c\vec{1}$  if and only if  $\mathbf{s} = c\vec{1}$ . This model is convenient for analysis because the expressed opinions can be written explicitly. Furthermore, several recent works in the literature have leveraged this model – see Section II for more.

The Friedkin-Johnsen (FJ) opinion dynamics model is specified by the discrete-time mapping  $\mathbf{s}(t) \rightarrow \mathbf{s}(t+1)$  as follows. We initialize  $\mathbf{s}(0) = \mathbf{s}$ , and iterate

$$s_i(t+1) = \frac{s_i(0) + \sum_{j \in N(i)} w_{ij} s_j(t)}{1 + \sum_{j \in N(i)} w_{ij}},$$

where  $w_{ij}$  is the weight associated with edge  $(i, j)$ , and is non-zero if and only if  $j \in N(i)$ . The expressed opinions  $\mathbf{z}$  are the fixed point of this mapping, given by

$$\mathbf{z} = (I + L)^{-1} \mathbf{s},$$

where  $I$  denotes the  $n \times n$  identity matrix. Notice that  $I + L \succcurlyeq I$  is necessarily invertible. Thus, there exist unique expressed opinions  $\mathbf{z}$  for any given  $\mathcal{G}$  and  $\mathbf{s}$ . Moreover, since the eigenvalues of  $(I + L)^{-1}$  are no greater than 1, the expressed opinions of the FJ dynamics are a contraction of the innate opinions. This observation also follows from the fact that the FJ model is purely *attractive* – opinions of connected individuals are always drawn to each other over time. One of the heuristics in this article will depend on this feature of the dynamics. However, exposure to substantially differing opinions in the real-world may yield no effect, or even strengthen one's original position. In Section VI we discuss how our results might be leveraged for such a class of richer opinion dynamics models, and relevant directions for future work.

#### B. Polarization and Disagreement

In practice, a perfect consensus is rare; therefore, we seek to understand “closeness” to consensus. Accordingly, we define *polarization* to be proportional to the variance of the expressed opinions. Large polarization indicates that the population is far from achieving a consensus, and vice-versa. Formally:

**Definition 1 (Polarization):** Given a vector of opinions  $\mathbf{x} = (x_1, \dots, x_n)$  and the mean of its entries  $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ , the *polarization* of  $\mathbf{x}$  is

$$P(\mathbf{x}) := \sum_{i=1}^n (x_i - \bar{x})^2 = \|\tilde{\mathbf{x}}\|^2,$$

where  $\tilde{\mathbf{x}} := \mathbf{x} - \bar{x}\vec{1}$  are the mean-centered opinions.

In particular,  $P(\mathbf{z})$  is *expressed* polarization, and  $P(\mathbf{s})$  is *innate* polarization.

It is useful to define *disagreement*, which captures distance from consensus on a local scale. Intuitively, if two vertices have very distinct opinions, their disagreement is large.

**Definition 2 (Disagreement):** For any vector of opinions  $\mathbf{x} = (x_1 \dots x_n)$ , the disagreement between vertices  $i$  and  $j$  is given by:

$$D_{ij}(\mathbf{x}) := (x_i - x_j)^2.$$

Again, between vertices  $i$  and  $j$ ,  $D_{ij}(\mathbf{z})$  is the *expressed* disagreement, while  $D_{ij}(\mathbf{s})$  is the *innate* disagreement. The two quantities above have been studied in several recent articles on social and information networks; see [1], [2], [3], [4], [7], [8], [9], [10], [14], [15] and references therein.

### IV. THEORETICAL RESULTS

We now present several theoretical results on polarization. We study how its magnitude depends on structural properties of the graph, and how it can vary as edges are modified.

#### A. Opinion Contraction and Polarization

This article is primarily concerned with polarization of expressed opinions,  $P(\mathbf{z})$ . However, the relationship between expressed and innate polarization depends on  $\mathcal{G}$ . Since the opinion dynamics model performs a contraction on the opinions, it follows that  $P(\mathbf{z}) \leq P(\mathbf{s})$ . In fact, more is true: the contraction ratio is controlled by structural properties of  $\mathcal{G}$ .

To present this result, we must introduce some notation. For any two disjoint subsets of vertices  $B_1$  and  $B_2$ , let  $E(B_1, B_2)$  denote the set of edges with one incident vertex in  $B_1$  and the other in  $B_2$ . The conductance of  $X \subset V$  is defined as

$$h_{\mathcal{G}}(X) := \frac{\sum_{(i,j) \in E(X, X^c)} w_{ij}}{\min\{\sum_{v \in X} d_v, \sum_{u \in X^c} d_u\}}.$$

The isoperimetric number (also known as the Cheeger constant) of a graph  $\mathcal{G}$  is given by

$$h_{\mathcal{G}} := \min_{X \subset V, 0 < |X| < |V|} h_{\mathcal{G}}(X), \quad (1)$$

as in [32], and will appear in the results. Note that  $h_{\mathcal{G}} \leq 1$ , since  $h_{\mathcal{G}}(X) = 1$  when  $X$  consists of a single vertex, and  $\mathcal{G}$  is disconnected if and only if  $h_{\mathcal{G}} = 0$ . The isoperimetric number of a graph indicates the presence of bottlenecks – it is small when there exists a large subset of vertices that are sparsely connected to others.

We now arrive at a first result on the contraction properties of the FJ model on polarization.

**Proposition 3:** Let  $d_{\min}$  and  $d_{\max}$  be the minimum and maximum weighted degrees in  $\mathcal{G}$ , and let  $h_{\mathcal{G}}$  be its isoperimetric number. Then,

$$\frac{P(\mathbf{s})}{(1 + (2d_{\max}) \wedge (\bar{w}n))^2} \leq P(\mathbf{z}) \leq \frac{P(\mathbf{s})}{(1 + \frac{1}{2}d_{\min}h_{\mathcal{G}}^2)^2}.$$

Proposition 3 quantifies the effects of the FJ model on polarization. In particular, if  $\mathcal{G}$  has strong expander properties (i.e.,  $h_{\mathcal{G}}$  is large), then we expect the expressed polarization to be small, relative to the innate polarization. The proof of this result can

be found in Appendix A, and follows from simple eigenvalue bounds and a version of Cheeger's inequality.

This result provides a tighter upper bound on polarization than that of [2]. The tightening is achieved by observing that the mean-centered innate opinions  $\tilde{\mathbf{s}}$  are orthogonal to the eigenvector of  $(I + L)^{-2}$  that has corresponding eigenvalue 1. In addition, we can use Proposition 3 to show that the complete graph  $K_n$ , with all edge weights equal to the maximal  $\bar{w}$ , is a global minimum for polarization.

**Corollary 4:** Fix innate opinions  $\mathbf{s}$ , and let  $\mathcal{G}$  be any graph on  $n$  vertices with maximal edge weight  $\bar{w}$ . Let  $\mathbf{z}_{K_n}$  and  $\mathbf{z}_{\mathcal{G}}$  denote the expressed opinions on  $K_n$  and  $\mathcal{G}$ , respectively. Then,

$$P(\mathbf{z}_{K_n}) \leq P(\mathbf{z}_{\mathcal{G}}).$$

Moreover,  $P(\mathbf{z}_{K_n}) = \frac{P(\mathbf{s})}{(1+\bar{w}n)^2}$ .

The key observation in the proof of Corollary 4 is that all non-zero eigenvalues of  $L_{K_n} = \bar{w}(nI - \mathbf{1}\mathbf{1}^T)$  (the Laplacian of the complete graph) are equal to  $\bar{w}n$ . Therefore, for any  $\mathcal{G}$ , the value of polarization on  $K_n$  achieves with equality the smallest lower bound from Proposition 3. This result also provides a useful reference point for closeness to global optimality.

### B. Given Opinions

We now turn to studying how the planner can decrease polarization by modifying the graph.

In a first setting, we assume that the innate opinions are known. If the planner can change (by adding or removing) the edge weight between at most  $k$  pairs of vertices, what is the least polarization they can achieve? Given a graph  $\mathcal{G}$ , innate opinions  $\mathbf{s}$ , and integer budget  $k > 0$ , we have

$$\begin{aligned} \min_{\mathcal{G}'} P(\mathbf{z}') \\ \text{s.t. } \|W - W'\|_0 \leq 2k, \end{aligned} \quad (2)$$

where the expressed opinions  $\mathbf{z}'$  correspond to  $\mathcal{G}'$ , which must also be an undirected graph with maximal edge weight  $\bar{w}$ . The factor of two in the constraint of (2) follows from our assumption of undirected graphs. The constraint naturally captures the assumption that it is costly for the planner to modify an edge, but upon committing to doing so, they may freely change the edge weight.<sup>4</sup>

Problem (2) may be challenging to solve efficiently since it is non-convex. Beyond the fact that the  $\ell_0$  norm gives a non-convex feasible set, the objective function is also not convex – see Fig. 1 for a small example. Therefore, relaxing the  $\ell_0$  constraint to  $\ell_1$  still yields a non-convex optimization problem. Instead of seeking an optimal set of  $k$  edges to add, we propose a greedy stepwise approach where the weight of  $k$  edges are saturated iteratively, one at a time. This simpler setting is tractable for analysis.

It seems intuitive that adding edge weight to  $\mathcal{G}$  promotes the flow of information, and thereby reduces polarization. However,

<sup>4</sup>The constraint could also bound the absolute difference in edge weights ( $\ell_1$  norm). This is a different problem – more similar to [1] – but an interesting direction for future work. In this case, we believe that the planner would distribute edge weight to maximize the least marginal return of polarization.

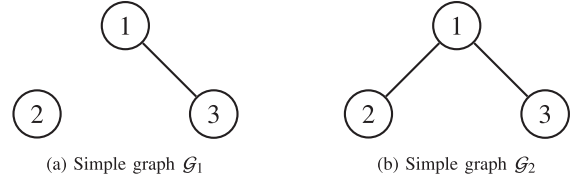


Fig. 1. A simple example of the non-convex objective function. With innate opinions  $\mathbf{s} = [0, 0.4, 1]$ , it can be seen that  $P(\frac{1}{2}[L_1 + L_2]) > \frac{1}{2}[P(L_1) + P(L_2)]$ . (Note the abuse of notation to illustrate  $P(\cdot)$ 's dependence on only the Laplacian.) In this particular example, the addition of any amount of weight to edge (1,3) increases polarization. (a) Simple graph  $\mathcal{G}_1$ . (b) Simple graph  $\mathcal{G}_2$ .

this is not the case in general. We will see that for most non-saturated edges, there exist some innate opinions for which the addition of weight to that edge will increase polarization. The exact expression for the change in polarization when adding edge weight is given in the following.

**Lemma 5:** Let  $\mathcal{G}(V, E)$  be an undirected graph yielding expressed opinions  $\mathbf{z}$ , and  $(i, j)$  be a pair of vertices with non-maximal weight, that is,  $w_{ij} < \bar{w}$ . Let  $\mathbf{v}_{ij} := \mathbf{e}_i - \mathbf{e}_j$ . For  $\delta \in (0, \bar{w} - w_{ij}]$ , we construct  $\mathcal{G}^+(V, E^+, W^+)$  according to  $w_{ij}^+ = w_{ij} + \delta$ , and  $E^+ = \{(i, j) : w_{ij}^+ > 0\}$ . If the expressed opinions on  $\mathcal{G}^+$  are given by  $\mathbf{z}^+ := (I + L^+)^{-1}\mathbf{s}$ , then

$$\begin{aligned} P(\mathbf{z}) - P(\mathbf{z}^+) = D_{ij}(\mathbf{z}) \left[ \frac{2\delta \tilde{\mathbf{z}}^T (I + L)^{-1} \mathbf{v}_{ij}}{\tilde{\mathbf{z}}^T \mathbf{v}_{ij} (1 + \delta \mathbf{v}_{ij}^T (I + L)^{-1} \mathbf{v}_{ij})} \right. \\ \left. - \frac{\delta^2 \mathbf{v}_{ij}^T (I + L)^{-2} \mathbf{v}_{ij}}{(1 + \delta \mathbf{v}_{ij}^T (I + L)^{-1} \mathbf{v}_{ij})^2} \right]. \end{aligned} \quad (3)$$

The proof of this result can be found in Appendix A. To discuss this result, it is useful to define the following.

**Definition 6** ( $\partial_{w_{ij}} P(L)$ ): Fix some innate opinions  $\mathbf{s}$ . Let  $\mathbf{z}_L$  denote the resulting expressed opinions when the underlying graph  $\mathcal{G}$  has Laplacian  $L$ . We write:

$$\partial_{w_{ij}} P(L) = \lim_{t \rightarrow 0^+} \frac{P(\mathbf{z}_{L+tL_{ij}}) - P(\mathbf{z}_L)}{t} \quad (4)$$

where  $L_{ij} = \mathbf{v}_{ij} \mathbf{v}_{ij}^T$ .

This definition allows us to analyze the first-order effects of edge modifications on polarization. Notice that even if a graph were unweighted, we can define this derivative for its equivalent *weighted* graph, where the weight of each existing edge equals one. In the following proposition, we derive a closed form expression for these partial derivatives.

**Proposition 7:** For fixed innate opinions  $\mathbf{s}$ , we have

$$\begin{aligned} \partial_{w_{ij}} P(L) &= -2\tilde{\mathbf{s}}^T (I + L)^{-2} L_{ij} (I + L)^{-1} \tilde{\mathbf{s}} \\ &= -2\tilde{\mathbf{z}}^T (I + L)^{-1} L_{ij} \tilde{\mathbf{z}}. \end{aligned}$$

This result allows us to re-write the right-hand side of (3):

$$\begin{aligned} -\delta \partial_{w_{ij}} P(L) &= \frac{\delta^2 \mathbf{v}_{ij}^T (I + L)^{-2} \mathbf{v}_{ij}}{(1 + \delta \mathbf{v}_{ij}^T (I + L)^{-1} \mathbf{v}_{ij})^2} (z_i - z_j)^2. \end{aligned}$$

Therefore, the necessary and sufficient condition for a reduction in polarization due to adding weight  $\delta$  to edge  $(i, j)$  is:

$$-\partial_{w_{ij}} P(L) > (z_i - z_j)^2 \frac{\mathbf{v}_{ij}^T (I + L)^{-2} \mathbf{v}_{ij}}{\delta^{-1} + \mathbf{v}_{ij}^T (I + L)^{-1} \mathbf{v}_{ij}},$$

which amounts to a steep enough first derivative.

Lemma 5 also allows us to study when polarization *increases* after adding weight to edge  $(i, j)$ . In particular, if  $\frac{\tilde{\mathbf{z}}^T (I+L)^{-1} \mathbf{v}_{ij}}{\tilde{\mathbf{z}}^T \mathbf{v}_{ij}} = 0$ , then  $P(\mathbf{z}^+) \geq P(\mathbf{z})$ . Notice that if  $\mathbf{v}_{ij}$  is not an eigenvector of  $L$ , then this modification can increase polarization when the mean-centered innate opinions  $\tilde{\mathbf{s}}$  lie on the  $(n-1)$ -dimensional subspace orthogonal to  $(I+L)^{-2} \mathbf{v}_{ij}$ . This condition is sufficient but not necessary – the example in Fig. 1 illustrates this point. Therefore, the planner cannot add edge weight arbitrarily and expect polarization to be reduced – the innate opinions can determine the effect's sign.

However, there are special cases in which polarization is always reduced, such as the following.

*Corollary 8:* If  $\mathcal{G}$ ,  $i$ , and  $j$  satisfy  $N(i) = N(j)$ , then polarization is always reduced by adding weight  $\delta$  to the edge  $(i, j)$ , and the difference is

$$P(\mathbf{z}) - P(\mathbf{z}^+) = (z_i - z_j)^2 \frac{2\delta(1 + \delta + d_i - w_{ij})}{(1 + 2\delta + d_i - w_{ij})^2}.$$

This result follows from proving that  $L\mathbf{v}_{ij} = (d_i - w_{ij})\mathbf{v}_{ij}$  under the assumptions; see Appendix A for full details.

Corollary 8 is somewhat counter-intuitive – if we strengthen connections between individuals who share the same set of neighbors, we may expect to form an ‘echo chamber’. However, the opinion dynamics show that the addition of weight to such an edge  $(i, j)$  will *only* affect the expressed opinions of vertices  $i$  and  $j$ . While this edge fails to have any global effect, it does indeed bring the opinions of its incident vertices closer together – hence reducing polarization. The limitation of these effects to only its incident vertices suggests that in practice, the return on polarization may be small.

Lemma 5 is also used for arriving at one of our main results.

*Theorem 9:* Let  $\mathbf{z}$ ,  $\mathbf{z}^+$ ,  $\delta$ , and  $\mathbf{v}_{ij}$  be as before. Then,

$$P(\mathbf{z}) - P(\mathbf{z}^+) \leq \frac{1 + \lambda_n(L)}{1 + 2\delta + \lambda_n(L)} (-\delta \partial_{w_{ij}} P(L)).$$

Furthermore, if there exists  $\epsilon > 0$  for which

$$\frac{\tilde{\mathbf{z}}^T (I + L)^{-1} \mathbf{v}_{ij}}{\tilde{\mathbf{z}}^T \mathbf{v}_{ij}} \geq \epsilon + \frac{\delta}{2\delta + (1 + \lambda_2(L))^2},$$

then we also have

$$P(\mathbf{z}) - P(\mathbf{z}^+) \geq \frac{2\delta\epsilon(z_i - z_j)^2}{1 + 2\delta}.$$

Theorem 9 directly motivates two heuristics for the planner. First, we see that the largest possible reduction in polarization is proportional to the first order effect  $-\delta \partial_{w_{ij}} P(L)$ . Therefore, it is natural for the planner to iteratively add maximal edge weight along the direction of steepest descent – a heuristic well-known as a *coordinate descent*. Additionally, for fixed  $\epsilon$ , the lower bound grows with the expressed disagreement. Therefore, edges

with large  $(z_i - z_j)^2$  are also good candidates for the planner to add weight to; we name this strategy *disagreement-seeking*.

The upper bound in Theorem 9 implies that there are diminishing returns with adding more weight to a single edge, as  $\frac{1 + \lambda_n(L)}{1 + 2\delta + \lambda_n(L)} < 1$ . Although  $P(\mathbf{z})$  is not globally convex, this shows that it is convex along the direction  $w_{ij}$ .

### C. Adversarial Opinions

In some cases, the planner may not reliably use the innate or expressed opinions. For instance, they may be difficult (even impossible) to measure, or vertices may be susceptible to takeovers; see [7], [8], [14], [15], [30], [31] for examples of the latter. Moreover, individuals' opinions may be multidimensional – capturing many distinct issues (e.g., firearm regulation, universal basic income, healthcare, etc.), all of which are shaped by the network's structure. Such cases may require the planner to take a robust approach: they seek to design a network structure that minimizes polarization for *any* possible vector of innate opinions.<sup>5</sup> Formally, they aim to solve:

$$\begin{aligned} \min_{\mathcal{G}'} \max_{\mathbf{s} \in \mathbb{R}^n: \|\mathbf{s}\|_2 \leq R} \tilde{\mathbf{s}}^T (I + L')^{-2} \tilde{\mathbf{s}} \\ \text{s.t. } \|W - W'\|_0 \leq 2k. \end{aligned} \quad (5)$$

Polarization in the resulting graph  $\mathcal{G}'$  will be robust to the choice of innate opinions, and this optimization problem yields different graph structures than problem (2). Again, the factor of two in the constraint follows from having undirected graphs.

This optimization problem can be interpreted as a game – an adversary selects  $\mathbf{s}$  from the  $n$ -dimensional sphere of radius  $R$ , and the planner evaluates polarization on this choice of  $\mathbf{s}$ . A similar problem appears in [15], who studies a ‘network defender’ that decreases vertices' susceptibility to the adversary. In contrast, we consider defending the network through modification of its structure. However, both our defender and theirs face the same adversary. This choice allows us to directly compare the effectiveness of these two defensive strategies. Although such an adversary may not be realistic, we believe this setting has numerous other justifications.

Note that the innate opinions now lie in the  $n$ -dimensional sphere, as opposed to the hypercube. This formulation allows us to relate the adversary's problem to spectral properties of the resultant graph  $\mathcal{G}'$ . In fact, problem (5) is equivalent to maximizing  $\lambda_2$ , the spectral gap of the Laplacian.

*Proposition 10:* The optimal solution  $\mathcal{G}'$  to (5) is the same as that of

$$\begin{aligned} \max_{\mathcal{G}'} \lambda_2(L') \\ \text{s.t. } \|W - W'\|_0 \leq 2k, \end{aligned} \quad (6)$$

If the optimal solution to (6) is  $L^*$ , then the optimal value of (5) is  $\frac{R}{(1 + \lambda_2(L^*))^2}$ .

For two graphs  $\mathcal{G}$  and  $\mathcal{G}'$ , if  $W \leq W'$  elementwise, then  $L \preceq L'$ , and therefore  $\lambda_2(L) \leq \lambda_2(L')$ . Therefore, the planner

<sup>5</sup>There is one other possible justification for this formulation – a robust (or minimax) optimization problem arises when the decision-maker is ambiguity averse, as is shown axiomatically by [33].

must only add edge weight to  $\mathcal{G}$ , as reducing weights cannot increase the spectral gap.<sup>6</sup> The spectral gap of the Laplacian is intimately tied to the synchronizability of various types of dynamical systems and the mixing time of Markov chains [16] and hence several studies seek to maximize it [17], [18]. In this article, where perfect synchronization is impossible, the spectral gap controls the best achievable consensus.

The proof of Proposition 10 follows from solving the inner maximization problem, for which the optimal solution is the eigenvector of  $L'$  corresponding to the second-smallest eigenvalue. This eigenvector is called the *Fiedler* vector of  $\mathcal{G}'$ , and describes a partition of vertices that approximates the normalized sparsest cut of  $\mathcal{G}$  [32].

For a graph with Laplacian  $L$ , Proposition 10 indicates that the *worst-case* polarization is equal to  $P(L) = \frac{R}{(1+\lambda_2(L))^2}$ . The adversary achieves this by choosing  $\tilde{\mathbf{s}}$  along the span of the Fiedler vector. The planner's effectiveness in (5) is controlled by  $P(L) - P(L')$ , the difference in *worst-case* polarization.

As in the previous setting, we approach this problem by iteratively choosing edges to saturate – starting from the initial graph until no further budget remains. Therefore, the principal results address how increasing an edge's weight affects the spectral gap and thereby polarization. This is quantified in Theorem 11, which relates changes in the spectral gap to elementwise differences in the Fiedler vector.

**Theorem 11:** Let  $\mathcal{G}$  be an undirected graph, and  $(i, j)$  be an edge with non-maximal weight, that is,  $w_{ij} < \bar{w}$ . Let also  $\mathbf{v}$  be the Fiedler vector of  $\mathcal{G}$  of unit magnitude with corresponding eigenvalue  $\lambda_2(L)$ . Recall that  $\lambda_3(L)$  is the third smallest eigenvalue of  $L$ , and define  $\beta = \lambda_3(L) - \lambda_2(L)$ .

For some  $\delta \in (0, \bar{w} - w_{ij}]$ , let  $\mathcal{G}^+$  be constructed by adding weight  $\delta$  to edge  $(i, j)$ . If  $\alpha = |v_i - v_j|$ , then we have that

$$\max \left\{ 1 - \frac{2\delta}{\beta}, 0 \right\} \delta \alpha^2 \leq \lambda_2(L^+) - \lambda_2(L) \leq \delta \alpha^2. \quad (7)$$

The proof follows from adapting the result of [35]. The bounds are tightest when  $\beta$  is largest, equivalently when  $\lambda_2(L)$  is the sole small eigenvalue of  $L$ .

This result motivates a simple heuristic for (6), which appears in [17]. The planner can iteratively compute the Fiedler vector and add weight to non-saturated edges whose incident vertices have large absolute difference in  $\mathbf{v}$ .

Corollary 12 quantifies the effects on polarization induced by the perturbation in Theorem 11.

**Corollary 12:** Let  $P(L) = \frac{R}{(1+\lambda_2(L))^2}$  be the worst-case polarization on a graph with Laplacian  $L$ . In the setting of Theorem 11, we have

$$\begin{aligned} P(L) - P(L^+) &\geq \frac{2R\delta}{(1+2\delta+\lambda_2(L))^3} \max \left\{ 1 - \frac{2\delta}{\beta}, 0 \right\} \alpha^2 \\ P(L) - P(L^+) &\leq \frac{4R(\delta \vee \delta^2)}{(1+\lambda_2(L))^3} \alpha^2. \end{aligned}$$

<sup>6</sup>We remark that this monotonicity of the spectral gap in the edge set does not hold for the normalized Laplacian  $\mathcal{L}$ , see for instance [34].

In contrast to the setting with full information, the worst-case polarization  $P(L)$  cannot increase when the planner increases an edge's weight. Recall that this follows from the monotonicity of the spectral gap in  $W$ . However, it is possible that the resulting graph  $\mathcal{G}^+$  has greater polarization for *some* particular innate opinions. The settings in (2) and (5) are distinct, and therefore the quantities compared before and after edge-weight addition are fundamentally different.

## V. EMPIRICAL SIMULATIONS

If we solved problems (2) or (6) naively, it would be necessary to test all  $\sum_{i=1}^k \binom{n}{i}$  possibilities. Given that computing polarization (or the spectral gap) requires  $O(n^3)$  time, we obtain a crude upper bound of  $O(kn^{2k+3})$ . Note that for fixed  $k$ , this rate is polynomial in  $n$  – albeit still not scalable. However, in subsequent experiments we choose  $k$  to grow linearly with  $n$ , which results in superexponential runtime. It is therefore extremely impractical to compute the optimal solution, and we resort to theoretically motivated heuristics.

In Sections IV-B and IV-C, we briefly discussed three heuristics for solving the planner's problem in a greedy, iterative fashion. Our theoretical results studied how polarization is reduced by *addition* of weight to a single edge. Therefore, all of the following heuristics are based on increasing edge weights. These are presented below – detailing the edge to be saturated (i.e. setting edge weight to  $\bar{w}$ ) at every step and briefly discussing the time complexity of each iteration. We will compare these approaches with a random baseline.

- *Random:* Add an edge from  $E^C$  chosen uniformly at random; this has runtime of  $O(\log(n))$ .
- *Disagreement Seeking (DS):*  $\operatorname{argmax}_{(i,j) \in E^C} (\bar{w} - w_{ij})(z_i - z_j)^2$ .

It takes  $O(n^3)$  time to compute the expressed opinions, and  $O(|E^C|)$  time to check all non-saturated edges.

- *Coordinate Descent (CD):*  $\operatorname{argmax}_{(i,j) \in E^C} -(\bar{w} - w_{ij})$

$\partial_{w_{ij}} P(\mathbf{z})$  Requires  $O(n^3)$  runtime to perform a matrix inversion and multiplication, and  $O(|E^C|)$  to find the maximizing pair.<sup>7</sup>

- *Fiedler Difference (FD):*  $\operatorname{argmax}_{(i,j) \in E^C} (\bar{w} - w_{ij})|v_i - v_j|$ ,

where  $\lambda_2 \mathbf{v} = L\mathbf{v}$

Takes  $O(n^3)$  time to compute the eigendecomposition of  $L$ , and  $O(|E^C|)$  to find the  $\operatorname{argmax}$ .

Notice two effects at play: the maximal weight that can be added ( $\bar{w} - w_{ij}$ ), and some measure of effectiveness per unit weight (disagreement, partial derivative, or absolute difference in Fiedler vector). Naturally, each heuristic attempts to maximize the two's product.

In addition, note that the three non-random heuristics have total runtime of  $O(k(n^3 + |E^C|))$ . The random baseline has runtime of only  $O(k \log(n))$ , but computing polarization at each

<sup>7</sup>Naively, one might think we need  $O(n^3|E^C|)$  time to find the optimum, as we perform a matrix multiplication to compute the gradient of every candidate edge. However, the matrix multiplication is extremely sparse, and can be reduced to operating on four entries of a fixed, pre-computed matrix.



TABLE I  
INITIAL NETWORKS FOR EVALUATION OF POLARIZATION-REDUCING  
HEURISTICS

Network	Vertices $n$	Edges $m$
Twitter	548	3638
Reddit	556	8969
Blogs	1222	16717
Erdős-Rényi	1000	9990
SBM	1000	13726
PA	1000	4883

step (for purposes of comparison) comes with an additional cost of  $O(kn^3)$ . We believe that the random heuristic is useful for two reasons. First, it captures a totally naive recommendation system, which does not curate a user's content exposure based on their opinions. Second, in two of the random graph models we study – Erdős-Rényi and stochastic block – the result of the random heuristic is another graph from the same model, but with slightly higher edge density. Therefore, this heuristic allows us to study how much *additional* polarization is reduced by adding edges in an informed, targeted, manner.

We now study the performance of these heuristics on six unweighted graphs. First we look at three real-world networks – sourced from Twitter, Reddit, and political blogs – and then three synthetic networks with different characteristics: the Erdős-Rényi, stochastic block, and preferential attachment models. Table I provides basic information about the graphs studied. In what follows, the planner's budget is given by  $k = \lfloor \frac{n}{2} \rfloor$ , such that on average each vertex receives one new edge. We plot the value of polarization with the planner's budget, along with the reference point  $P(\mathbf{z}_{K_n})$ , which represents the global minimum of polarization.

Table II shows three quantities: expressed polarization, spectral gap, and assortativity of innate opinions. Expressed polarization is the principal concern of this study, and through Proposition 10 is closely related to the spectral gap. Assortativity is introduced by [36], and captures homophily in a network – which has been shown to control the speed of consensus-forming [37]. In particular, assortativity lies in  $[-1, 1]$ , and measures the correlation of an attribute across edges. In this article, this metric is evaluated for the innate opinions.

In general, the random baseline decreases polarization the least, and both the DS and CD heuristics outperform the Fiedler vector-based strategy. This is expected, as the FD heuristic is blind to the innate opinions, and uses strictly less information. However, we observe that DS and CD tend to result in negative values of homophily, while the FD heuristic does not share this tendency. As an interesting implication, it does not appear that a reduction in polarization requires negative values of homophily. Namely, it may not be necessary to directly connect the most polarized individuals in a society to reduce its level of polarization.

In the figures that follow, vertices are colored according to their innate opinions. Graphs are plotted using the python module `networkx` [38]. Vertices are placed in two-dimensional space using force-directed algorithms, in which vertices repel each other and edges behave like springs in tension. Therefore,

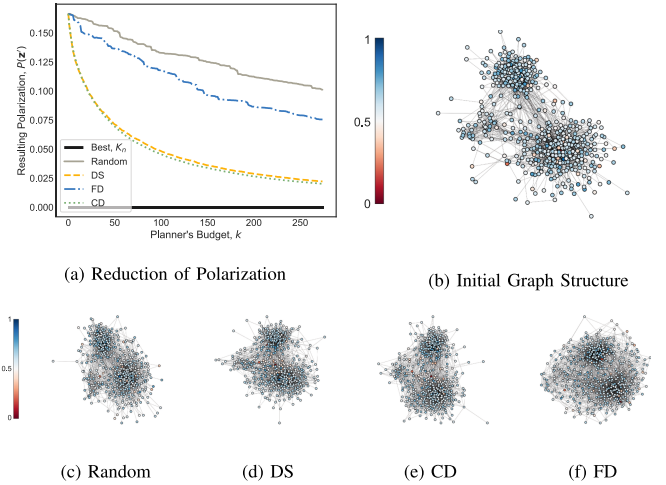


Fig. 2. Evaluation of the planner's heuristics on the Twitter network. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)-(f) visualize the network after the planner has exhausted their budget according to each heuristic.

the vertex layout reflects their relative attraction. The same random seed for initial node placement is used for every graph type studied. All code and data is publicly available here.

#### A. Real-World Networks

The Twitter and Reddit datasets used in this section were first collected by [39], and used by both [14] and [10] in recent work. An additional dataset comprised of political blogs was collected by [40] and used in [7], [31].

*a) Twitter:* This network reflects individuals who tweeted about a Delhi assembly debate in 2013. The network is shown in Fig. 2(b), and mainly consists of two communities.

Fig. 2(a) shows the reduction in polarization achieved by the planner when applying each of the heuristics. Notably, all heuristics outperform our simple baseline. For the two best-performing heuristics, the first 50 edges modified reduce polarization by about a factor of two, and the subsequent 50 achieve a similar fractional reduction. This highlights both the substantial effect of minimal modifications to the graph, and the diminishing returns.

The networks resulting from the planner's heuristics are shown in Fig. 2(c)–(f). There are notable reductions in the strength of community structures. While less effective in reducing polarization, the Fiedler vector-based heuristic (FD) appears to smooth out communities the most.

*b) Reddit:* This network was generated by following Reddit users who posted in a politics forum. Three isolated vertices are removed in preprocessing. Fig. 3(b) shows that the initial network appears to be tightly clustered, and Table II shows that it exhibits an extremely small level of polarization.

For any non-baseline heuristic, the full budget reduces polarization by almost a factor of four. For the best-performing heuristics, this reduction is by nearly an order of magnitude. We observe greatly diminishing returns, with the most significant reduction achieved with the first few edges modified. Moreover, the best-performing heuristics come close to achieving the globally optimal solution with the full budget.



TABLE II

VALUES FOR EXPRESSED POLARIZATION, SPECTRAL GAP, AND INNATE ASSORTATIVITY COMPUTED BEFORE AND AFTER THE PLANNER APPLIES EACH HEURISTIC TO SIX NETWORKS. THE PLANNER ADDS  $k = \lfloor \frac{n}{2} \rfloor$  EDGES – AN AVERAGE OF ONE NEW EDGE PER VERTEX. THE BEST-PERFORMING HEURISTICS ARE HIGHLIGHTED IN BOLD. APPENDIX B CONTAINS ADDITIONAL FIGURES SHOWING CHANGES IN THE SPECTRAL GAP AND ASSORTATIVITY WITH THE PLANNER'S BUDGET

Quantity	Heuristic	Real-World Networks			Synthetic Networks		
		Twitter	Reddit	Blogs	Erdős-Rényi	SBM	PA
Expressed Polarization	Initial	0.1664	0.0053	36.6	0.2422	3.526	1.71
	Random	0.1011	0.0035	22.1	0.2185	2.579	1.35
	DS	0.0221	<b>0.0006</b>	<b>8.2</b>	0.1432	1.774	0.62
	CD	<b>0.0200</b>	<b>0.0006</b>	<b>8.2</b>	<b>0.1420</b>	<b>1.767</b>	<b>0.61</b>
	FD	0.0754	0.0013	15.1	0.2011	1.814	1.23
Spectral Gap	Initial	0.439	0.960	0.169	7.381	4.58	2.85
	Random	0.689	0.977	0.300	8.183	5.51	3.19
	DS	0.792	0.972	1.393	7.439	6.67	3.10
	CD	0.799	2.820	1.258	7.435	6.79	3.21
	FD	<b>2.052</b>	<b>9.170</b>	<b>2.327</b>	<b>12.046</b>	<b>6.87</b>	<b>4.02</b>
Assortativity of Innate Opinions	Initial	0.023	-0.007	0.811	-0.016	0.687	0.025
	Random	0.018	-0.005	0.779	-0.015	0.661	0.029
	DS	-0.143	-0.142	0.747	-0.114	0.606	-0.138
	CD	-0.090	-0.093	0.747	-0.102	0.618	-0.114
	FD	0.029	-0.007	0.780	-0.013	0.635	0.026

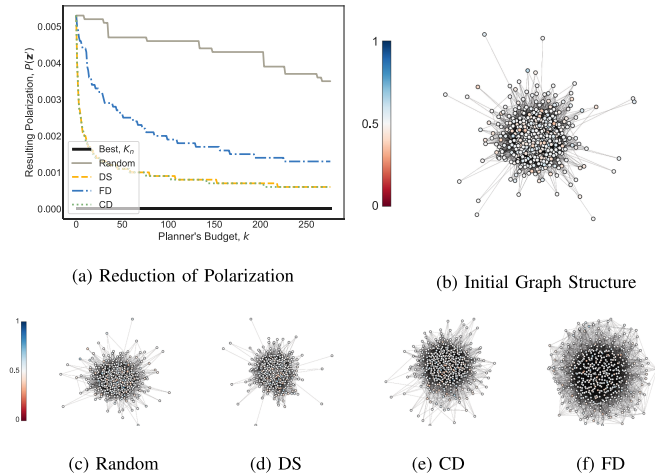


Fig. 3. Evaluation of the planner's heuristics on the Reddit network. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)–(f) visualize the network after the planner has exhausted their budget according to each heuristic.

Only minor changes are observed in the resulting graph structures. Fig. 3(c), (d), and (e) look almost identical to the initial network. In contrast, the graph in Fig. 3(f) does not have as dense a core, and appears to be more evenly connected. Since maximizing the spectral gap results in the graph behaving similarly to an expander, which (informally) is equally well-connected across all cuts, this is to be expected.

*c) Blogs:* This network was collected by aggregating online directories of political blogs around the 2004 US elections. Note that vertices in this network represent blogs – not individuals as in the previous datasets. Each blog was identified as either ‘conservative’ or ‘liberal,’ which we encode by innate opinions of 0 or 1, respectively. Observe in Table II that this network exhibits extremely large values of polarization and homophily, and a small spectral gap.

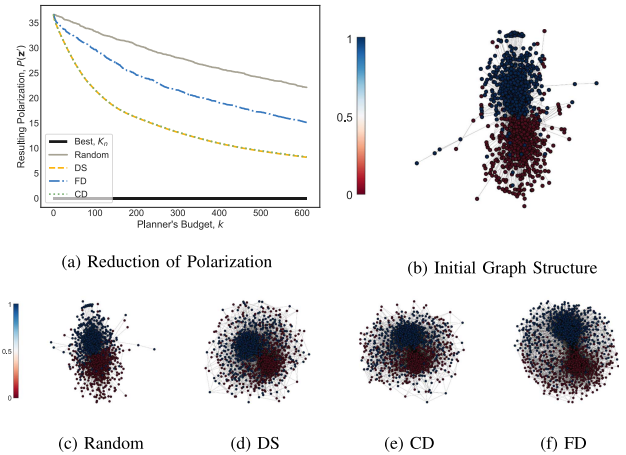


Fig. 4. Evaluation of the planner's heuristics on the political blogs network. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)–(f) visualize the network after the planner has exhausted their budget according to each heuristic.

We find consistent reductions in polarization with all heuristics – including the baseline. This network is unique in that the community structure largely mirrors the innate opinions. That is, the mean-centered innate opinions vector is highly collinear with the Fiedler vector. Hence, both the DS and FD heuristics will choose to add inter-community edges. Since a large fraction of the non-edges span the two communities, a randomly chosen edge is also likely to bridge the two.

Fig. 4(c) shares with Fig. 4(b) a tightly-knit core, with a few vertices at the extremities. In contrast, Figs. 4(d)–(f) depict networks that are more uniformly connected. As before, we find feature this to be most observable with the FD heuristic.

## B. Synthetic Datasets

These heuristics are also applied to three canonical models of random graphs. In the first two models, the number of edges

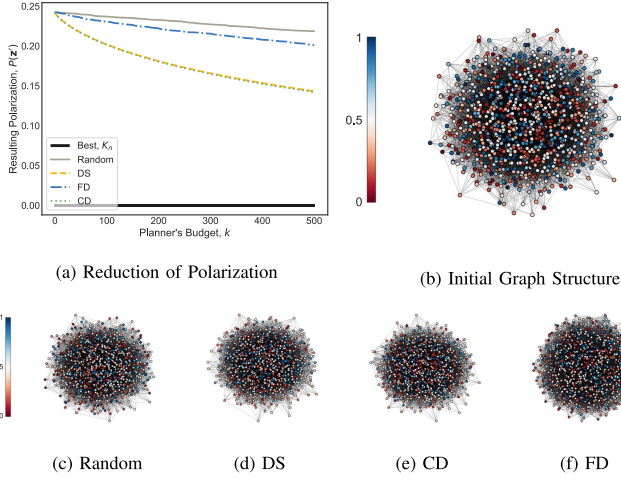


Fig. 5. Evaluation of the planner's heuristics on the Erdős-Rényi graph. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)–(f) visualize the network after the planner has exhausted their budget according to each heuristic.

grows quadratically in the number of vertices (for fixed parameters). However, in our final model, the number of edges is always linear in the number of vertices. Therefore, one may expect that the impact of the planner's  $O(n)$  edges is greatest in the sparser model – but we will see that this is not the case.

*a) Erdős-Rényi:* A graph from this model connects each pair of vertices independently with a fixed probability  $p \in [0, 1]$ . We take  $n = 1000$  and  $p = 0.02$ , although the results are qualitatively similar for different values. The innate opinions are independent uniform random variables in  $[0, 1]$ .

This model produces homogeneous, well-connected networks, which are good spectral approximations of the complete graph  $K_n$  [41]. This can be seen through the large initial spectral gap in Table II. Therefore, according to Proposition 3 it is natural to expect polarization to be small. Nonetheless, all heuristics fail to significantly reduce polarization. A comparison with the random baseline is particularly interesting in this model, as it results in another Erdős-Rényi graph, but with a slightly larger value of  $p$ .

Few changes can be seen among Figs. 5(c)–(f). However, there are few vertices with extreme opinions in the fringes of Fig. 5(d). Instead, these vertices tend to be concentrated in the center of the graph. This aligns with the most negative assortativity seen in Table II. Notably, this is not seen in Fig. 5(f) or (c).

*b) Stochastic Block Model:* A graph drawn from a stochastic block model can replicate community structures, and is shown in Fig. 6(b). This random graph on  $n = 1000$  vertices with two equal-sized communities is generated by mirroring the methodology in [14]. Specifically, the probability of inter-community edges is given by  $q = 0.005$ , and the probability of intra-community edges is  $p = 0.05$ . Since  $p > q$ , we expect to see strong communities. The innate opinions of vertices in each community are drawn independently from either  $\text{Beta}(1, 5)$  or  $\text{Beta}(5, 1)$ , such that the distribution of opinions mirrors the graph's community structure.

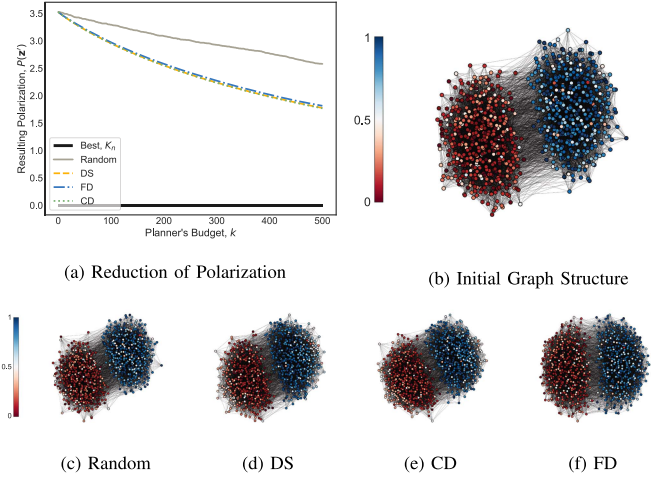


Fig. 6. Evaluation of the planner's heuristics on the stochastic block model graph. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)–(f) visualize the network after the planner has exhausted their budget according to each heuristic.

In Fig. 6(a), a nearly identical reduction in polarization can be seen for all non-random heuristics. This occurs because the mean-centered innate opinions are highly collinear with the Fiedler vector, which partitions the graph into its two communities. Therefore, both the DS and FD strategies will add edges between the two communities. Similarly to the Erdős-Rényi setting, the random baseline yields another stochastic block graph, but with slightly larger parameters  $p$  and  $q$ .

Qualitatively, all three heuristics can be seen to bring the two communities closer together. However, in Fig. 6(d) and (e), the vertices with extreme opinions are brought closer to the center. As before, this is not observed in Fig. 6(f).

*c) Preferential Attachment Model:* This model generates graphs with power-law degree distribution, often known as scale-free or Barabási-Albert networks [42]. Again, we follow a similar procedure to [14], with  $n = 1000$  vertices added sequentially. Each incoming vertex connects to at most  $m = 5$  vertices, where the likelihood of connecting to a particular vertex is proportional to its degree.

This graph tends to exhibit a small, highly interconnected core, and many vertices with low degree. This structure is not conducive to low polarization, as we see in Fig. 7(a). The best-performing heuristics manage to reduce polarization by just over a factor of two, whereas the others see only negligible fractional reductions. Notably, the FD heuristic only slightly outperforms the baseline. These observations are a result of the Friedkin-Johnsen model – higher-degree nodes experience the smallest marginal effects of increased edge weight. Since the preferential attachment model yields a heavy-tailed degree distribution, a larger fraction of nodes are resilient to the planner's modifications. These nodes will also exert large amounts of influence on their neighbors due to their high degree. These structural properties of the preferential attachment graph may dampen the planner's effectiveness.

Qualitatively, Fig. 7(d)–(f) appear similar to the original network in Fig. 7(b). We do not see strong changes in the

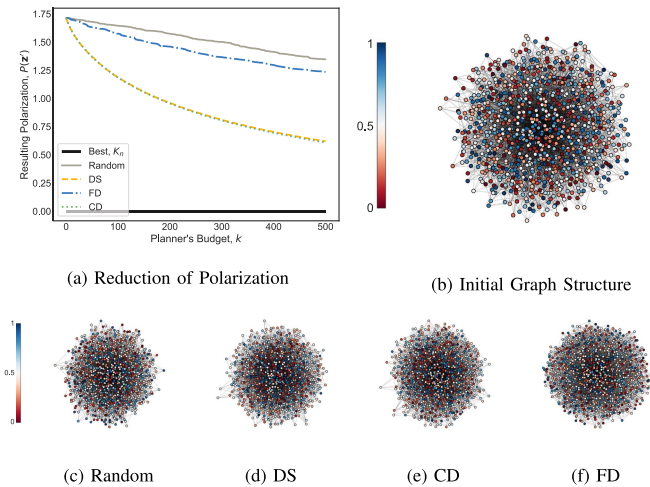


Fig. 7. Evaluation of the planner's heuristics on the preferential attachment graph. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)–(f) visualize the network after the planner has exhausted their budget according to each heuristic.

structure, which aligns with the minor differences in homophily and spectral gap in Table II.

## VI. DISCUSSION AND CONCLUSION

In this article, we analyze the connection between structures of social and information networks and polarization.

First, we establish a relationship between the ratio of *expressed* to *innate* polarization. This ratio is controlled by structural properties of the graph, such as the degree profile and isoperimetric number (i.e. Cheeger constant). In particular, the worst-case polarization depends directly on the spectral gap of the Laplacian. Consequentially, we show that the complete graph achieves the global minimum for polarization. This result aligns with one's intuition – bottlenecks in the graph are liabilities to a consensus.

Next, this article presents two variations of the planner's problem – one in which the innate opinions of the population are known, and another in which they are chosen adversarially. In the first, an expression is derived for the exact difference in polarization when weight is added to a single edge. We find that strengthening the connections between vertices with large *expressed* disagreement reduces polarization. However, it is seen as costly for individuals to interact with differently-minded others [11]. Therefore, reaching a consensus, while arguably beneficial for the population, may prove costly to individuals. We also present a second setting wherein the planner defends the network against adversarially-controlled innate opinions. Here, we prove that the planner aims to maximize the spectral gap. We then show the effectiveness of a strategy based on the Fiedler vector  $\mathbf{v}$  – the eigenvector corresponding to the spectral gap. Intuitively, this vector partitions the graph based on the signs of its elements, and the planner should strengthen edges across the cut.

Finally, we evaluate the performance of four heuristics on several real-world and synthetic networks. We find that all

strategies may smooth out community structures – often referred to as 'echo-chambers'. Furthermore, when there are no strong communities present, the Fiedler vector-based strategy is able to reduce polarization without simultaneously reducing homophily. With this approach, a reduction in polarization did not necessitate direct connections between opposite-minded individuals. However, this strategy performed significantly worse in several networks. We believe that the difference reflects how much of the polarization is driven by the particular values of opinions. For instance, all three heuristics perform similarly when the profile of opinions mirrors the graph structure, and therefore both contribute similarly to the level of polarization. Specifically, all heuristics behave similarly when the mean-centered innate opinions  $\tilde{\mathbf{s}}$  are highly aligned with the Fiedler vector – this observation can be seen most easily in the blogs and stochastic block model networks.

There are several interesting directions for future theoretical work. First, this article has only derived bounds for single-edge modifications. It is an open problem to characterize the effects on polarization of more substantial perturbations to the graph structure. Furthermore, it may be possible to study the planner's effectiveness within various classes of random graph models. For instance, what fraction of non-edges in an Erdős-Rényi graph reduce polarization when added?

At first glance, the results article are severely limited by the model of opinion dynamics. Experimental research has shown that exposure to differing opinions may increase polarization [12]. Motivated by these observations, many models incorporate non-attractive forces between opinions – see [8], [24] for extensions of the FJ model, and [23], [25], [26] for geometrically-inspired approaches. Within the broader problem of understanding how social network structures relate to polarization, this article provides only a first step – the analytical tractability of the FJ model comes at the expense of expressibility. Nonetheless, we believe the results in this article may be generalizable to a wider class of opinion dynamics models that exhibit attraction – which includes all of the above examples. For instance, one could modify a 'disagreement-seeking' heuristic to only consider non-saturated edges between individuals within each others' radius of attraction. The study of polarization-reducing strategies in these more complex models of opinion interaction is a rich and fruitful area for future work.

Several networks showed large reductions in polarization with a small number of edge modifications. However, in the Erdős-Rényi and preferential attachment networks, this article's heuristics did not have as strong of a performance. Beyond our speculation, it remains to be understood what properties of these networks may limit the planner's effectiveness, or what minimal budget is necessary for a fixed fractional reduction in polarization. Moreover, it is not yet clear if this observation is a feature of the heuristics or the graph itself – are we closely approximating the true optimal solution?

In this study, we have shown that strengthening ties between disagreeing individuals is an effective strategy for reducing social polarization. Therefore, if polarization is instead increasing as society becomes increasingly connected, then both



individuals and social media platforms may be failing to contribute to discourse between opposing perspectives.

#### APPENDIX A PROOFS

First, we specify notation. Let  $I$  denote the identity matrix,  $\vec{1}$  the all-ones vector, and  $\mathbf{e}_i$  the  $i$ -th standard basis vector – all of appropriate dimension. Additionally, for  $\mathbf{x} \in \mathbb{R}^n$ , we write  $\bar{x} := \frac{1}{n} \sum_{j=1}^n x_j$  to denote the mean of its entries and  $\tilde{\mathbf{x}} := \mathbf{x} - \bar{x}\vec{1}$  for the mean-centered version of  $\mathbf{x}$ . For a square matrix  $A \in \mathbb{R}^{n \times n}$ , we write  $A_i$  for the  $i$ -th column of  $A$ . The eigenvalues of  $A$  in descending order are given by  $\lambda_n(A) \geq \lambda_{n-1}(A) \geq \dots \geq \lambda_1(A)$ . We frequently use the notation  $\lambda_{\max}(A) = \lambda_n(A)$  and  $\lambda_{\min}(A) = \lambda_1(A)$  for the largest and smallest eigenvalues.

Given an initial graph  $\mathcal{G}$  and any other graph  $\mathcal{G}'$ , define  $T \equiv T(\mathcal{G}; \mathcal{G}') \in \mathbb{R}^{n \times n}$  to be

$$T := (I + L)^{-1}(I + L'), \quad (8)$$

where  $L$  and  $L'$  denote the combinatorial Laplacians of  $\mathcal{G}$  and  $\mathcal{G}'$ , respectively. The dependence of  $T$  on  $\mathcal{G}$  and  $\mathcal{G}'$  will be clear from context and hence omitted. The expressed opinions  $\mathbf{z}'$  can be computed in terms of  $T$  and the original expressed opinions as follows:

$$\mathbf{z}' = T^{-1}\mathbf{z}.$$

This matrix is also useful in allowing us to express the new value of polarization in terms of the expressed opinions on the initial graph. After some algebra, we have that

$$P(\mathbf{z}') = \tilde{\mathbf{z}}^T (T^{-1})^T T^{-1} \tilde{\mathbf{z}}, \quad (9)$$

where we used (8). The spectrum of  $T$  will be critical for theoretical results.

Recall the definition of the isoperimetric number (also known as the Cheeger constant) of a graph from (1). The following simple Lemma is useful in subsequent proofs.

**Lemma 13:** Let  $d_{\max}$  and  $d_{\min}$  denote the maximum and minimum weighted degrees of  $\mathcal{G}$ . Additionally, let  $L$  be the combinatorial Laplacian of  $\mathcal{G}$ , and let  $\lambda_n \geq \lambda_{n-1} \geq \dots \geq \lambda_2 \geq \lambda_1 = 0$  denote its eigenvalues in decreasing order. Then, we have both

$$\frac{1}{2} d_{\min} h_{\mathcal{G}}^2 \leq \lambda_2 \leq 2 d_{\max} h_{\mathcal{G}}, \quad (10)$$

and

$$\lambda_n \leq (2 d_{\max}) \wedge (\bar{w}n). \quad (11)$$

*Proof:* For the normalized Laplacian  $\mathcal{L}$ , the well-known Cheeger inequality (see, e.g., [32]) gives that

$$\frac{h_{\mathcal{G}}^2}{2} \leq \lambda_2(\mathcal{L}) \leq 2 h_{\mathcal{G}}.$$

Notice that the eigenvalues of  $L = D^{1/2} \mathcal{L} D^{1/2}$  are equal to those of  $\mathcal{L}D$ . Additionally, the ordered eigenvalues of  $D$  are simply the degrees of  $\mathcal{G}$  in descending order. Since both  $\mathcal{L}$  and  $D$  are positive semidefinite and Hermitian, we can apply a Weyl multiplicative inequality from [43] for both of

$$\lambda_{i+j-n}(\mathcal{L}D) \leq \lambda_i(\mathcal{L})\lambda_j(D), \quad \text{if } i+j-n \geq 1 \quad (12)$$

$$\lambda_i(\mathcal{L})\lambda_j(D) \leq \lambda_{i+j-1}(\mathcal{L}D), \quad \text{if } i+j-1 \leq n. \quad (13)$$

Choosing  $i = 2$  and  $j = n$  in (12) gives that

$$\lambda_2(\mathcal{L}D) \leq \lambda_2(\mathcal{L})d_{\max} \leq 2 h_{\mathcal{G}} d_{\max}.$$

With  $i = 2$  and  $j = 1$  in (13), we have that

$$\lambda_2(\mathcal{L}D) \geq \lambda_2(\mathcal{L})d_{\min} \geq \frac{1}{2} h_{\mathcal{G}}^2 d_{\min}.$$

Combining the previous two displays gives (10).

The inequality (11) can be proved using the triangle inequality. The largest eigenvalue of  $\mathcal{L}$  is equal to the operator norm of  $D - A$ . Since the norm of both  $D$  and  $A$  are upper bounded by  $d_{\max}$ , we conclude that  $\lambda_n(L) \leq 2d_{\max}$ .

Let  $L_{K_n} = \bar{w}(nI - \vec{1}\vec{1}^T)$  denote the combinatorial Laplacian of the complete graph, where all edge weights are equal to  $\bar{w}$ . Since for any  $L$  we have  $L_{K_n} \succcurlyeq L$ , then  $\bar{w}n = \lambda_n(L_{K_n}) \geq \lambda_n(L)$ . ■

#### A. Proofs of Section IV-A

*Proof of Proposition 3:* We seek to write  $P(\mathbf{z})$  in a way that  $P(\mathbf{s})$  appears. Recall that  $\mathbf{z} = (I + L)^{-1}\mathbf{s}$  and also  $\tilde{\mathbf{z}} = (I + L)^{-1}\tilde{\mathbf{s}}$ . Therefore

$$P(\mathbf{z}) = \tilde{\mathbf{z}}^T \tilde{\mathbf{z}} = \tilde{\mathbf{s}}^T (I + L)^{-2} \tilde{\mathbf{s}}. \quad (14)$$

Towards the lower bound in the claim, we may use an eigenvalue bound to obtain that

$$P(\mathbf{z}) \geq \lambda_{\min}((I + L)^{-2}) \tilde{\mathbf{s}}^T \tilde{\mathbf{s}} = (1 + \lambda_{\max}(L))^{-2} P(\mathbf{s}).$$

From (11) we have that  $\lambda_{\max}(L) \leq (2d_{\max}) \wedge \bar{w}n$ ; plugging this into the display above we obtain the claimed lower bound.

For the upper bound, first note that the eigenvector corresponding to the largest eigenvalue of  $(I + L)^{-2}$  is  $\vec{1}$ . Since  $\tilde{\mathbf{s}}$  is orthogonal to  $\vec{1}$ , we have from (14) that

$$P(\mathbf{z}) \leq \lambda_{n-1}((I + L)^{-2}) \tilde{\mathbf{s}}^T \tilde{\mathbf{s}} = (1 + \lambda_2(L))^{-2} P(\mathbf{s}).$$

From (10) we have that  $\lambda_2(L) \geq (1/2)d_{\min}h_{\mathcal{G}}^2$ ; plugging this into the display above we obtain the desired upper bound. ■

*Proof of Corollary 4:* Take any graph  $\mathcal{G}$  and innate opinions  $\mathbf{s}$ . Proposition 3 implies that

$$P(\mathbf{z}_{\mathcal{G}}) \geq P(\mathbf{s})(1 + (2d_{\max}) \wedge (\bar{w}n))^{-2} \geq P(\mathbf{s})(1 + \bar{w}n)^{-2}.$$

Turning to the complete graph  $K_n$ , recall that the spectrum of its Laplacian has 0 as an eigenvalue with eigenvector  $\vec{1}$ . It also has eigenvalue  $\bar{w}n$  with multiplicity  $n - 1$  and eigenspace containing all vectors orthogonal to  $\vec{1}$ . Since  $\tilde{\mathbf{s}}^T \vec{1} = 0$ , we have  $(I + L_{K_n})^{-1} \tilde{\mathbf{s}} = (1 + \bar{w}n)^{-1} \tilde{\mathbf{s}}$ . Recalling the definition of polarization, we obtain  $P(\mathbf{z}_{K_n}) = \|(I + L_{K_n})^{-1} \tilde{\mathbf{s}}\|^2 = (1 + \bar{w}n)^{-2} \|\tilde{\mathbf{s}}\|^2 = (1 + \bar{w}n)^{-2} P(\mathbf{s})$ . Comparing with the display above, we see that  $K_n$  minimizes polarization over all graphs with maximal weight  $\bar{w}$ . ■

#### B. Proofs of Section IV-B

*Proof of Lemma 5:* To obtain the claim, we expand  $P(\mathbf{z}^+)$  in a way that  $P(\mathbf{z})$  appears. First, note that  $L^+ = L + \delta L_{ij}$  and  $L_{ij} = \mathbf{v}_{ij} \mathbf{v}_{ij}^T$ , and hence we have that  $T = I + \delta(I + L)^{-1} \mathbf{v}_{ij} \mathbf{v}_{ij}^T$ . The Sherman-Morrison formula thus gives that

$T^{-1} = I - \frac{\delta}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} (I+L)^{-1} \mathbf{v}_{ij} \mathbf{v}_{ij}^T$ . Plugging this into (9), we obtain

$$\begin{aligned} P(\mathbf{z}^+) &= \left\| \left( I - \frac{\delta}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} (I+L)^{-1} \mathbf{v}_{ij} \mathbf{v}_{ij}^T \right) \tilde{\mathbf{z}} \right\|^2 \\ &= \tilde{\mathbf{z}}^T \tilde{\mathbf{z}} - \frac{2\delta \tilde{\mathbf{z}}^T (I+L)^{-1} \mathbf{v}_{ij} \mathbf{v}_{ij}^T \tilde{\mathbf{z}}}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} \\ &\quad + \frac{(\tilde{\mathbf{z}}^T \mathbf{v}_{ij})^2}{(1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij})^2}. \end{aligned}$$

Noting that  $P(\mathbf{z}) = \tilde{\mathbf{z}}^T \tilde{\mathbf{z}}$ , and  $D_{ij}(\mathbf{z}) = (\tilde{\mathbf{z}}^T \mathbf{v}_{ij})^2$  leads to the desired expression after rearranging. ■

*Proof of Proposition 7:* For simplicity of notation, let  $A = I + L$ . Then, for any  $t > 0$ , we have

$$\frac{P(\mathbf{z}_{L+tL_{ij}}) - P(\mathbf{z}_L)}{t} = \frac{\tilde{\mathbf{s}}^T \left[ (A + t\mathbf{v}_{ij} \mathbf{v}_{ij}^T)^{-2} - A^{-2} \right] \tilde{\mathbf{s}}}{t}$$

Using the Sherman-Morrison formula, we can compute that

$$\begin{aligned} \left[ (A + t\mathbf{v}_{ij} \mathbf{v}_{ij}^T)^{-1} \right]^2 &= \left[ A^{-1} - \frac{tA^{-1} \mathbf{v}_{ij} \mathbf{v}_{ij}^T A^{-1}}{1 + t\mathbf{v}_{ij}^T A^{-1} \mathbf{v}_{ij}} \right]^2 \\ &= A^{-2} - 2t \frac{A^{-2} \mathbf{v}_{ij} \mathbf{v}_{ij}^T A^{-1}}{1 + t\mathbf{v}_{ij}^T A^{-1} \mathbf{v}_{ij}} + o(t) \end{aligned} \quad (15)$$

where  $\frac{o(t)}{t} = o(1) \rightarrow_{t \rightarrow 0} 0$ . Plugging (15) into (4) and taking the limit concludes. ■

*Proof of Corollary 8:* Recall that  $\mathbf{v}_{ij} = \mathbf{e}_i - \mathbf{e}_j$ . Since  $N(i) = N(j)$ , a direct computation gives  $L\mathbf{v}_{ij} = (d_i - w_{ij})\mathbf{v}_{ij}$ . Consequently, we have  $(I+L)^{-1}\mathbf{v}_{ij} = \frac{1}{1+d_i-w_{ij}}\mathbf{v}_{ij}$ . Plugging this into Lemma 5 and simplifying yields the desired result. ■

*Proof of Theorem 9:* The proof of this Theorem follows from bounding the terms in Lemma 5.

First, we show the upper bound. Notice that  $\frac{\delta^2 \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} \geq 0$ , so this term can be dropped. Through an eigenvalue bound we also find that

$$\frac{1}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} \leq \frac{1 + \lambda_n(L)}{1 + 2\delta + \lambda_n(L)}.$$

Plugging these two observations into (3) and rearranging to find  $\partial_{w_{ij}} P(L)$  concludes.

For the lower bound, we have the following inequalities.

$$\begin{aligned} \frac{\delta \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} &\leq \frac{\delta \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}} \\ &\leq \frac{2\delta}{2\delta + (1 + \lambda_2(L))^2}. \end{aligned}$$

Therefore, by assumption and Lemma 5, we have that

$$P(\mathbf{z}) - P(\mathbf{z}^+) \leq \frac{\delta(z_i - z_j)^2}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} [2\epsilon] \leq \frac{2\delta\epsilon(z_i - z_j)^2}{1 + 2\delta},$$

where we used  $\mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij} \leq 2$ . ■

### C. Proofs of Section IV-C

*Proof of Proposition 10:* This proof requires only that we solve explicitly the adversary's optimization problem.

By construction,  $\tilde{\mathbf{s}}$  is orthogonal to  $\tilde{\mathbf{1}}$ . As a result, the optimal solution for the adversary is  $\sqrt{R}\mathbf{v}_2$ , where  $\mathbf{v}_2$  is the eigenvector corresponding to  $\lambda_2(L')$ , and the optimal value is:

$$\max_{\mathbf{s} \in \mathbb{R}^n: \|\mathbf{s}\|_2 \leq R} \tilde{\mathbf{s}}^T (I+L')^{-2} \tilde{\mathbf{s}} = \frac{R}{(1 + \lambda_2(L'))^2}.$$

To minimize this quantity, it follows that the planner maximizes the spectral gap of  $L'$ . ■

*Proof of Theorem 11:* This proof uses a variation of a result by [35]. The original result states that if a simple (unweighted, undirected) graph  $\mathcal{G}_s^+$  is constructed by adding a non-edge  $(i, j)$  to another simple graph  $\mathcal{G}_s$ , we have

$$\lambda_2(L_s^+) \geq \min \left\{ \lambda_2(L_s) + \frac{\epsilon\alpha^2}{\epsilon + (2 - \alpha^2)}, \lambda_3(L_s) - \epsilon \right\}$$

$$\lambda_2(L_s^+) \leq \min \{ \lambda_2(L_s) + \alpha^2, \lambda_3(L_s) \},$$

where  $\alpha^2 = (v_i - v_j)^2$ , and  $\mathbf{v}$  is the eigenvector of  $L$  corresponding to  $\lambda_2(L)$ .

It is possible to adapt the original proof to consider the case where weight  $\delta$  is added to edge  $(i, j)$ . This result would yield:

$$\lambda_2(L^+) \geq \min \left\{ \lambda_2(L) + \frac{\epsilon\delta\alpha^2}{\epsilon + \delta(2 - \alpha^2)}, \lambda_3(L) - \epsilon \right\}$$

$$\lambda_2(L^+) \leq \min \{ \lambda_2(L) + \delta\alpha^2, \lambda_3(L) \},$$

The tightest lower bound is achieved by choosing

$$\epsilon^* = \frac{\beta - 2\delta}{2} + \left( \left( \frac{\beta - 2\delta}{2} \right)^2 + \beta\delta(2 - \alpha^2) \right)^{1/2},$$

where  $\beta = \lambda_3(L) - \lambda_2(L)$ , so that both terms in the minimum are equal. First, we note that  $\epsilon^* \geq \beta - 2\delta$ , with equality when  $\alpha = 2$ . Additionally, the first term in the minimum is increasing in  $\epsilon$ , so therefore we have

$$\frac{\beta - 2\delta}{\beta} \delta \alpha^2 \leq \frac{(\beta - 2\delta)\delta\alpha^2}{\beta - 2\delta + \delta(2 - \alpha^2)} \leq \lambda_2(L^+) - \lambda_2(L) \leq \alpha^2,$$

as claimed since  $\alpha^2 \geq 0$ . Finally, note that  $\lambda_2(L^+) \geq \lambda_2(L)$  as  $L^+ - L = \delta L_{ij} \succcurlyeq 0$ . ■

*Proof of Corollary 12:* Recall that we defined  $P(L) = \frac{R}{(1 + \lambda_2(L))^2}$ . We first prove the upper bound by using (7):

$$\begin{aligned} \frac{P(L) - P(L^+)}{R} &\leq \frac{1}{(1 + \lambda_2(L))^2} - \frac{1}{(1 + \lambda_2(L) + \delta\alpha^2)^2} \\ &\leq \frac{\delta^2\alpha^4 + 2(1 + \lambda_2(L))\delta\alpha^2}{(1 + \lambda_2(L))^2 (1 + \lambda_2(L) + \delta\alpha^2)^2}. \end{aligned}$$

Since  $\delta\alpha^2 \geq 0$  and  $\alpha^2 \leq 2 \leq 2(1 + \lambda_2(L))$ , we write  $\alpha^4 \leq 2(1 + \lambda_2(L))\alpha^2$ , and arrive at the desired

$$\frac{1}{(1 + \lambda_2(L))^2} - \frac{1}{(1 + \lambda_2(L^+))^2} \leq \frac{4\alpha^2 (\delta \vee \delta^2)}{(1 + \lambda_2(L))^3}.$$

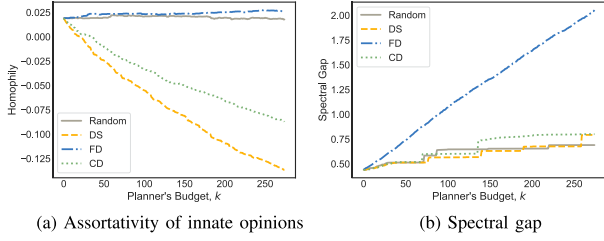


Fig. 8. The planner's effect on the Twitter network.

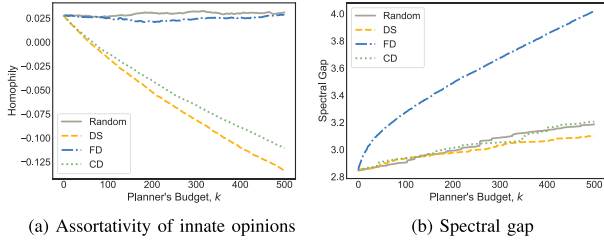


Fig. 13. The planner's effect on the preferential attachment graph.

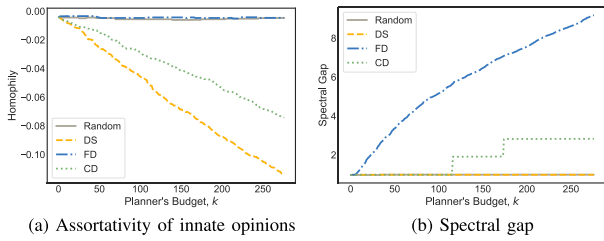


Fig. 9. The planner's effect on the Reddit network.

The lower bound follows similarly by (10). For simplicity of notation, let  $c = \max\{1 - \frac{2\delta}{\beta}, 0\}$ . Then,

$$\begin{aligned} \frac{P(L) - P(L^+)}{R} &\geq \frac{1}{(1 + \lambda_2(L))^2} - \frac{1}{(1 + \lambda_2(L) + c\delta\alpha^2)^2} \\ &\geq \frac{c^2\delta^2\alpha^4 + 2c\delta\alpha^2(1 + \lambda_2(L))}{(1 + \lambda_2(L))^2(1 + \lambda_2(L) + c\delta\alpha^2)^2}. \end{aligned}$$

Observe that  $c^2\delta^2\alpha^4 \geq 0$ , so this term can be dropped. Furthermore,  $c\delta\alpha^2 \leq 2\delta$ , which gives us the result:

$$\frac{P(L) - P(L^+)}{R} \geq \frac{2c\delta\alpha^2}{(1 + 2\delta + \lambda_2(L))^3},$$

■

## APPENDIX B ADDITIONAL FIGURES

In this short section, Figs. 8–13 show how homophily (i.e., assortativity of innate opinions) and the spectral gap are affected by the planner's modifications. These provide greater detail than the initial and final values found in Table II.

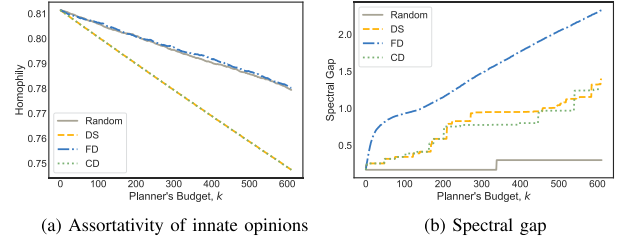


Fig. 10. The planner's effect on the political blogs network.

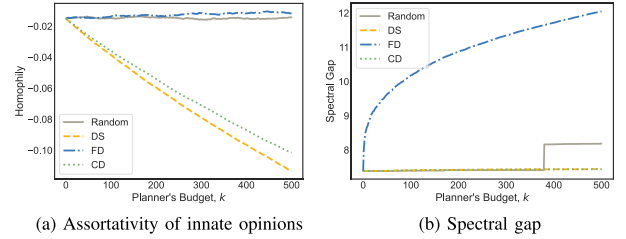


Fig. 11. The planner's effect on the Erdős-Rényi graph.

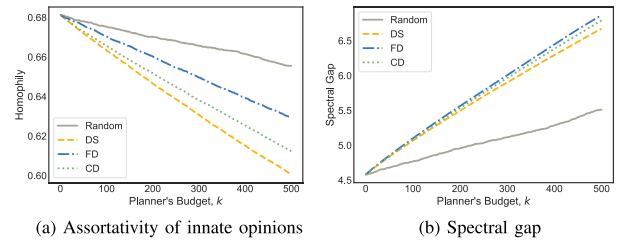


Fig. 12. The planner's effect on the stochastic block graph.

## REFERENCES

- [1] U. Chitra and C. Musco, "Analyzing the impact of filter bubbles on social network polarization," in *Proc. 13th Int. Conf. Web Search Data Mining*, 2020, pp. 115–123.
- [2] N. Bhalla, A. Lechowicz, and C. Musco, "Local edge dynamics and opinion polarization," in *Proc. 16th Conf. Web Search Data Mining*, 2023, pp. 6–14.
- [3] F. P. Santos, Y. Lelkes, and S. A. Levin, "Link recommendation algorithms and dynamics of polarization in online social networks," *Proc. Nat. Acad. Sci.*, vol. 118, no. 50, 2021, Art. no. e2102141118.
- [4] X. Chen, J. Lijffijt, and T. D. Bie, "Quantifying and minimizing risk of conflict in social networks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1197–1205.
- [5] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Reducing controversy by connecting opposing views," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 81–90.
- [6] S. Haddadan, C. Menghini, M. Riondato, and E. Upfal, "Republik: Reducing polarized bubble radius with link insertions," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, 2021, pp. 139–147.
- [7] A. Matakos, E. Terzi, and P. Tsaparas, "Measuring and moderating opinion polarization in social networks," *Data Mining Knowl. Discov.*, vol. 31, no. 5, pp. 1480–1505, 2017.
- [8] I. Rahaman and P. Hosein, "A model for optimizing article recommendation for reducing polarization," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2021, pp. 107–110.
- [9] L. Zhu, Q. Bao, and Z. Zhang, "Minimizing polarization and disagreement in social networks via link recommendation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 2072–2084.



- [10] C. Musco, C. Musco, and C. E. Tsourakakis, "Minimizing polarization and disagreement in social networks," in *Proc. World Wide Web Conf.*, 2018, pp. 369–378.
- [11] D. Bindel, J. Kleinberg, and S. Oren, "How bad is forming your own opinion?," *Games Econ. Behav.*, vol. 92, pp. 248–265, 2015.
- [12] C. A. Bail et al., "Exposure to opposing views on social media can increase political polarization," *Proc. Nat. Acad. Sci.*, vol. 115, no. 37, pp. 9216–9221, 2018.
- [13] S. Ballelli, L. Getoor, D. G. Goldstein, and D. J. Watts, "Reducing opinion polarization: Effects of exposure to similar people with differing political views," *Proc. Nat. Acad. Sci.*, vol. 118, no. 52, 2021, Art. no. e2112552118.
- [14] M. F. Chen and M. Z. Rácz, "An adversarial model of network disruption: Maximizing disagreement and polarization in social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 728–739, Mar./Apr. 2022.
- [15] J. Gaitonde, J. Kleinberg, and E. Tardos, "Adversarial perturbations of opinion dynamics in networks," in *Proc. 21st ACM Conf. Econ. Computation*, 2020, pp. 471–472.
- [16] L. Donetti, F. Neri, and M. A. Muñoz, "Optimal network topologies: Expanders, cages, Ramanujan graphs, entangled networks and all that," *J. Stat. Mechanics: Theory Experiment*, vol. 2006, no. 8, 2006, Art. no. P08007.
- [17] H. Wang and P. V. Mieghem, "Algebraic connectivity optimization via link addition," in *Proc. 3D Int. ICST Conf. Bio-Inspired Models Netw., Inf., Comput. Syst.*, 2010.
- [18] T. Watanabe and N. Masuda, "Enhancing the spectral gap of networks by node removal," *Phys. Rev. E - Stat., Nonlinear, Soft Matter Phys.*, vol. 82, no. 4, pp. 1–7, 2010.
- [19] M. H. DeGroot, "Reaching a consensus," *J. Amer. Stat. Assoc.*, vol. 69, no. 345, pp. 118–121, 1974.
- [20] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *J. Math. Sociol.*, vol. 15, no. 3/4, pp. 193–206, 1990.
- [21] I. Conjeaud, P. Lorenz-Spreen, and A. Kalogeratos, "Degroot-based opinion formation under a global steering mechanism," 2022, *arXiv:2210.12274*.
- [22] E. Biondi, C. Boldrini, A. Passarella, and M. Conti, "Dynamics of opinion polarization," 2022, *arXiv:2206.06134*.
- [23] R. Hegselmann et al., "Opinion dynamics and bounded confidence models, analysis, and simulation," *J. Artif. Societies Social Simul.*, vol. 5, no. 3, 2002.
- [24] E. Cornacchia, N. Singer, and E. Abbe, "Polarization in attraction-repulsion models," in *Proc. IEEE Int. Symp. Inf. Theory* 2020, pp. 2765–2770.
- [25] J. Hazla, Y. Jin, E. Mossel, and G. Ramnarayan, "A geometric model of opinion polarization," *Math. Oper. Res.*, 2023.
- [26] J. Gaitonde, J. Kleinberg, and E. Tardos, "Polarization in geometric opinion dynamics," in *Proc. 22nd ACM Conf. Econ. Computation*, 2021, pp. 499–519.
- [27] C. Altafini, "Consensus problems on networks with antagonistic interactions," *IEEE Trans. Autom. Control*, vol. 58, no. 4, pp. 935–946, Apr. 2013.
- [28] Y. Mao, S. Bolouki, and E. Akyol, "Spread of information with confirmation bias in cyber-social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 688–700, Apr./Jun. 2020.
- [29] J. Qin, Q. Ma, Y. Shi, and L. Wang, "Recent advances in consensus of multi-agent systems: A brief survey," *IEEE Trans. Ind. Electron.*, vol. 64, no. 6, pp. 4972–4983, Jun. 2017.
- [30] A. Gionis, E. Terzi, and P. Tsaparas, "Opinion maximization in social networks," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 387–395.
- [31] A. Matakos, S. Tu, and A. Gionis, "Tell me something my friends do not know: Diversity maximization in social networks," *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3697–3726, 2020.
- [32] F. R. Chung, *Spectral Graph Theory*, no. 92. Providence, RI, USA: Amer. Math. Soc., 1997.
- [33] I. Gilboa and D. Schmeidler, "Maxmin expected utility with a non-unique prior," *J. Math. Econ.*, vol. 18, no. 2, pp. 141–153, 1989.
- [34] R. Eldan, M. Z. Rácz, and T. Schramm, "Braess's paradox for the spectral gap in random graphs and delocalization of eigenvectors," *Random Struct. Algorithms*, vol. 50, no. 4, pp. 584–611, 2017.
- [35] C. Maas, "Transportation in graphs and the admittance spectrum," *Discrete Appl. Math.*, vol. 16, no. 1, pp. 31–49, 1987.
- [36] M. E. Newman, "Mixing patterns in networks," *Phys. Rev. E*, vol. 67, no. 2, 2003, Art. no. 026126.
- [37] B. Golub and M. O. Jackson, "How homophily affects the speed of learning and best-response dynamics," *Quart. J. Econ.*, vol. 127, no. 3, pp. 1287–1338, 2012.
- [38] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11–15.
- [39] A. De, S. Bhattacharya, P. Bhattacharya, N. Ganguly, and S. Chakrabarti, "Learning a linear influence model from transient opinion dynamics," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 401–410.
- [40] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: Divided they blog," in *Proc. 3rd Int. Workshop Link Discov.*, 2005, pp. 36–43.
- [41] C. Hoffman, M. Kahle, and E. Paquette, "Spectral gaps of random graphs and applications," *Int. Math. Res. Notices*, vol. 2021, no. 11, pp. 8353–8404, 2021.
- [42] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [43] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1994.



Princeton's Howard B. Wentz, Jr. Junior Faculty Award, a Princeton SEAS Innovation Award, and an Excellence in Teaching Award.



**Miklos Z. Rácz** received the Ph.D. degree in statistics from the University of California, Berkeley, Berkeley, CA, USA. He is currently an assistant professor with ORFE Department, Princeton University, Princeton, NJ, USA, and an Associated Faculty Member with the Center for Statistics and Machine Learning. Before coming to Princeton, he was a Postdoc with the Theory Group, Microsoft Research, Redmond. His research interests include interface of probability, statistics, computer science, and information theory. His research and teaching has been recognized by Princeton's Howard B. Wentz, Jr. Junior Faculty Award, a Princeton SEAS Innovation Award, and an Excellence in Teaching Award.

**Daniel E. Rigobon** received the B.S. degree in mechanical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2018. He is currently working toward the Ph.D. degree with ORFE Department, Princeton University, Princeton, NJ, USA. His research interests include the assessment and prevention of fragility in networks, algorithmic fairness in learning, and more generally, computational social science. In 2018, he was the recipient of the John C. and Elizabeth J. Chato Award from MIT, and the Presidents' Fellowship from Princeton.