Salloum, Ali; Chen, Ted Hsuan Yun; Kivelä, Mikko

Separating Polarization from Noise

Separating Polarization from Noise: Comparison and Normalization of Structural Polarization Measures

ALI SALLOUM, Aalto University, Finland

TED HSUAN YUN CHEN, Aalto University and University of Helsinki, Finland

MIKKO KIVELÄ, Aalto University, Finland

Quantifying the amount of polarization is crucial for understanding and studying political polarization in political and social systems. Several methods are used commonly to measure polarization in social networks by purely inspecting their structure. We analyse eight of such methods and show that all of them yield high polarization scores even for random networks with similar density and degree distributions to typical real-world networks. Further, some of the methods are sensitive to degree distributions and relative sizes of the polarized groups. We propose normalization to the existing scores and a minimal set of tests that a score should pass in order for it to be suitable for separating polarized networks from random noise. The performance of the scores increased by 38%-220% after normalization in a classification task of 203 networks. Further, we find that the choice of method is not as important as normalization, after which most of the methods have better performance than the best-performing method before normalization. This work opens up the possibility to critically assess and compare the features and performance of different methods for measuring structural polarization.

CCS Concepts: • **Human-centered computing → Collaborative and social computing**; **Collaborative and social computing design and evaluation methods**; **Social network analysis**; • **Applied computing → *Sociology*.**

Additional Key Words and Phrases: network science; polarization; political polarization; computational social science; normalization; twitter; networks; sociology; community detection; clustering; statistical significance

1 INTRODUCTION

Political polarization has long been a question in the social sciences [7, 26], with mainstream interest growing following observed political divisions in the 2000 and 2004 US Presidential elections [26]. Polarization, which is generally understood in the social science literature as the division of individuals into coherent and strongly opposed groups based on their opinion of one or more issues [21, 26], has deleterious consequences for social systems. These undesirable outcomes include increased divisiveness and animosity [51], policy gridlock [42], and even decreased political representation [7]. Recent sociopolitical challenges have further strengthened academic interest in this area, as political polarization has been associated with difficulties in resolving pressing societal issues such as climate change [79], immigration and race relations [39], and recently the COVID-19

Authors' addresses: Ali Salloum, ali.salloum@aalto.fi, Aalto University, Konemiehentie 2, 02150, Espoo, Finland; Ted Hsuan Yun Chen, ted.hsuanyun.chen@gmail.com, Aalto University and University of Helsinki, Unioninkatu 37, 00170, Helsinki, Finland; Mikko Kivelä, mikko.kivela@aalto.fi, Aalto University, Konemiehentie 2, 02150, Espoo, Finland.

pandemic [49]. In the context of social computing and computer mediated communication, political polarization on social media has been shown to constrain communication and eases the spread of misinformation [40].

Extensive research efforts have been put toward mitigating political polarization in computer mediated communication. This body of work ranges from studies exploring the role of polarization in social media [11, 14, 15, 18, 20, 64, 66] to platform design aimed at attenuating polarization [5, 60, 61]. A fundamental requirement that underlies these efforts is that we are able to identify polarized topics and systems, and monitor changes over time. *How will we know where intervention is required? And how will we know if our interventions are successful?* Much research has been done in the area of measuring political polarization. Traditional methods primarily relied on survey-based approaches, which tend to measure distributional properties of responses to questions on public opinion surveys, such as bimodality and dispersion [21]. With the increasing availability and richness of publicly-observable social data, recent work from the computational social science and network science fields have shifted polarization measurement in two new directions [32]. The first area of work are content-based approaches [e.g. 10, 20], which have become widely used following developments in natural language processing tools that allow researchers to detect antagonistic positions between groups on social media.

Another fruitful area of work focuses on structural aspects of polarization inferred from network representations of social or political systems. These structural polarization measures tend to be based on the logic of identifying what would be observable features of systems that are characterized by polarization. Because polarization is a system-level phenomenon, these features are defined at the network-level, making them different from node-level (i.e. individual) behavioral mechanisms. While node-level mechanisms such as homophily can contribute to polarization [9, 41], they do not necessitate it, as individuals will at any time exhibit a multitude of behavioral tendencies. Most importantly, following the definition of polarization outlined above, structural measures generally take separation between structurally-identified communities to represent polarization between groups in the system. Additionally, different measures tend to be motivated by additional aspects of political polarization, such as, for example, the difficulty of information to escape echo chambers [e.g. 32]. Because these structural polarization measures can flexibly capture theoretically-grounded aspects of political polarization, especially at a relatively low cost compared to content-based and survey-based approaches, they appear to be attractive measures for applied work.

Indeed, applications of structural polarization scores to real networks have spanned many domains, including the studies of political party dynamics [6, 45, 57, 59, 75], cohesion and voting behavior [3, 12, 17, 75], political events [15, 18, 58, 67, 76], and online echo chambers [15, 16, 31]. They have also been used to detect and study the presence of 'controversy' in communication across different topics [32]—in fact, some structural polarization scores are named as controversy scores—as polarized groupings can be interpreted as a manifestation of a controversial issue. Despite their widespread application, there are few systematic assessments of how well these structural polarization measures perform in key aspects of measurement validity [74], such as predicting whether a network is polarized based on human labeling, and whether they are invariant to basic network statistics such as size and densiy. Exceptions to this include a small number of studies that compare the performance of some scores in classifying polarized political systems [23, 32] or those that show certain scores are invariant to network size [64]. On the whole, the body of evidence is sparse.

Further, beyond simply the lack of evidence, there are in fact theoretical reasons to expect systematic shortcomings with the current approach. Consider that the typical method for measuring structural polarization relies on the joint tasks of (1) graph clustering to find two groups, and then (2) measuring how separate the two groups are. A key characteristic of this approach is that most

clustering methods start from the assumption that the networks have separate groups, and because these clustering algorithms are optimized to do so, they can find convincing group structures even in completely random networks [36, 47, 77]. Moreover, the quality function of these methods can be sensitive to elementary features such as the size and density of these networks [36, 77]. Given such a starting point, it is difficult to develop a measure that would separate non-polarized networks from polarized ones. Scores based on ostensibly intuitive community structures are potentially poor measures of polarization because they yield high polarization scores for non-polarized networks.

We address these concerns by analyzing eight different scores that have been proposed and used for quantifying structural polarization. We test how well they separate random networks from polarized networks and how sensitive they are to various network features: number of nodes, average degree, degree distribution, and relative size of the two groups. Further, we compute these polarization values for a large array of political networks with varying degrees of polarization, and use null models to see how much of the score values are explained by these different network features. We find that none of the measures tested here are able to systematically produce zero scores (i.e., low or non-polarization) for random networks without any planted group structure, which we take to be a reasonable requirement for a polarization measure [35]. Further, they are sensitive to variations in basic network features, with the average degree being particularly challenging for these measures.

Our study makes a number of contributions to the literature. First, our results indicate that it is difficult to interpret any of the polarization scores in isolation. Given a score value it can be impossible to tell even whether the network is more polarized than a random network unless the score is reflected against the network features. Further, analysing the extent to which the network is polarized or making comparisons claiming that one network is more polarized than another one is not straightforward with the current measures. Second, we present a straightforward solution to the problem. We show that normalizing polarization scores against a distribution of scores from their corresponding configuration models improves their performance in a classification task of 203 labeled networks. Finally, we make our testing code and data publicly available on GitHub [70] and Zenodo [71].

The rest of this paper is organized as follows. We first briefly discuss prior research and describe the current methods for measuring polarization in networks. Next, we introduce the methods and data used in this study. We then study the performance of modern polarization scores both on real networks and synthetic networks. Finally, we present normalization to the existing scores. We conclude by summarizing our results in the discussion section.

2 NETWORK-BASED POLARIZATION SCORES

Fig. 1. The polarization measurement pipeline

Polarization can be understood as the division of individuals into coherent and strongly opposed groups based on their opinion of one or more issues [21, 26, 52]. Polarization reflects strongly on social network structures by creating two internally well-connected groups that are sparsely connected to each other [9], which means that the amount of polarization in a system can be measured by observing the structure of interactions within that system. This logic informs many structural polarization scores, which are obtained following the procedure shown in Fig. 1: (1)

social network construction, (2) finding the two groups via graph partitioning, and (3) quantifying the separation between the two groups.

When measuring structural polarization in a social system, constructing a network representation of the system that allows us to identify the existence of polarized group structures apart from other types of interactions is a crucial step. It defines the appropriate bounds of the system, both in terms of which units are included and which interactions are measured. A social network is suitable for polarization analysis with the methods described here if the link between two nodes indicates a positive relationship, such as friendship, preference similarity, endorsement, or agreement [2, 14, 30, 32]. Bounding the network to the appropriate set of interactions is key, as it has been observed that simultaneously measuring unrelated topics will yield lower structural polarization scores regardless of how polarized each of the topics are [12].

Network-based polarization scores share the requirement that the network needs to be partitioned, with the logic being that the partitioned units are somehow closer to others in their in-group than to those from out-groups [32]. In some cases, information external to the network structure can be used to infer the groups in the network (e.g. node partisanship labels [11]). However, the group membership or position of each unit is often not known or is difficult to otherwise infer, making graph partition algorithms necessary. Graph partitioning is an ill-posed problem [28], with a large number of algorithms available that try to find slightly different types of clusters in networks. In this paper, we use three different partition algorithms: METIS [43], regularized spectral clustering [78], and modularity [62] optimization algorithm. These methods give similar overall results in our analysis (see Appendix B); for brevity we will show only results for METIS, which searches for two large groups such that there is a minimum number of connections between them.

The main difference between structural polarization scores is how they compute the amount of polarization given a graph and its partition. As noted above, structural polarization between groups is measured by the separation between structurally-identified communities in the network. Here, separation is usually characterized by the discrepancy between interaction patterns that occur within groups and those that cross groups, but scores differ in the kind of interactions they highlight. In this study, we examine eight different structural polarization scores. A brief summary of all polarization scores examined in this study can be found in Table 1. We selected these scores because there is considerable variation in the kinds of interactions they highlight, and because they are measures that have been used in applied work across various fields, including computational social science [18], political science [37], communications [11], and policy studies [12].

At its simplest, structural polarization is measured by comparing the difference in the frequency of external to internal links using the *EI-index* [46] and similar density-based scores [12]. These scores disregard the internal structure of the groups and how the links between them are placed. The *Betweeness Centrality Controversy* (BCC) score [32] alternatively considers the difference in the edge betweenness centrality of external and internal links. Another approach is to use random walk simulations to determine how likely information is to remain within groups or reach external groups (i.e. *Random Walk Controversy* and *Adaptive Random Walk Controversy*; RWC and ARWC) [18, 32, 66, 69]. We also explore the performance of a polarization measure based on community boundaries (*Boundary Polarization*; BP) [35], where a high concentration of high-degree nodes in the boundary of communities implies lower polarization as the influential users are seen as bridging the gap between communities. Lastly, we study a measure based on label propagation (*Dipole Polarization*; DP) [58], where the influence of high-degree nodes is believed to spread via the neighbors in the network, and the distance of the quantified influence of each community is proportional to the polarization. Full definitions of each of the eight structural polarization scores we study are included in Appendix A.

Table 1. Polarization scores used in this study

Polarization score	Domain	Parameters
Random Walk Controversy (RWC) [32]	[-1, 1]	# of influencers in each group
Adaptive Random Walk Controversy (ARWC)	[-1, 1]	% of influencers in each group
Betweenness Centrality Controversy (BCC) [32]	[0, 1]	Kernel for density estimation
Boundary Polarization (BP) [35]	[-0.5, 0.5]	-
Dipole polarization (DP) [58]	[0, 1]	% of influencers in each group
E-I Index (EI) [46]	[-1, 1]	-
Adaptive E-I Index (AEI) [12]	[-1, 1]	-
Modularity (Q) [75]	[-0.5, 1]	-

There are several possible issues with the approach to measuring structural polarization described above. First, if not designed carefully, such polarization scores can be very sensitive to ostensibly unrelated network features such as number of nodes, density, and degree heterogeneity. The scores can also behave in untransparent ways if the groups sizes are not balanced. Finally, graph partitioning algorithms can find clustering structures even in randomized networks, giving them very high polarization scores, especially if they are sparse. These are all problems that have already been found in related problem of partitioning networks to arbitrary number of clusters (as opposed to two) and evaluating the partition [4, 36, 47, 77]. This kind of sensitivity to basic network features and cluster sizes means that the scores are not in an absolute scale where they could be compared across different networks unless these confounding factors are taken into consideration. Problems with polarization scores are not only a theoretical possiblity, but practical problems with the structural polarization score framework have been reported recently. The RWC score, which has been recently described as state-of-the-art [19, 23] and used as the main measure [15, 18, 67], has poor accuracy in separating polarized networks from non-polarized ones [23].

We illustrate some of the problems in measuring structural polarization in practice using two networks (shown in Fig. 2). The first network (#*leadersdebate*) demonstrates the general problem of all scores studied here. Here, the observed network has an RWC value of 0.57. After fixing the observed degree sequence of the network and randomizing everything else, the shuffled networks still scores positively, with an averaged RWC value of 0.27. That is, approximately half of the polarization score value is explained by the size and the degree distribution of the network.

The second network (#*translaw*), with RWC value of 0.68, displays a serious problem related to hubs. A randomized network can have a higher polarization than the observed one due to the absorbing power of the hubs. In other words, a random network with one or more hubs can keep the random walker trapped inside the starting partition even in a non-polarized network. This leads to higher values of RWC as in the case of the *#translaw* network, where the randomized versions obtained an average score of 0.74. Note that this is also higher than the score for the first network, which has a clearly visible division to two clusters in the visualization. As will become evident in the following sections, this is likely due to the larger size and higher density of the first network. We will next systematically analyse how the various network features affect all of the eight structural polarization scores.

3 METHODS

Our primary aim in this paper is to assess how well commonly-used structural polarization measures perform on important aspects of measurement validity [74]. We begin by examining the extent to
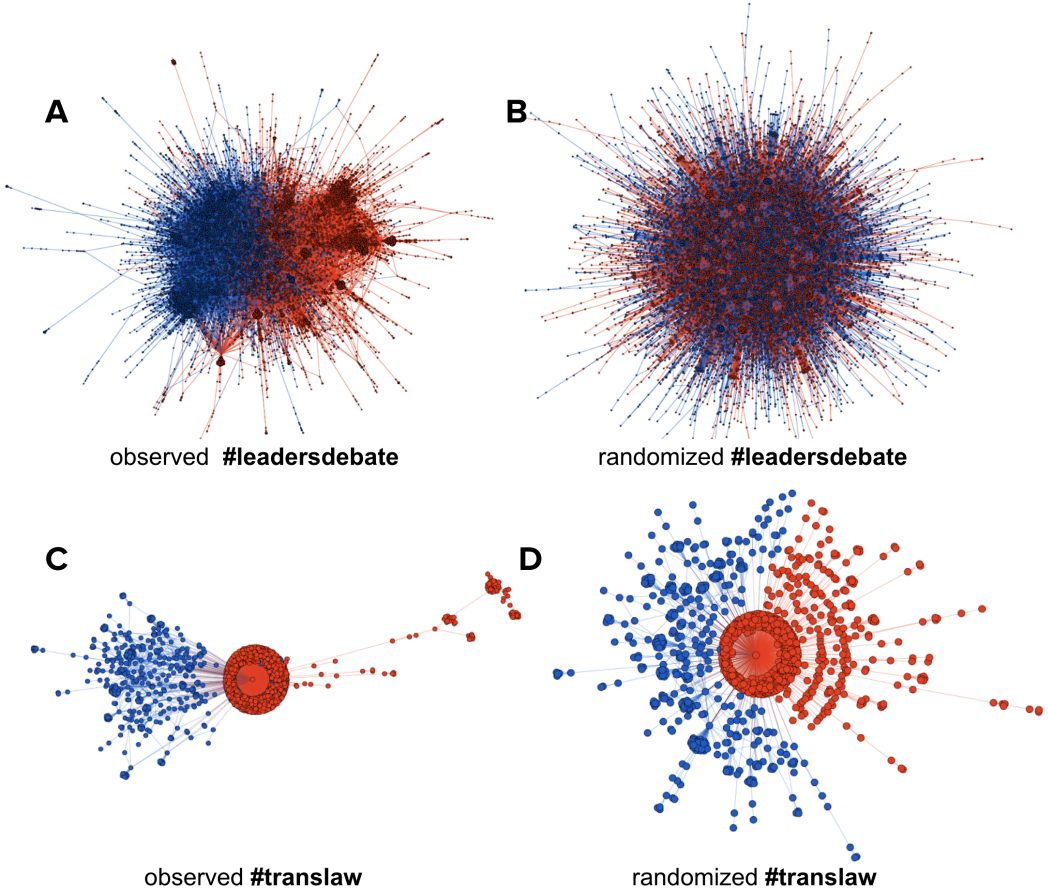
Fig. 2. **A** The network represents a Twitter conversation network for #leadersdebate. Garimella et al. [32] has labeled this topic polarized as it refers to the debate during the U.K. national elections in May 2015. **C** A Twitter endorsement network around #translaw during the Finnish Parliamentary Elections in 2019. (**B**, **D**) For both networks we also visualized their randomized versions obtained with configuration model. The following RWC values ($k = 10$) were computed: observed #leadersdebate 0.57, randomized #leadersdebate 0.27, observed #translaw 0.68 and randomized #translaw 0.74. RWC value of zero indicates no polarization and one indicates maximum polarization. Randomized networks still have positive polarization, and for the randomized #translaw network the value is even higher than the observed network's score. Only the giant component of the network is visualized.

which these eight structural polarization scores are driven by basic network properties using null models based on random networks. These models are built in a way that they fix some structural properties but are otherwise maximally random, i.e., they give equal probability of sampling every network while holding that property constant. There are two main use cases. First, we will systematically analyse various network features by sweeping through the parameter spaces of these models. Second, we can match some of the features of the observed networks and randomize everything else, giving us an estimate of how much of the scores is a consequences of these structural features. Valid measures should not be systematically biased by these structural features.

We use (1) an Erdős-Rényi model [24] (fixing the number of links), (2) a configuration model [29, 56] (fixing the degree sequence), and (3) a model for fixing degree-degree sequences [48]. All of these models fix the number of nodes. One can see these models as a sequence where each of them shuffle less of the data than the previous one [34, 48], because fixing degree-degree sequences automatically fixes the degree distribution, which in turn fixes the number of links. To emphasize this relationship between the models, we label them using the dk-series framework [48, 63]. The dk-distribution of graph G incorporates all degree correlations within d-sized subgraphs, hence allowing us to 'shuffle' the original graph while keeping the desired prescribed properties regarding network topology. Increasing values of d captures more properties of the network at the expense of more complex probability distributions. In the limit of large d, the complete representation of the input graph is obtained. The intuition is very similar to Taylor series. The more terms you include in the series (corresponding to higher values of d), the closer you get to the exact value of the function (i.e. the original network). Within this framework, the Erdős-Rényi network fixing the number of links (or equivalently, the average node degree) is $d = 0$, the configuration model fixing the degree sequence is $d = 1$, and the model fixing the degree-degree sequence is $d = 2$. Each randomized network is repartitioned before computing its polarization score by applying the same graph partitioning algorithm as for the corresponding original network.

In addition to models fixing some observable network properties, we use the stochastic block model [38] to generate networks with two groups. This model is used in Section 4.2.3 to explore how unbalanced group sizes affect the score values.

3.1 Data
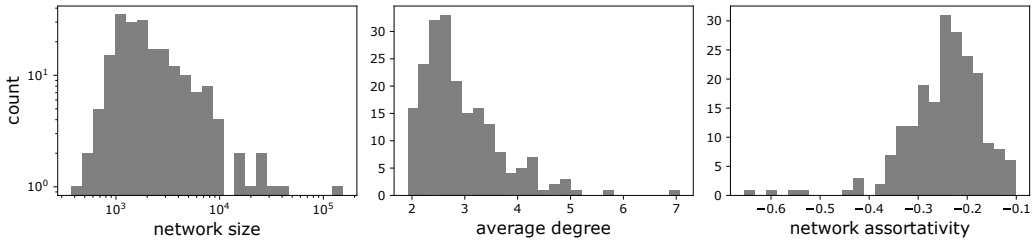

Fig. 3. Distributons of basic statistics of the 203 observed networks. The average network studied had approximately 4000 nodes and an average degree of 3. Network assortativity refers to degree assortativity.

In addition to networks simulated with models, we use real-world data from Twitter from three sources for a total of 203 networks. First, from Chen et al., we used 150 networks constructed from the most popular hashtags during the 2019 Finnish Parliamentary Elections [12]. These were constructed based on single hashtags (e.g. *#police, #nature, #immigration*). Second, from the same source, we included 33 topic networks, which were constructed from sets of hashtags related to larger topics, such as climate change or education [see Appendix A in 12]. Third, we used 20 topic networks from Garimella et al.'s study on the RWC [32]. Each of the 203 resulting network has a node set containing all users who posted an original tweet with a specific set of hashtags and all users who retweeted at least one of these tweets. Undirected ties on the network indicate the connected nodes have at least one instance of retweeting between them on the given topic.

Finally, we process all network prior to assessment by (1) obtaining the giant component of the network as has been done previously [32], (2) removing self-loops, and (3) removing parallel edges. The latter two steps did not have a significant effect on polarization values. The average

network in this study had approximately 4000 nodes, an average degree of 3 and tended to be slightly assortative. Complete summary distributions for the networks included in our study are presented in Fig. 3.

4   RESULTS

4.1   Real Networks

We first compare the observed network data to random networks that are shuffled in a way that we keep some of the features of the networks. As expected, the more features we keep, the more similar the scores are to the ones for original networks (see Fig. 4).

For BP, Q, EI, and AEI the scores produced by the random models cover the observed score for most networks (black bar that corresponds to the observed score of a single network is covered by the other colors that correspond to the scores produced by different random models), indicating that the number of links and size of the networks ($d = 0$) are already enough to predict much of the observed score. For RWC, ARWC, BC, and DP, the more features are kept, the higher (and therfore closer to the original value) the scores tends to be. In general, the change in scores after randomization follows a pattern where both low and high original scores can get very low values for the model keeping the average degree ($d = 0$). The degree sequence ($d = 1$) and degree-degree sequence ($d = 2$) can in many cases explain most of the observed scores, and in some cases the scores for these random networks are even larger than for the original networks. We also have included an alternative way to visualize the polarization values of the randomized networks in Appendix B.

Fig. 4 gives a detailed view of the poplarization scores. It can be used to read scores for each of the original networks and the corresponding random networks. There are four methods for which the $d = 0$ model already explains most of the observed scores, and for the rest the degree sequence ($d = 1$) is usually a very good predictor.

To get a more detailed view on the characteristics that are related with the polarization scores, we show the RWC score as a function of both network size and average degree in Fig. 5. Network size is correlated in a way that smaller networks have higher RWC scores (Spearman correlation coefficient -0.42). After shuffling the real network with fixing the original degree sequence, the smaller and sparser networks have higher RWC scores (respective Spearman correlation coefficients -0.67 and -0.68). Although randomized networks had lower RWC scores than the original networks, the averaged RWC for all networks with average degree less than four was approximately 0.45 even after randomization.

4.2   Synthetic Random Networks

Based on Section 4.1, we see that polarization scores are heavily affected by elementary network statistics like network size, density, and degree distribution. We will next explore more systematically which factors explain the high polarization scores in randomized networks. In addition, to the aforementioned statistics, we analyse the effect of unequal community sizes as real network are likely to have a range of polarized group sizes.

4.2.1   Network size and density. Even an ideal structural polarization score can be correlated with network size and average degree in a specific set of real-world networks, but it should not be biased by these elementary network features in the set of all networks. As a consequence, structural polarization scores should get neutral values for random networks with any size and average degree that are generated without any explicit group structure that is present in the creation process. To assess how the scores perform on this metric, we computed each score for Erdős–Rényi graphs of
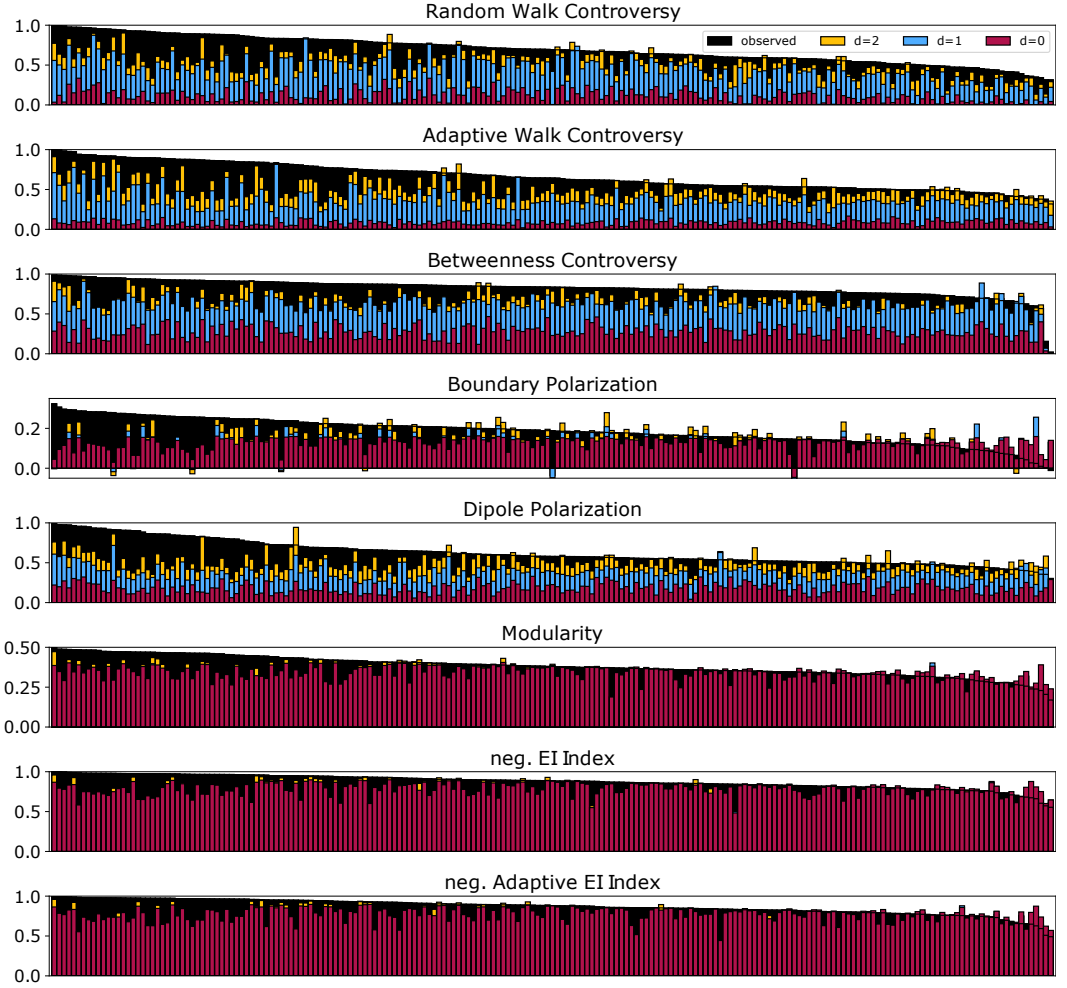
Fig. 4. Polarization scores for the 203 observed networks and their shuffled versions. Each bar corresponds a score, and scores for a network and its randomized versions are on top of each other, ordered from bottom to top in the following order: observed network (black) and randomized networks where degree-degree sequence ($d = 2$, yellow), degree sequence ($d = 1$, blue), or average degree ($d = 0$, red) is preserved. An interpretation for the figure is that, the amount of color that is shown tells how much of the total bar height (the score value) is explained by the corresponding network feature. Note that in some cases, the randomized networks produce higher scores than the original network and in this case the black bar is fully covered by the colored bar(s). In this case we draw a black horizontal line on top of the colored bars indicating the height of the black bar. See Appendix B for similar results obtained by other partition algorithms.

varying sizes and densities. Our results shown in Fig. 6 indicate that all scores are affected by at least one of these elementary network features.

First, we find that network size generally did not contribute to polarization scores, with RWC being the sole exception. It was affected by the number of nodes in the network, giving lower values for larger graphs with the same average degree (see Fig. 6). For networks with enough nodes, the RWC is very close to zero for all values of average degree, but this comes at the cost of the score
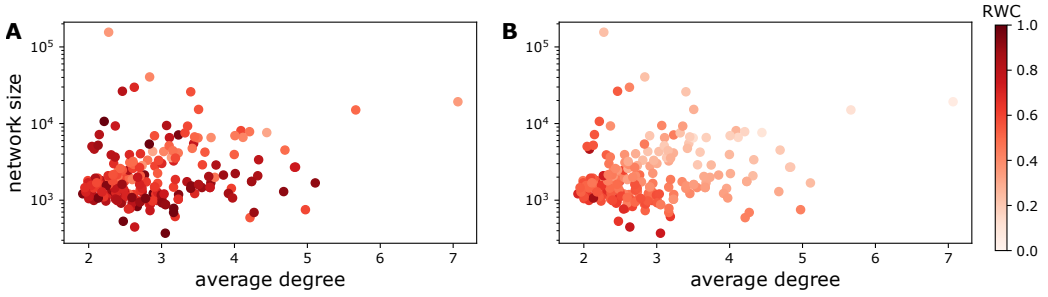
Fig. 5. **A** The RWC-scores for the 203 observed networks as a function of average degree and network size. **B** The same score for configuration model $d = 1$ of the real network.

being sensitive to network size. On the other hand, despite being a similar random walk-based score, the ARWC is invariant to network size. This highlights the difference in their construction. Specifically, the RWC takes as a parameter a fixed number of influencers in the system, meaning that the number of influencers *as a proportion of the network* varies by network size, leading to inconsistent variation in RWC across networks. The ARWC removes this dependence by setting the number of influencers as a function of network size (i.e., as a proportion of network size it remains constant). We discuss the difference between the RWC and ARWC in Appendix A.

Specific instances of RWC aside, all scores are dependent on average degree, and only approach non-polarized levels when the network's average degree exceeds a certain value. For instance, BCC gives zero polarization for random graphs only when the average degree is approximately 10. This is quite strict condition to have especially for Twitter networks. BP decreases almost linearly as a function of density. It reaches the value zero when the average degree of the network is between 5 and 6. The negative values for larger degrees indicate that nodes in the boundary are more likely to connect to the other group. Morales et al., i.e., author behind the DP score, pointed out how their score suffers from the "minimum polarization problem" due to the nonzero standard deviations of the opinion distributions [58].

Dependence between density and modularity has been studied theoretically before for the case where the number of clusters is not restricted to two like in polarization scores. Previous research has shown that, sparse random graphs (and scale-free networks) can have very high modularity scores [36], and that, with high probability, modularity cannot be zero if the average degree is bounded [54]. It is therefore known that using a network's modularity score to measure the amount to which it is clustered is inadvisable. This notion has been speculated to apply for the use of modularity as a polarization score [35]. Our results confirm this notion, and show that modularity behaves similarly for the case where the number of clusters is limited to two, with the difference that the maximum value in our simulations goes up to only approximately 0.5. Here, it should be noted that none of the other scores seem to be immune to this problem.

4.2.2 Heterogeneity of degree sequence. The role of the underlying degree sequence is essential to study as political communication networks tend to have a few nodes with relatively high number of edges. For networks produced by the Erdős–Rényi model, the degree distribution is Binomial centered at the average degree $\langle k \rangle = p(n-1)$, where p is the probability that two nodes are connected and n is the number of nodes in network. In contrast, many real networks' degrees follow fat-tailed distributions which can have considerably higher variance [8, 13, 72]. To analyze the effect of degree distribution, we simulate random graphs whose degree sequences were drawn
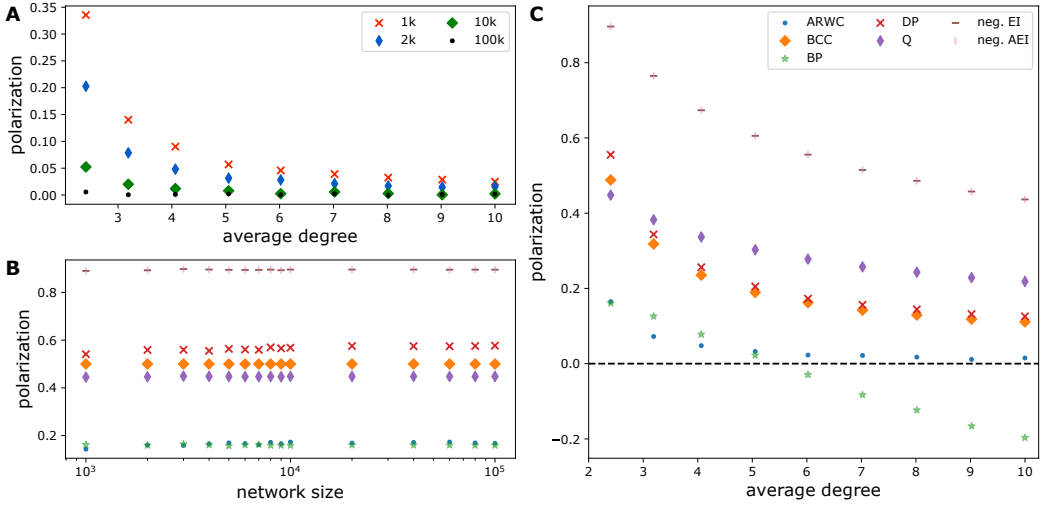
Fig. 6. **A** The expected RWC values as a function of average degree for ER-networks. The symbols denote the network size as indicated in the legend. **B** The expected values of all polarization scores (except RWC) as a function of size for ER networks with average degree 2.4. See panel **C** for legend. **C** The expected values of all polarization scores (except RWC) as a function of average degree for ER networks with 4000 nodes. The dashed line indicates zero polarization. See Appendix B for the polarization scores as a function of both number of nodes and average degree.
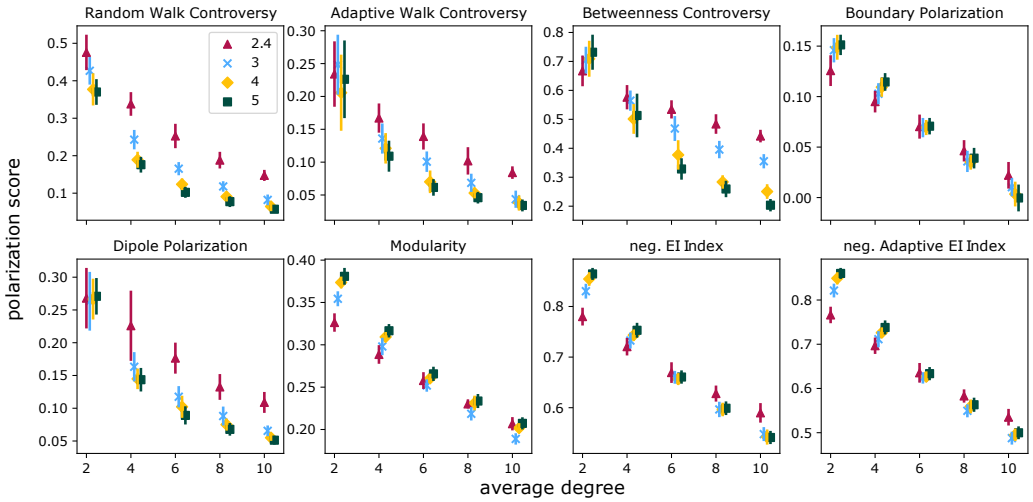

Fig. 7. The effect of degree heterogeneity to polarization scores for the 8 scores in simulated scale-free networks with 1000 nodes. We show the expected polarization score and its standard deviation as a function of average degree. The symbols correspond to different exponent γ values as indicated by the legend. See Figs. 14-15 in Appendix B for similar results for larger networks.

from a power law distribution $P(k) \propto k^{-\gamma}$ [25]. We vary the exponent γ, which allows us to explore
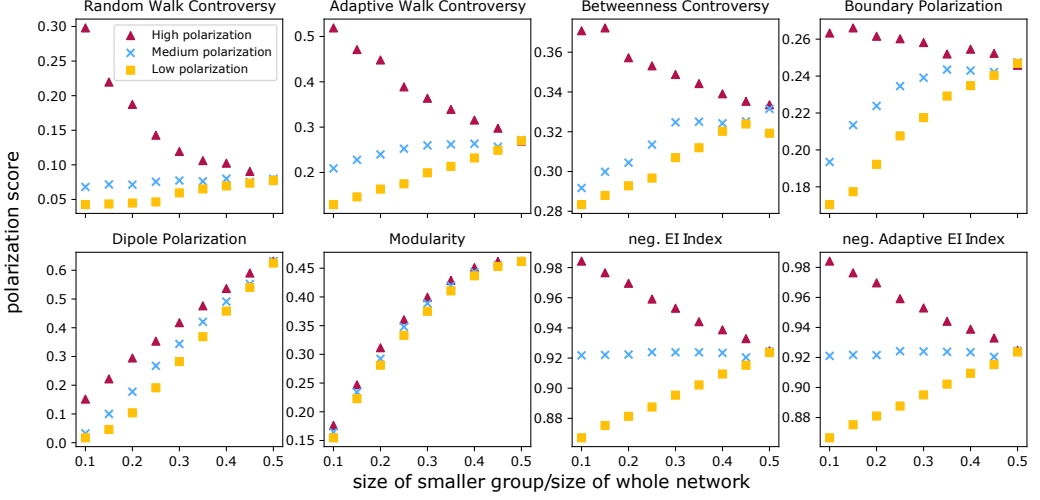
Fig. 8. Polarization scores as the function of relative group size $n_{small}/(n_{small} + n_{large})$ in an SBM model described in the main text. The different markers correspond to the three schemes of adding inter-block links.

the amount of variation in the degree distributions. The lower the value of γ is, the higher is the variation and larger are the hubs.

The RWC, ARWC, BCC, and DP give higher scores for more heterogeneous networks, and there is a slight opposite effect for the other scores given a low average degree (see Fig. 7). The average degree affects polarization scores for networks with heterogeneous degree sequences as well. The observation that the level of polarization approaches zero only when the network becomes denser still holds for all scores, but for the scores that are lifted by degree heterogeneity, much denser networks are required in order for the scores to become close to zero.

4.2.3 Communities of different sizes. Polarized groups within a system are often imbalanced in size, making it important for scores to perform consistently across a range of group size balance. To assess this metric, we used the stochastic block model to generate a set of polarized networks differing by group size imbalance and level of polarization. All networks were fixed to have size $n = 10000$, and the size of the smaller group ranges from between 10% to 50% of the whole network. Within-group links were generated for each node to have an average within-group degree of k_{in}. To obtain different levels of polarization, we generated between-group links at three levels of density. For *low polarization* networks, links were generated with the expectation of adding $k_{out} = k_{in}/c$ between-group links per node from the larger group. Conversely, *high polarization* networks have an expected k_{out} between-group links per node from the smaller group. A third set of networks, with *medium polarization*, have an expected $k_{out} \times n/2$ total between-group links. Our networks were generated with $k_{in} = 4.5$ and $c = 25$, but our results are robust to a reasonable range of different values.

Our results indicate that all scores depend on group size imbalance at least for the high and low polarization schemes (see Fig. 8). The EI and AEI scores are relatively insensitive to the group size imbalances as their values change by only a few percentage points. For all scores except DP and Q, simulated level of polarization affects the extent to which they are dependent on group size imbalances; at certain levels of simulated polarization, the dependence disappears. For EI and AEI this level is exactly the one present in the medium polarization networks. Finally, it is worth noting

that DP has an intrinsic scaling factor which penalizes the difference between groups. Specifically, its coefficient is designed to contribute maximally to the polarization only when the communities are equally-sized, thus the linear relationship between imbalance and score.

5 EVALUATION AND NORMALIZATION OF THE POLARIZATION SCORES

As our analysis suggests, nonzero polarization values arise in randomized networks due to the scores being sensitive to network size, number of links, and degree distribution. These features in themselves are not likely to be linked to polarization. The number of nodes and the number of links depend on unrelated system features such as overall activity in the network and and design choices such as link sampling schemes. Further, the fat-tailed degree distributions we often observe in the data have been attributed to preferential-attachment mechanisms [8].

However, even if scores do not return zero values for random networks or are biased by some features of the data, they can still be useful if they are able to separate polarized networks from non-polarized ones by systematically returning higher scores for the former set. In this section, we assess the predictive validity of the scores against a manually labeled set of 203 polarized and non-polarized networks, which we introduced in Section 3.1. The 20 networks from Garimella et al.'s study are already manually labeled [32], so we are able to directly use these external labels. The 183 networks from Chen et al.'s study had not been labeled [12], so we manually labeled them for our present task.

We labeled these networks based on the content of tweets, which ensures that our labeling is independent of the structural polarization measures. We checked whether any of the pre-defined features, which are selected based on our definition of polarization, were present in a reasonable number of tweets before marking the network polarized. The labeling process, which we describe below, was performed before the main analysis, and resulted in a balanced data set with around 47% of the networks labeld as polarized. To the extent that the basic network features we examined are not indicative of polarization, we should see an increase in the classification performance of the scores when the effects induced by them are removed.

We recognize that polarization labels based on content alone are not necessarily the ground truth on polarization in the system. Instead, content-based labeling is another method for quantifying polarization, which is based on a set of criteria different from those used in structural polarization scores. Because it is another label of the same underlying latent construct, content-based labeling is useful for assessing the validity of structural polarization scores. If a structural score correlates well with our content-based labels, they are said to have high *convergent validity*, which is a form of measurement validity[1]. This means that it can be seen as a better measure of the latent overall polarization in the system compared to a less correlated structural score.

5.1 Data Labeling

We labeled our networks before performing the main analysis. All networks with at least one hashtag containing the substring 'election' were classified as polarized. We manually sampled tweets from each network for confirmation. For each network, we applied a four-stage process for labeling in the following order:

(1) Sample uniformly 5 days from which tweets are read.
(2) Display all the tweets from each sampled day.
(3) Sample 20 users from each sampled day and display all their tweets during the sampled days.
(4) Partition the network and distinguish 10 highest degree nodes from both groups. Display all their tweets.
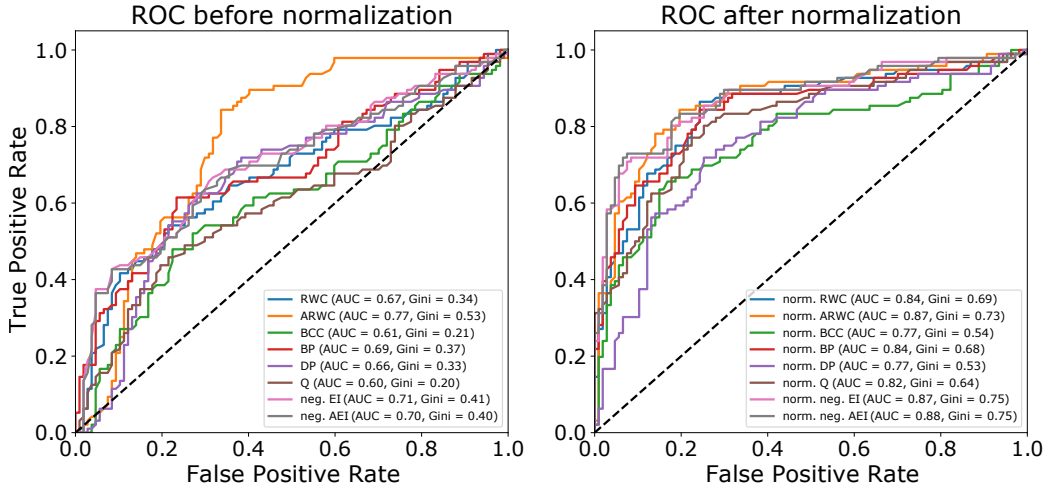
Fig. 9. ROC curves, Gini coefficient values, and AUC values for the task of predicting manually curated labeling of polarized and non-polarized networks. The results shown (left) for the score values before the normalization and (right) after the normalization with denoised scores $\hat{\Phi}(G)$ (see main text). Figures are based on all the 203 empirical networks described in section 3.1. Results for the different normalization scheme $\hat{\Phi}_z(G)$ including the standardization are shown in the Fig. 16 in Appendix B. The ROC curves for the alternative graph partitioning methods are also reported in Figs. 22 and 23 in Appendix B.

After displaying the tweets that were obtained by the described process, we checked whether any of the following features were present in a reasonable number of tweets.

- us-versus-them mentality, signs of disagreement, dispute, or friction
- strongly discrediting the out-group or strongly supporting the in-group from both sides
- direct, negative, or strong criticism of political adversaries or political actors from both sides
- completely opposite opinions, beliefs, or points of view on a political or social topic

Based on the content of the sampled tweets together with domain knowledge, a researcher classified the network either polarized or non-polarized. If the sample was too vague to be labeled, we repeated the process to gain a clearer view into the general content of tweets.

5.2 Classification Performance of the Polarization Scores

We adapt a typical framework where there is a decision threshold for the score value under which all networks are predicted to be non-polarized and above which they are prediced to be polarized. Each threshold value produces a false positive rate and true positive rate, and varying the threshold gives us an ROC curve (shown in Fig. 9) characterizing the overall performance of the scores in the prediction task. This makes the evaluation independent on the selected type of classifier. We also derive single scores to quantify the overall performance. The *Gini coefficient* measures the performance improvement compared to a uniformly random guess and is defined as the area between the diagonal line and the ROC curve, normalized so 1 indicates perfect prediction and 0 indicates random guess. We also report the unnormalized area under curve (AUC).

The Gini coefficients for the scores vary between 0.20 and 0.53 with Q and BCC performing the worst. ARWC performs the best with a wide margin to the second best AEI and EI (with coefficient values 0.40 and 0.41). The non-adaptive RWC has a Gini coefficient of 0.34, which is better than prior work shows [23], but still generally poor. Notably, the ARWC score performs very well if
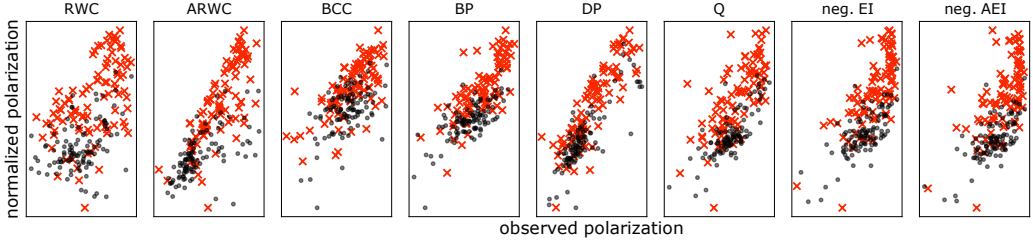
Fig. 10. Scatter plot of denoised ($\hat{\Phi}$) and observed polarization scores for the 203 networks described in Section 3.1. Red crosses are networks labeled as polarized and black points are networks labeled as non-polarized. An outlier was removed from the plots. Note that the scales for the scores are different and not shown in this figure. See Fig. 19 in Appendix B for the equivalent scatter plot for the standardized scores ($\hat{\Phi}_z$).

we accept a high false positive rate. That is, if a system has a very small ARWC score, it is a good indication of the network not being polarized. In contrast, a large ARWC score (or any other measure) does not necessarily indicate that the network is polarized. As our prior results show, a high score might be due to the effects of small and sparse networks with hubs as opposed to anything related to polarized group structures.

5.3 Normalization of the scores

To remove the effect of network size and the degree distribution, we computed the averaged polarization score for multiple instances of the network shuffled with the configuration model ($d = 1$ in the dk-framework), and subtracted it from the observed polarization score value. That is, given a network G and a polarization score Φ, we define a normalized score as

$$\hat{\Phi}(G) = \Phi(G) - \langle \Phi(G_{CM}) \rangle,$$

where $\Phi(G)$ is the polarization score of observed network and $\langle \Phi(G_{CM}) \rangle$ is the expected polarization score of graphs generated by the configuration model. This score corrects for the expected effect of the size and degree distribution of the network (i.e. removes the blue part from the observed score shown previously in Fig. 4). Thus we call it *the denoised polarization score*. It does not consider the fact that there is some fluctuations in the score values in the configuration model. We correct for this in another, slightly different normalization scheme, where we divide the normalized score by the standard deviation of the score value distribution for the configuration model:

$$\hat{\Phi}_z(G) = \frac{\Phi(G) - \langle \Phi(G_{CM}) \rangle}{\sqrt{\langle \Phi(G_{CM})^2 \rangle - \langle \Phi(G_{CM}) \rangle^2}}.$$

We call this normalization *standardized and denoised polarization score*. Note that the distribution of polarization value over the ensemble of null random graphs is not necessarily Gaussian. If the values $\Phi(G_{CM})$ followed Gaussian distribution, then the statistical significance testing could be performed with the standard normal distribution, and $\hat{\Phi}_z(G_{CM})$ would be the appropriate test statistic (the z-score). An approximate value for a significance can be obtained with large number of samples. The same normalization has been proposed for modularity to mitigate the resolution limit [55]. In that work, the proposed quality function gives higher values for statistically rarer observations.

Fig. 9 shows the ROC curves and related coefficient values for the denoised scores $\hat{\Phi}(G)$ (see the qualitatively similar results in Appendix B for $\hat{\Phi}_z(G)$). The performance of all of the scores increased after normalization as indicated by the overall lifted ROC curves and improved Gini coefficients
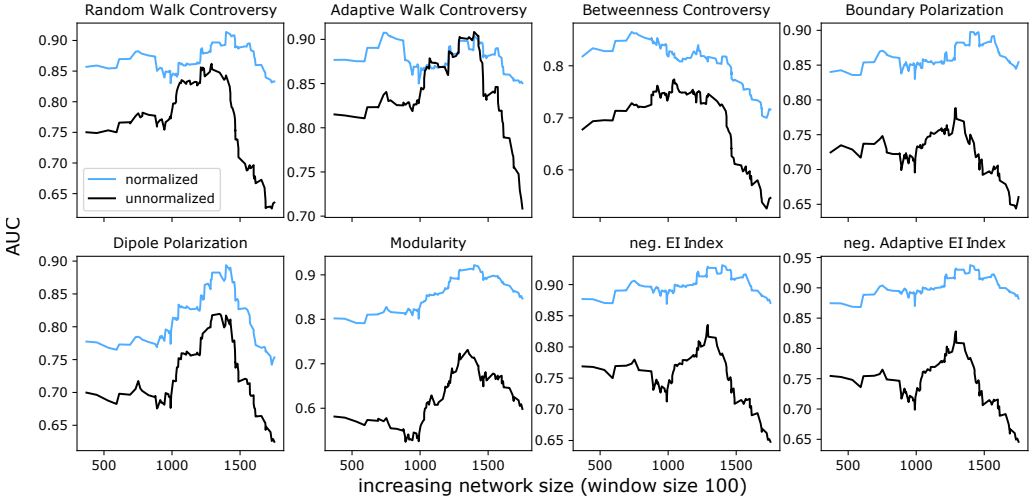
Fig. 11. Quantifying how network's average degree affects the performance. We group the data such that there are 100 networks with consecutive sizes in our data, and create a set of such windows by varying the size range. We then evaluate the AUC for the moving window of 100 networks. Generally, all the networks benefits from the normalization across all the polarization methods. The same analysis for average degree is included in Appendix B. Also the details on the scale of the windows are included in Appendix B.

(and AUC values). Improvements in Gini coefficients after normalization ranges between 38% - 220% depending on the measure used. The ARWC remains among the best performing scores, along with the AEI and EI. The AEI in particular, perfoms best under conservative estimates of polarization (i.e., low false positive rates). The BCC is still among the worst after the normalization, along with DP. However, all post-normalization scores notably outperform the best unnormalized score (ARWC).

Fig. 10 further illustrates the dramatic change in the predictive power of the scores after normalization. It shows the normalized score values as a function of the observed score values. With most scores, the two types of networks are mixed together when sorted with the observed scores (on the x-axis). Normalization largely lifts the polarized networks above the cloud of non-polarized ones, making it possible to separate these two groups (on the y-axis). This finding holds for all the polarization measures analyzed here.

Note that in practice, to implement the normalization procedure, one needs to randomize the network multiple times with the configuration model and compute the score $\Phi(G_{CM})$ to get estimates for the mean value of the score $\langle\Phi(G_{CM})\rangle$ (and $\langle\Phi(G_{CM})^2\rangle$ for $\hat{\Phi}_z(G)$). Here we sampled the networks 500 times which was more than enough samples as they lead to error of the means ranging from 0.01 to 0.05.

To determine which types of networks benefits the most from the normalization, we plotted the AUC as a function of network size and average degree for each polarization score. This was done by evaluating the performance for subsets of networks with a fixed window size of 100 (shown in Fig. 11). The results show how the performances for all the polarization methods are better and more stable after the normalization independent on the network size. Only ARWC has a short region where the performances of both normalized and unnormalized scores overlap. The same analysis for average degree is included in Appendix B.

We also tested whether combining the results from all polarization methods improves the accuracy of predicting whether a network is polarized. Our first strategy was to take the average of all the scores and use that as a new polarization score. The second strategy was to train a simple decision tree classifier where the input vector contained all eight scores obtained for a network. The AUC for the average of unnormalized values was 0.71, and for normalized values it increased to 0.87. Although the averaged normalized score outperformed some of the single normalized polarization scores (e.g. BP, DP, and Q), it did not outperform the best-performing ones (e.g. ARWC and AEI). Regarding the decision tree, the AUC of the pre-normalization classifier was 0.78, whereas for the post-normalization one, the AUC increased to 0.90. Our results show that strategies based on combined scores can in some cases offer improvements over single polarization scores, but only minimally. It is up to the researcher to decide if these gains are worth the cost of additional work and loss of transparency associated with training machine learning models.

6 DISCUSSION

Measuring polarization is important for social science research, including the social computing and computer mediated communication fields. Structural polarization measures offer an ostensibly promising approach, but we identified a number of undesirable properties associated with all eight commonly-used measures studied here. These measures can be high for random networks and they are sensitive to various network features such as size, average degree, and degree distribution. These properties pose a clear problem for polarization research. Considerable research effort has been put into polarization identification and platform design for attenuating polarization, but if the measurement of polarization is systematically biased by basic network features, our ability to make valid inferences are greatly reduced. For example, consider Bail et al.'s study that found increasing exposure to opposing views increased political polarization [5]. The study did not rely on structural polarization measures, but had this study been conducted in the field using the RWC to measure polarization, the increased activity that likely would have resulted from the intervention could have decreased polarization scores, resulting in the exact opposite conclusion being drawn.

Based on our results, we strongly recommend applying the normalization procedures introduced in Section 5.3 for applied work using any of the network-based polarization scores included here. Doing so removes a substantial amount of noise arising from the network's local properties. For our test networks, classification performance improved by 38%-220% (Gini coefficient) depending on the measure. Notably, the differences in performance across polarization scores were minor after normalization. In fact, the AEI and EI, which are the simplest and least computationally-demanding scores, were among the best performing scores.

In order for us to draw qualitative conclusions based on the score values we should understand the scale of the score, e.g., what values constitute medium or high polarization. The unnormalized score values often have interpretation described in the logic of their definition. Despite their relatively high accuracy, normalized scores are less amenable to this kind of direct interpretation. If this is needed, a plausible alternative is to report both the score itself and its expected value in one or more random network models. This way, one has a sense of how much of the score is explained by the various features of the network.

Our work has implications for additional structural polarization scores not studied here, including those in development. It is clear from our results that structural scores obtained via the consecutive procedures of network clustering and score computation are susceptible to being sensitive to various network features in a way that is not apparent from the score's definition. Our argument (and others' before us [35]), backed up by the results that normalization increases the performance of the scores, is that these sensitivities bias the scores. At a minimum, one should have a clear idea of how the given score behaves in relation to basic network statistics and group sizes. To

facilitate such examination, we have made our analysis pipeline and data publicly available [70, 71]. There could be other possible sources of bias, so our benchmarking framework should be taken as a minimal test that is not necessarily sufficient.

More broadly, the fact that all eight scores we tested were affected to some extent by the same problems suggests that the approach of separating polarization scoring into independent clustering and score calculation phases might be flawed. This is part of the wider problem where clustering methods start from the assumption that the given network contains clusters and can find them even in random data. A solution to this problem could be to break the independence of the clustering phase and the score calculation phase, using instead clustering methods that can test if the clusters could be explained by random networks [47, 65]. Scores can be set to zero if no significant clusters are found. This reduces the false positive rate, which was especially problematic with the best-performing ARWC method.

Our study presents some limitations which can be avenues for future research. First, our results are based purely on simulations, and a more theoretical approach could be taken to understand the individual scores better. This work can build on the prior work on fundamental problems in graph clustering which, as illustrated here, are reflecting onto polarization scores. In this context, modularity is a well-studied example of how an apparently reasonably defined method can have many underlying problems [4, 27, 36, 53, 54]. Given this, analyzing modularity from the perspective of limiting the number of clusters to two could be done as an extension to the previous literature on modularity optimisation with an arbitrary number of communities. Even if modularity is not important as a structural polarization score, this analysis could shed light on the kind of phenomena to expect when scoring networks with two clusters.

Second, public opinion on politics can be multisided. This means that instead of having only two groups, there can be multiple cohesive clusters that are segragated in some way in the network. However, the majority of polarization measures, including the structural measures analyzed here are defined exclusively for two cluster, with the exception of modularity and the EI-index. Conceptual and technical work that generalizes polarization to the multiside context is therefore useful. This is a nascent area of study [68], with some extensions to structural measures [33, 50]. Such generalizations are likely to retain the same problems as their two-sided variants, because more degrees of freedom in the number of groups for the clustering algorithms will lead to better clusters (as measured with the internal evaluation metrics of the methods). As discussed in section 4.2.1, previous work on modularity can again be useful here, as it indicates that the issues high score values in random networks is even worse when the number of clusters is not limited.

Further, a clear limitation of the current work is the number and variety of labeled network data that was used. While the number of network is enough to statistically show that normalization improves the score performance, a more fine-grained view of the problem could be achieved with more networks. Similarly, the generalizability of the classification results could be improved by widening the range of network size and density but more importantly by including different types of social networks. Here, it is worth noting that our approach to labeling the content might not be as clear cut for other contexts such as non-political communication. Finally, the analysis regarding the different-sized clusters can be improved. Although our results indicated that all scores depend on group size imbalance at least for the low and high polarization schemes, other techniques for simulating polarization between the communities should be examined.

Despite the issues we raised in this paper, structural polarization measures as an approach remains useful. In addition to being based on emergent behavior directly observed in the system under study, they facilitate an accessible approach to studying polarization. A network-based approach generally has low language processing requirements, making it equally easy to apply to different linguistic contexts. Additionally, it has been argued that there is an uneven development of natural

language processing tools across languages [22], which presents an additional barrier to content-based polarization measures. Content-based polarization measures often require language-specific resources such as sentiment lexicons, which are often costly to build [73]. Similarly, survey-based measures, especially from original data sources, are costly to obtain. In light of these sometimes considerable barriers to research, structural polarization measures provide an accessile alternative for applied researchers.

To facilitate the use of structural polarization measures, we introduced in this paper a minimal set of tests that a structural polarization score should pass in order for it to be able to distinguish polarization from noise. This should serve as a benchmark for future developments of such scores. The fact that the all of the current scores perform relatively poorly indicates that there is a need for alternative approaches to the typical scoring approach. The normalization procedure we introduced here is a patch that alleviates this problem. There is space for other, possibly fundamentally different approaches, to be innovated for measuring structural polarization.

ACKNOWLEDGMENTS

REFERENCES

[1] Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review* 95, 3 (2001), 529–546.

[2] Leman Akoglu. 2014. Quantifying Political Polarity Based on Bipartite Opinion Networks. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014). https://ojs.aaai.org/index.php/ICWSM/article/view/14524

[3] Alessia Amelio and Clara Pizzuti. 2012. Analyzing Voting Behavior in Italian Parliament: Group Cohesion and Evolution. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 140–146. https://doi.org/10.1109/ASONAM.2012.33

[4] James P Bagrow. 2012. Communities and bottlenecks: Trees and treelike networks have high modularity. *Physical Review E* 85, 6 (2012), 066118.

[5] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[6] Delia Baldassarri and Peter Bearman. 2007. Dynamics of political polarization. *American sociological review* 72, 5 (2007), 784–811.

[7] Delia Baldassarri and Andrew Gelman. 2008. Partisans without constraint: Political polarization and trends in American public opinion. *Amer. J. Sociology* 114, 2 (2008), 408–446.

[8] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512. https://doi.org/10.1126/science.286.5439.509

[9] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. 2020. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters* 124, 4 (2020), 048301.

[10] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. 2020. Learning political polarization on social media using neural networks. *IEEE Access* 8 (2020), 47177–47187.

[11] Jonathan Bright. 2018. Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication* 23, 1 (2018), 17–33.

[12] Ted Hsuan Yun Chen, Ali Salloum, Antti Gronow, Tuomas Ylä-Anttila, and Mikko Kivelä. 2020. Polarization of Climate Politics Results from Partisan Sorting: Evidence from Finnish Twittersphere. *arXiv:2007.02706* (2020).

[13] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[14] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Political Polarization on Twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847

[15] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. 2020. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 130–140.

[16] Wesley Cota, Silvio C Ferreira, Romualdo Pastor-Satorras, and Michele Starnini. 2019. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science* 8, 1 (2019), 35.

[17] Carlo Dal Maso, Gabriele Pompa, Michelangelo Puliga, Gianni Riotta, and Alessandro Chessa. 2014. Voting behavior, coalitions and government strength through a complex network analysis. *PloS one* 9, 12 (2014), e116046.

[18] Kareem Darwish. 2019. Quantifying Polarization on Twitter: The Kavanaugh Nomination. In *Social Informatics*. Springer International Publishing, Cham, 188–201.

[19] Juan Manuel Ortiz de Zarate, Marco Di Giovanni, Esteban Zindel Feuerstein, and Marco Brambilla. 2020. Measuring Controversy in Social Networks Through NLP. In *String Processing and Information Retrieval*. Springer International Publishing, Cham, 194–209.

[20] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2970–3005.

[21] Paul DiMaggio, John Evans, and Bethany Bryson. 1996. Have American's social attitudes become more polarized? *American journal of Sociology* 102, 3 (1996), 690–755.

[22] Fahim Djatmiko, Ridi Ferdiana, and Muhammad Faris. 2019. A Review of Sentiment Analysis for Non-English Language. In *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*. IEEE, 448–451.

[23] Hanif Emamgholizadeh, Milad Nourizade, Mir Saman Tajbakhsh, Mahdieh Hashminezhad, and Farzaneh Nasr Esfahani. 2020. A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). *Social Network Analysis and Mining* 10, 1 (2020), 1–20.

[24] P Erdős and A Rényi. 1959. On random graphs I. *Publicationes Mathematicae* 6 (1959), 290–297.

[25] Dario Fasino, Arianna Tonetto, and Francesco Tudisco. [n.d.]. Generating large scale-free networks with the Chung–Lu random graph model. *Networks* ([n. d.]). https://doi.org/10.1002/net.22012

[26] Morris P Fiorina and Samuel J Abrams. 2008. Political polarization in the American public. *Annu. Rev. Polit. Sci.* 11 (2008), 563–588.

[27] Santo Fortunato and Marc Barthelemy. 2007. Resolution limit in community detection. *Proceedings of the national academy of sciences* 104, 1 (2007), 36–41.

[28] Santo Fortunato and Darko Hric. 2016. Community detection in networks: A user guide. *Physics reports* 659 (2016), 1–44.

[29] Bailey K Fosdick, Daniel B Larremore, Joel Nishimura, and Johan Ugander. 2018. Configuring random graph models with fixed degree sequences. *SIAM Rev.* 60, 2 (2018), 315–355.

[30] Kiran Garimella. 2018. *Polarization on Social Media*. Ph.D. Dissertation. Aalto University. http://urn.fi/URN:ISBN:978-952-60-7833-5

[31] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 913–922. https://doi.org/10.1145/3178876.3186139

[32] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[33] Noé Gaumont, Maziyar Panahi, and David Chavalarias. 2018. Reconstruction of the socio-semantic dynamics of political activist Twitter networks—Method and application to the 2017 French presidential election. *PloS one* 13, 9 (2018), e0201879.

[34] Laetitia Gauvin, Mathieu Génois, Márton Karsai, Mikko Kivelä, Taro Takaguchi, Eugenio Valdano, and Christian L Vestergaard. 2018. Randomized reference models for temporal networks. *arXiv:1806.04032* (2018).

[35] P.H. Guerra, Wagner Meira Jr, Claire Cardie, and R. Kleinberg. 2013. A measure of polarization on social media NetworksBased on community boundaries. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* (01 2013), 215–224.

[36] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. 2004. Modularity from fluctuations in random graphs and complex networks. *Physical Review E* 70, 2 (2004), 025101.

[37] Eszter Hargittai, Jason Gallo, and Matthew Kane. 2008. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice* 134, 1-2 (2008), 67–86.

[38] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social networks* 5, 2 (1983), 109–137.

[39] Michael Hout and Christopher Maggio. 2020. Immigration, Race, and Political Polarization. https://doi.org/10.31235/osf.io/p7q2w

[40] Luca Iandoli, Simonetta Primario, and Giuseppe Zollo. 2021. The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change* 170 (2021), 120924.

[41] Lorien Jasny, Joseph Waggle, and Dana R Fisher. 2015. An empirical examination of echo chambers in US climate policy networks. *Nature Climate Change* 5, 8 (2015), 782–786.

[42] David R Jones. 2001. Party polarization and legislative gridlock. *Political Research Quarterly* 54, 1 (2001), 125–141.

[43] George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing* 20, 1 (1998), 359–392.

[44] B. W. Kernighan and S. Lin. 1970. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal* 49, 2 (1970), 291–307. https://doi.org/10.1002/j.1538-7305.1970.tb01770.x

[45] Justin H Kirkland and Justin H Gross. 2014. Measurement and theory in legislative networks: The evolving topology of Congressional collaboration. *Social Networks* 36 (2014), 97–109.

[46] David Krackhardt and Robert N Stern. 1988. Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly* (1988), 123–140.

[47] Andrea Lancichinetti, Filippo Radicchi, and José J Ramasco. 2010. Statistical significance of communities in networks. *Physical Review E* 81, 4 (2010), 046110.

[48] Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. 2006. Systematic topology analysis and generation using degree correlations. *ACM SIGCOMM Computer Communication Review* 36, 4 (2006), 135–146.

[49] Christos Makridis and Jonathan T Rothwell. 2020. The real cost of political polarization: evidence from the COVID-19 pandemic. *Available at SSRN 3638373* (2020).

[50] Moritz Markgraf and Manfred Schoch. 2019. Quantification of Echo Chambers: A Methodological Framework Considering Multi-party Systems. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*. https://aisel.aisnet.org/ecis2019_rp/91/

[51] Lilliana Mason. 2015. "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science* 59, 1 (2015), 128–145.

[52] Jennifer McCoy, Tahmina Rahman, and Murat Somer. 2018. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist* 62, 1 (2018), 16–42.

[53] Colin McDiarmid and Fiona Skerman. 2017. Modularity of regular and treelike graphs. *Journal of Complex Networks* 6, 4 (10 2017), 596–619. https://doi.org/10.1093/comnet/cnx046 arXiv:https://academic.oup.com/comnet/article-pdf/6/4/596/25451962/cnx046.pdf

[54] Colin McDiarmid and Fiona Skerman. 2020. Modularity of Erdős-Rényi random graphs. *Random Structures & Algorithms* 57, 1 (2020), 211–243.

[55] Atsushi Miyauchi and Yasushi Kawase. 2016. Z-score-based modularity for community detection in networks. *PloS one* 11, 1 (2016), e0147805.

[56] Michael Molloy and Bruce Reed. 1995. A critical point for random graphs with a given degree sequence. *Random structures & algorithms* 6, 2-3 (1995), 161–180.

[57] James Moody and Peter J Mucha. 2013. Portrait of political party polarization. *Network Science* 1, 1 (2013), 119–121.

[58] Alfredo Jose Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25, 3 (2015), 033114.

[59] Zachary P Neal. 2020. A sign of the times? Weak and strong polarization in the US Congress, 1973–2016. *Social Networks* 60 (2020), 103–112.

[60] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. 2018. Social media is polarized, social media is polarized: towards a new design agenda for mitigating polarization. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 957–970.

[61] Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. 2019. (Re) Design to Mitigate Political Polarization: Reflecting Habermas' ideal communication space in the United States of America and Finland. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–25.

[62] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.

[63] Chiara Orsini, Marija M Dankulov, Pol Colomer-de Simón, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E Bassler, Zoltán Toroczkai, Marián Boguná, Guido Caldarelli, et al. 2015. Quantifying randomness in real networks. *Nature communications* 6, 1 (2015), 1–10.

[64] Mert Ozer, Mehmet Yigit Yildirim, and Hasan Davulcu. 2017. Measuring the Polarization Effects of Bot Accounts in the US Gun Control Debate on Social Media. *Proceedings of ACM Conference (Conference'17)* (2017).

[65] Tiago P Peixoto. 2013. Parsimonious module inference in large networks. *Physical review letters* 110, 14 (2013), 148701.

[66] Abdullateef Rabab'ah, Mahmoud Al-Ayyoub, Yaser Jararweh, and Mohammed Al-Kabi. 2016. Measuring the Controversy
     Level of Arabic Trending Topics on Twitter. In *2016 7th International Conference on Information and Communication
     Systems (ICICS)*. https://doi.org/10.1109/IACS.2016.7476097
[67] Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, and Cansın Bayrak. 2020. Embeddings-Based
     Clustering for Target Specific Stances: The Case of a Polarized Turkey. arXiv:2005.09649 [cs.SI]
[68] Andres Reiljan. 2020. 'Fear and loathing across party lines'(also) in Europe: Affective polarisation in European party
     systems. *European journal of political research* 59, 2 (2020), 376–396.
[69] Anna Rumshisky, Mikhail Gronas, Peter Potash, Mikhail Dubov, Alexey Romanov, Saurabh Kulshreshtha, and Alex
     Gribov. 2017. Combining Network and Language Indicators for Tracking Conflict Intensity. In *Social Informatics*,
     Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Springer International Publishing, Cham, 391–404.
[70] Ali Salloum. 2021. Code and Data for Separating Controversy from Noise: Comparison and Normalization of Structural
     Polarization Measures. https://github.com/alesalloum/normalized_polarization.
[71] Ali Salloum, Ted Hsuan Yun Chen, and Mikko Kivelä. 2021. Data for Separating Controversy from Noise: Comparison
     and Normalization of Structural Polarization Measures. https://doi.org/10.5281/zenodo.4434245.
[72] Matteo Serafino, Giulio Cimini, Amos Maritan, Andrea Rinaldo, Samir Suweis, Jayanth R Banavar, and Guido Caldarelli.
     2021. True scale-free networks hidden by finite size effects. *Proceedings of the National Academy of Sciences* 118, 2
     (2021).
[73] Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining
     systems. *Information fusion* 36 (2017), 10–25.
[74] William MK Trochim and James P Donnelly. 2001. *Research methods knowledge base*. Vol. 2. Atomic Dog Pub.
[75] Andrew Scott Waugh, Liuyi Pei, James H. Fowler, Peter J. Mucha, and Mason A. Porter. 2011. Party Polarization in
     Congress: A Network Science Approach. arXiv:0907.3509 [physics.soc-ph]
[76] Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in egypt on
     twitter. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*.
     290–297.
[77] Pan Zhang and Cristopher Moore. 2014. Scalable detection of statistically significant communities and hierarchies,
     using message passing for modularity. *Proceedings of the National Academy of Sciences* 111, 51 (2014), 18144–18149.
[78] Yilin Zhang and Karl Rohe. 2018. Understanding Regularized Spectral Clustering via Graph Conductance. In *Proceedings
     of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran
     Associates Inc., Red Hook, NY, USA, 10654–10663.
[79] Jack Zhou. 2016. Boomerangs versus javelins: how polarization constrains communication on climate change. *Envi-
     ronmental Politics* 25, 5 (2016), 788–811.

A   POLARIZATION MEASURES

Here we briefly introduce the definitions of polarization measures studied in the main body of
the paper. Each score here assumes two disjoint sets A and B. Here a network refers to the giant
component only. Let $V = A \cup B$ be the set of nodes and E be the set of edges in network. The
membership c_i of the nodes are determined during the network partition stage. In a polarized
network, the sets are expected to represent the opposing communities or sides. Some of the scores,
such as Random Walk Controversy and Boundary Polarzation, are also designed to capture potential
antipolarization behavior.

(1) Random Walk Controversy
    This measure captures the intuition of how likely a random user on either side is to be
    exposed to dominant content produced by influencers from the opposing side. From both sets
    k nodes with the highest degrees are selected and labeled as influencers. The high-degreeness
    of a node is assumed to indicate a large number of received endorsements on the specific
    topic, thus called influencers.
    A random walk begins randomly from either side with equal probability and terminates only
    when it arrives in any influencer node (absorbing state). Based on the distribution of starting
    and ending sides of the multiple random walks, the score is computed as

$$P_{RWC} = p_{AA}p_{BB} - p_{AB}p_{BA},$$

where p_{AB} is the conditional probability for a random walk ending in side B given that it had started from side A. The other probabilities in the formula are computed similarly. The polarization score P_{RWC} takes values between -1 and 1. Fully polarized network has a P_{RWC} of 1, whereas a non-polarized network is expected to have $P_{RWC} = 0$. If the users are exposed more likely to the content produced by influencers of the opposing group, the P_{RWC} becomes negative.

As there was no general network dependent rule for choosing the parameter k in [32], we chose a single value $k = 10$ for all of the networks.

(2) Adaptive Random Walk Controversy
The Random Walk Controversy measure is very sensitive to the number of influencers k. As no strategy for selecting the parameter k based on the network was presented in the article it was defined [31], we devised such a strategy to adaptively change k depending the network based on an initial sensitivity analysis of the score. Instead of selecting a fixed number of influencers from both sides, the number of influencers in a community depends on its size. By labeling the fraction K of the highest-degree nodes as influencers from each side, i.e., by selecting $k_A = K/n_A$ for the community A and $k_B = K/n_B$ for community B with fixed K, the polarization measure P_{ARWC} scales with the number of nodes in community. We used $K = 0.01$. It should be noted that the actual values for the ARWC score (and RWC score) are sensitive to these parameter choices making comparison of results difficult if different values are used, but the qualitative behavior relative to random networks as described in the article is not sensitive to small changes in the actual parameter value K (and k for RWC).

(3) Betweenness Centrality Controversy
This measure is based on the distribution of edge betweenness centralities. If the two sides are strongly separated, then the links on the boundary are expected to have high edge betweenness centralities. The intuition is that, in highly polarized network, links connecting the opposing communities have a critical role in the network topology. The centrality c_B for each edge present in the network is defined as follows

$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)},$$

where $\sigma(s,t)$ denotes the total number of shortest paths between nodes s and t and $\sigma(s,t|e)$ is the number of those paths that include edge e.

Then KL-divergence d_{KL} is computed between the distribution of edge centralities for edges in the cut and the distribution of edge centralities for the rest of edges. The PDFs for KL are estimated by kernel density estimation. The measure seeks to quantify polarization by comparing the centralities of boundary and non-boundary links. The measure is defined as

$$P_{BCC} = 1 - e^{-d_{KL}},$$

The score approaches 1 as the level of separation increases. For networks in which the centrality values of links between two communities do not differ significantly, P_{BCC} produces values close to 0.

(4) Boundary Polarization
This measure assumes that a low concentration of high-degree nodes in the boundary of communities implies polarization. The underlying intuition is that the further some

authoritative or influential user is from the boundary, the larger is the amount of antagonism present in the network. Two sets, C and I, are defined for the score. The node $s \in A$ belongs to set C_A if and only if it is linked to at least one node of the other side ($t \in B$) and it is linked to a node $w \in A$ that is not connected to any other node of side B. For the whole network, we have $C = C_A \cup C_B$, as both sides have their own boundary nodes naturally. The non-boundary nodes are called internal nodes and are obtained by $I_A = A - C_A$. The sets of internal nodes of both communities are then combined by $I = I_A \cup I_B$.

The measure is defined as

$$P_{BP} = \frac{1}{|C|} \sum_{s \in C} \frac{d_I(s)}{d_C(s) + d_I(s)} - 0.5,$$

where d_C is the number of edges between the node s and nodes in C and d_I is the number of edges between the same node s and nodes in I. The score is normalized by the cardinality of set of boundary nodes C. The values of P_{BP} range from -0.5 to 0.5, where 0.5 indicates maximum polarization. Non-polarized network is expected to have values close to zero, whereas negative values indicate that the boundary nodes of community A are more likely to connect to the other side B than to own side.

(5) Dipole Polarization

This measure applies label propagation for quantifying the distance between the influencers of each side. Its intuition is that a network is perfectly polarized when divided in two communities of the same size and opposite opinions. First, the top-k% highest-degree nodes from both sides are selected. These nodes are assigned the "extreme opinion scores" of -1 or 1 depending on which side they belong to. For the influencer nodes, the r_t is fixed to its the extreme value for all steps t. All the other nodes begin with a neutral opinion score $r_{t=0} = 0$. The opinion scores r of the rest of nodes in the network are then updated by label propagation as follows

$$r_t(s) = \frac{\sum_v W_{sv} r_{t-1}(s)}{d(s)},$$

where $W_{sv} = 1$ if there is an edge between the nodes s and v, $r_{t-1}(s)$ is the opinion score of node s at previous step and $d(s)$ is the degree of node s. This process is repeated until the convergence of opinion scores.

Denote the average or the gravity center of positive and negative opinion scores by gc^+ and gc^-. The distance between the means of the opposite opinion score distributions is then $d = \frac{|gc^+ - gc^-|}{2}$. For the final polarization score P_{DP}, the distance d is multiplied by $(1 - \Delta a)$ to penalize the potential difference in the community sizes. The Δa can be obtained either by a) taking the difference of definite integrals of the opposite opinion score distributions or b) by computing the absolute difference of the normalized community sizes. The latter is simply obtained by $\Delta a = \frac{|n^+ - n^-|}{|A \cup B|}$, where n^+ denotes the number of nodes having a positive opinion score and n^- denotes the number of nodes having a negative opinion score.

The final polarization is calculated by

$$P_{DP} = (1 - \Delta a) d.$$

The value of P_{DP} can only be its maximum when the label-propagation based communities have equal sizes. The closer the means of opinion score distributions of both communities are, the lower polarization.

(6) Modularity

Modularity is one of the most popular scores for describing discrepancy in a social network. Modularity measures how different the communities are from the corresponding communities of the ensemble of random graphs obtained by configuration model. The polarization based on modularity is simply the formula of modularity that is used to evaluate the quality of communities.

$$P_Q = \frac{1}{2|E|} \sum_{ij} (W_{ij} - \frac{k_i k_j}{2|E|}) \delta(c_i, c_j),$$

where $|E|$ is the number of edges, W_{ij} is the element of adjacency matrix and k_i is the degree of node i. The value of $\delta(c_i, c_j)$ equals to one only when the nodes i and j belong to the same community, otherwise it is zero.

(7) E-I Index

This simple measure, also known as *Krackhardt E/I Ratio*, computes the relative density of internal connections within a community compared to the number of connections that community has externally. For two communities, it can be defined as

$$P_{EI} = \frac{|C|}{|C'|},$$

where C is the cut-set $\{(s, t) \in E | s \in A, t \in B\}$ and C' is the complement of that set ($C' = E/C$).

(8) Adaptive E-I Index

This measure is an extension of E-I Index as it accounts for different community sizes by using the density of links within each community. The Adaptive E-I Index becomes E-I Index when both of the communities have equal number of nodes. The measure is defined as

$$P_{AEI} = \frac{\sigma_{AA} + \sigma_{BB} - (\sigma_{AB} + \sigma_{BA})}{\sigma_{AA} + \sigma_{BB} + (\sigma_{AB} + \sigma_{BA})},$$

where σ_{AA} is the ratio of actual and potential links within the community A (similary for σ_{BB}) and σ_{AB} is the observed number of links between the communities A and B divided by the number of all potential links.

B ADDITIONAL FIGURES AND ANALYSIS

In this appendix, we include figures that summarize results of additional analysis. In section B.1, we include alternative illustrations and additional analysis to support our arguments. In section B.2, we include results obtained using different partitioning methods.

B.1 Alternative visualisations and additional analysis

Fig. 12 is an alternative plot for illustrating our observation in Section 4.1. Fig. 13 is an alternative plot for illustrating our observation in Section 4.2.1.

The effects of heterogeneous degree sequences on polarization measures were also studied for larger networks (Figs. 14 and 15). To see the performance stability across networks with different average degrees, see Figs. 17 and 18. Finally, the Figs. 16 and 19 show the performance of the standardized denoised polarization scores.

B.2 Results for alternative graph partitioning methods

In addition to METIS, we perfomed our analysis using two alternative clustering methods: regularized spectral clustering and modularity optimization. While the number of links between the two groups is used by METIS as an optimisation criteria, the intuition behind spectral clustering is related to finding groups where a random walker would remain in the starting cluster as long as possible. Further, modularity measures the excess fraction of links inside the groups as compared to a null model. The clusters obtained by METIS had already high modularity values. Therefore, for optimizing the modularity, we used the partition produced by METIS as the pre-partition for which a fine-tuning was performed: We then optimised the partition for maximum modularity by a greedy stochastic optimization method which consecutively tries to swap the cluster of a random node and accepts it if the value of the target function improves [44]. A reasonable convergence was achieved when the number of swaps was equal to two times the number of nodes in the network. Figures 20 and 21 display the noise bar analysis equivalent to Fig. 4 in the main text for these two additional methods, but only considering the configuration model ($d = 1$). Figures 22 and 23 are alternatives for Fig. 9 displaying the ROC curves related to the classification task presented in Sec. 5.
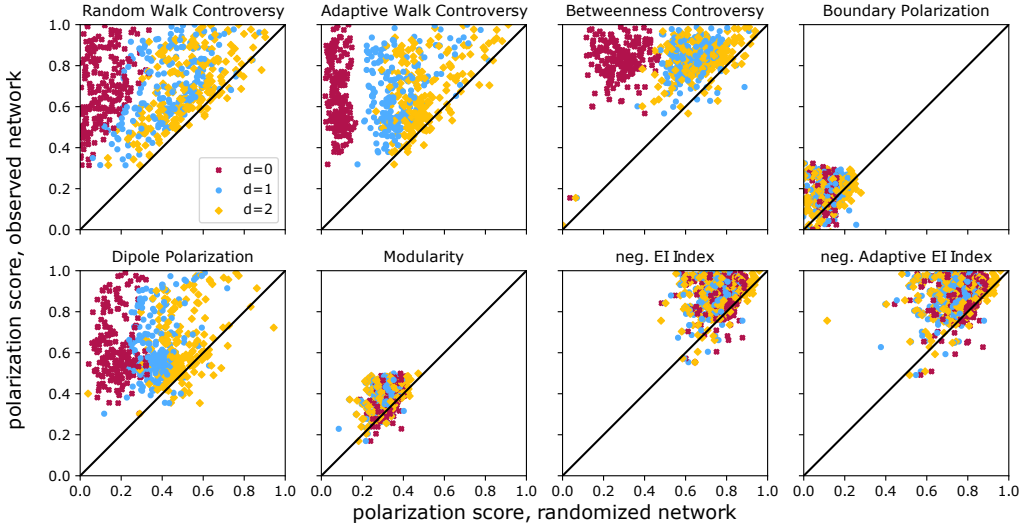
Fig. 12. The observed polarization score as a function of the score computed for a randomized version of that network. Each point corresponds to the average polarization score of 500 randomizations of the original network. The colors represent the three different random models (see the legend in upper left panel). Note that the negative values of Boundary Polarization have been left out in this visualisation. Figure contains data for all the 203 networks. Error bars for single points are ommitted because they are mostly too small to be visible.
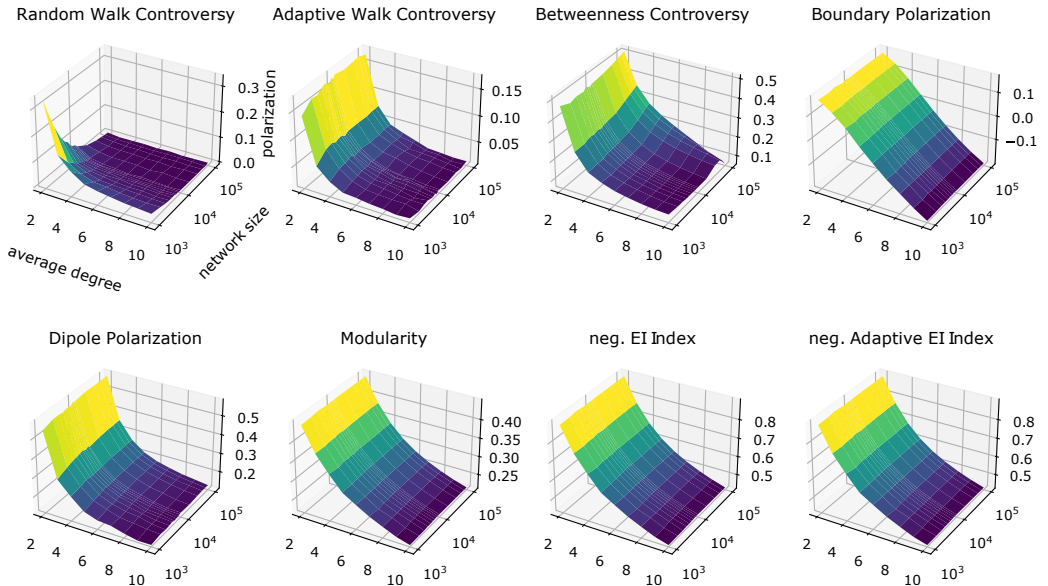
Fig. 13. 3D plots displaying the effect of network size and average degree on polarization. For ER-networks, only RWC gives lower polarization as network grows but fixed parameter is required then. Many of the scores approaches zero polarization only after becoming very dense.
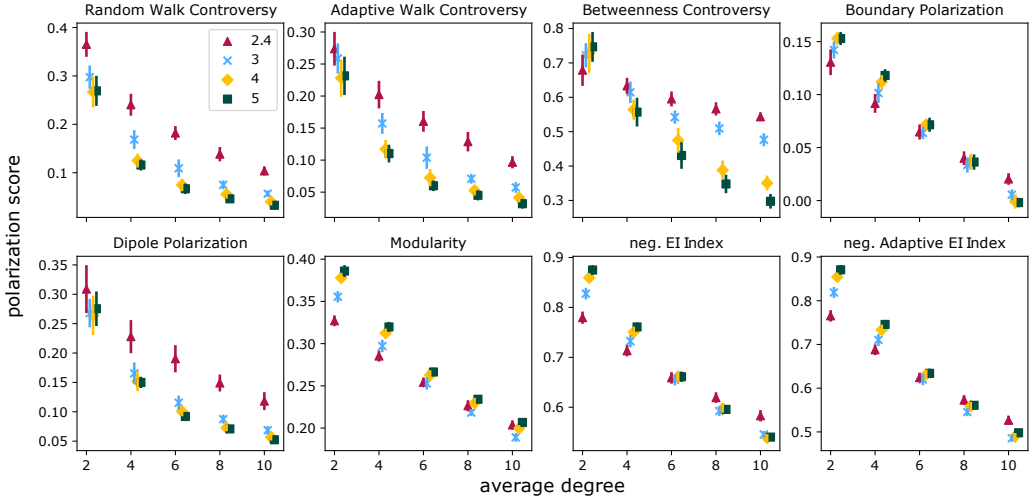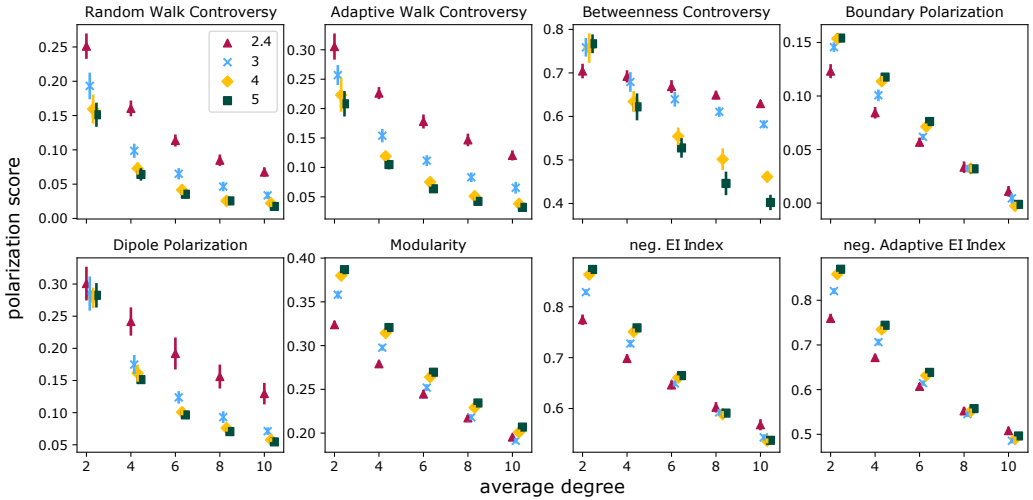
Fig. 14. The effect of degree heterogeneity to polarization scores for the 8 scores in simulated scale-free networks with 2000 nodes. We show the expected polarization score and its standard deviation as a function of average degree. The symbols correspond to different exponent γ values as indicated by the legend. The RWC values are lower for larger networks which was also shown to be the case for ER-networks in Section 4.2.1. The rest of the scores are not dependent on network size. See Fig. 7 in the main text for the network with 1000 nodes

Fig. 15. The effect of degree heterogeneity to polarization scores for the 8 scores in simulated scale-free networks with 5000 nodes. We show the expected polarization score and its standard deviation as a function of average degree. The symbols correspond to different exponent γ values as indicated by the legend. See Fig. 7 in the main text for the network with 1000 nodes.
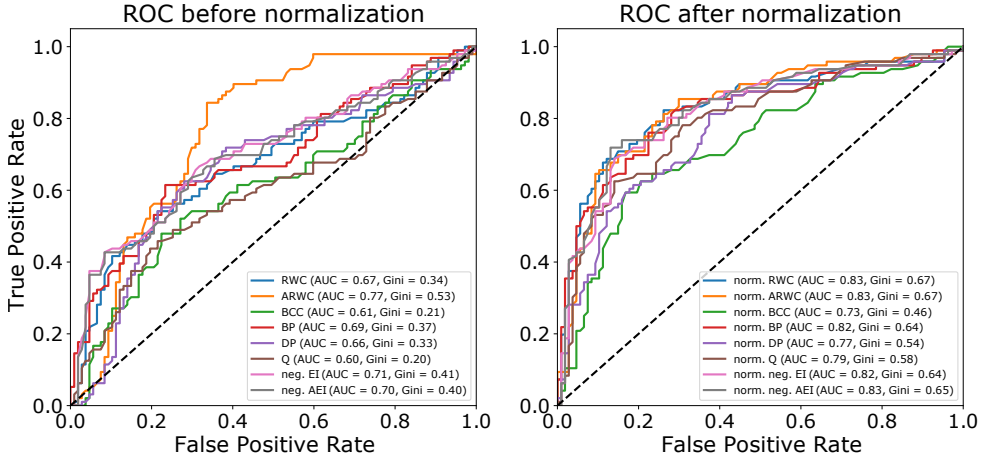
Fig. 16. ROC curves, Gini coefficient values, and AUC values for the task of predicting manually curated labeling of polarized and non-polarized networks. The results shown (left) for the score values before the normalization and (right) after the normalization with standardized and denoised scores $\hat{\Phi}_z(G)$ (See Section 5.3). Figures are based on all the 203 empirical networks described in section 3.1. The ROC curves for denoised polarization scores are in Fig. 9.
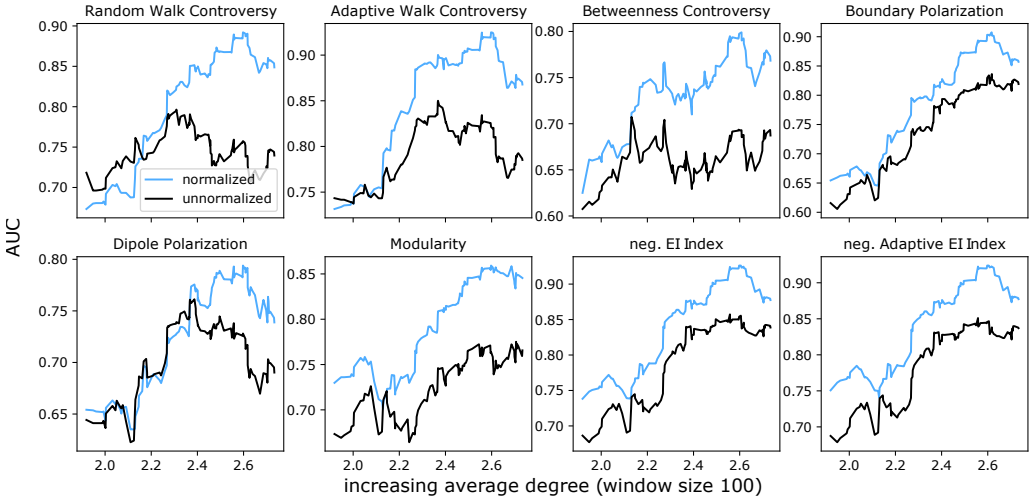

Fig. 17. Quantifying how network's average degree affects the performance. We group the data such that there are 100 networks with consecutive degrees in our data, and create a set of such windows by varying the degree range. We then evaluate the AUC for the moving window of 100 networks. The plots show how the performances of both normalized and unnormalized scores become higher as the average degree increases for all the polarization methods. However, the normalization still improves the overall accuracy, especially for the less sparse networks. For instance, the normalization of RWC score improves the AUC approximately 0.10 units for networks with an average degree of 2.4 or higher.
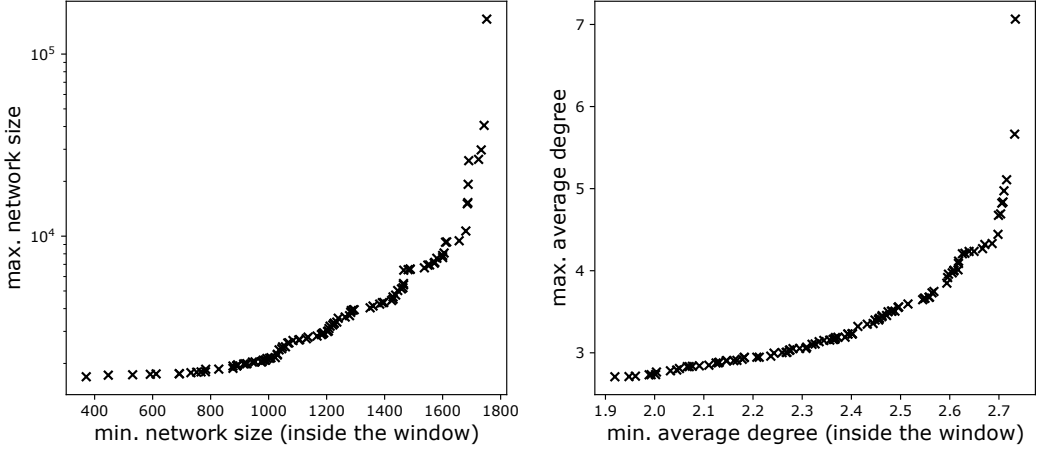
Fig. 18. Additional information about the windows for Fig. 11 and Fig. 17. The smallest value of the window is on x-axis and, respectively, the largest value of the window is located on y-axis.
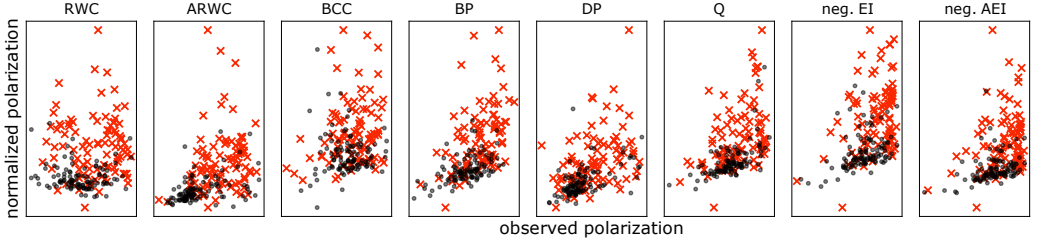
.

Fig. 19. Similar scatter plot as in Fig. 10 but for denoised polarization scores with standardization. Red crosses are networks labeled as polarized and black points are networks labeled as non-polarized. An outlier was removed from the plots. Note that the scales for the scores are different and not shown in this figure.
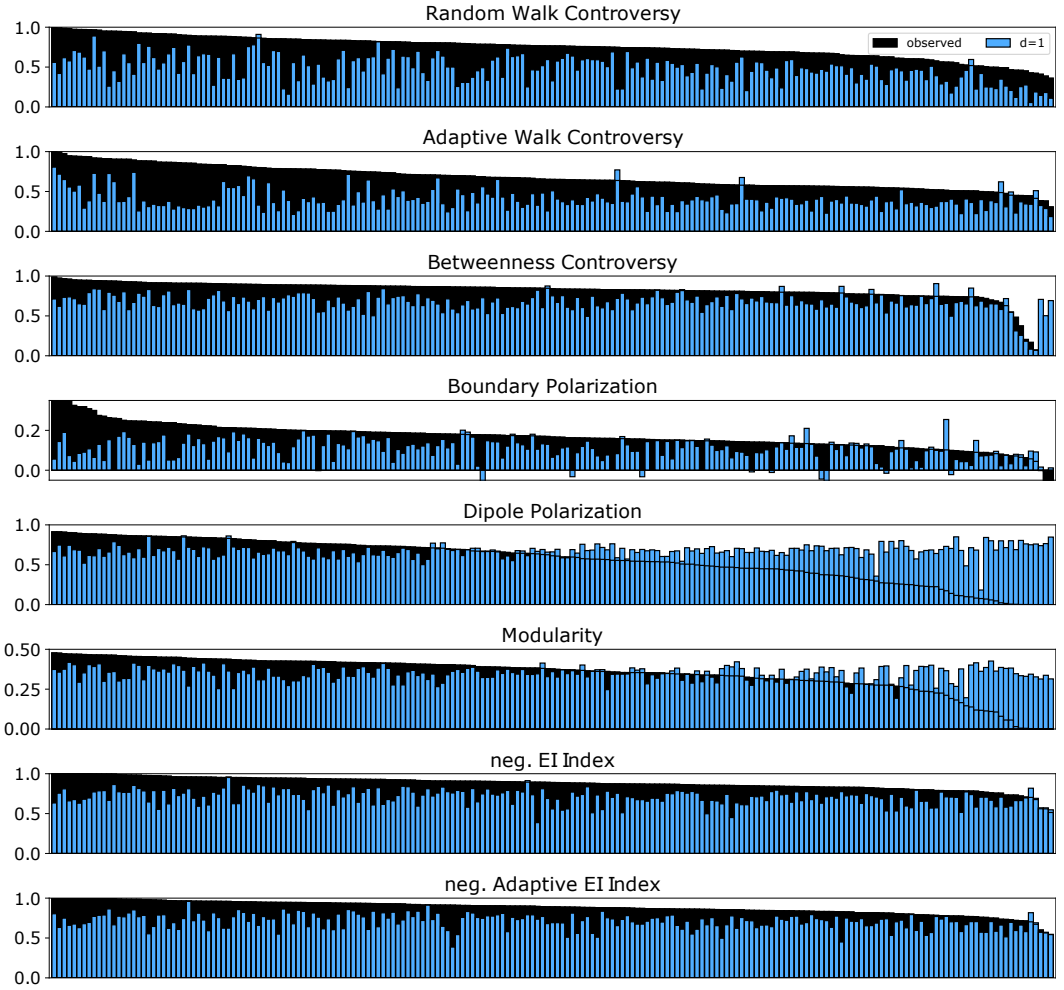
Fig. 20. Polarization scores for the 203 observed networks and their shuffled versions. Each bar corresponds a score, and scores for a network and its randomized versions are on top of each other: observed network is represented with black bars and scores competed for random networks where degree sequence is preserved ($d = 1$) are shown in blue. An interpretation for the figure is that, the amount of blue that is shown tells how much of the total bar height (the score value) is explained by the degree distribution and the amount of black that is shown is not explained by it. Note that in some cases, the randomized networks produce higher scores than the original network and in this case the black bar is fully covered by the blue bar. In this case we draw a black horizontal line on top of the blue bar indicating the height of the black bar. The difference of this figure to Fig. 4 in the main text is that groups in this figure are produced with spectral clustering and only the null model for degree sequence is shown.
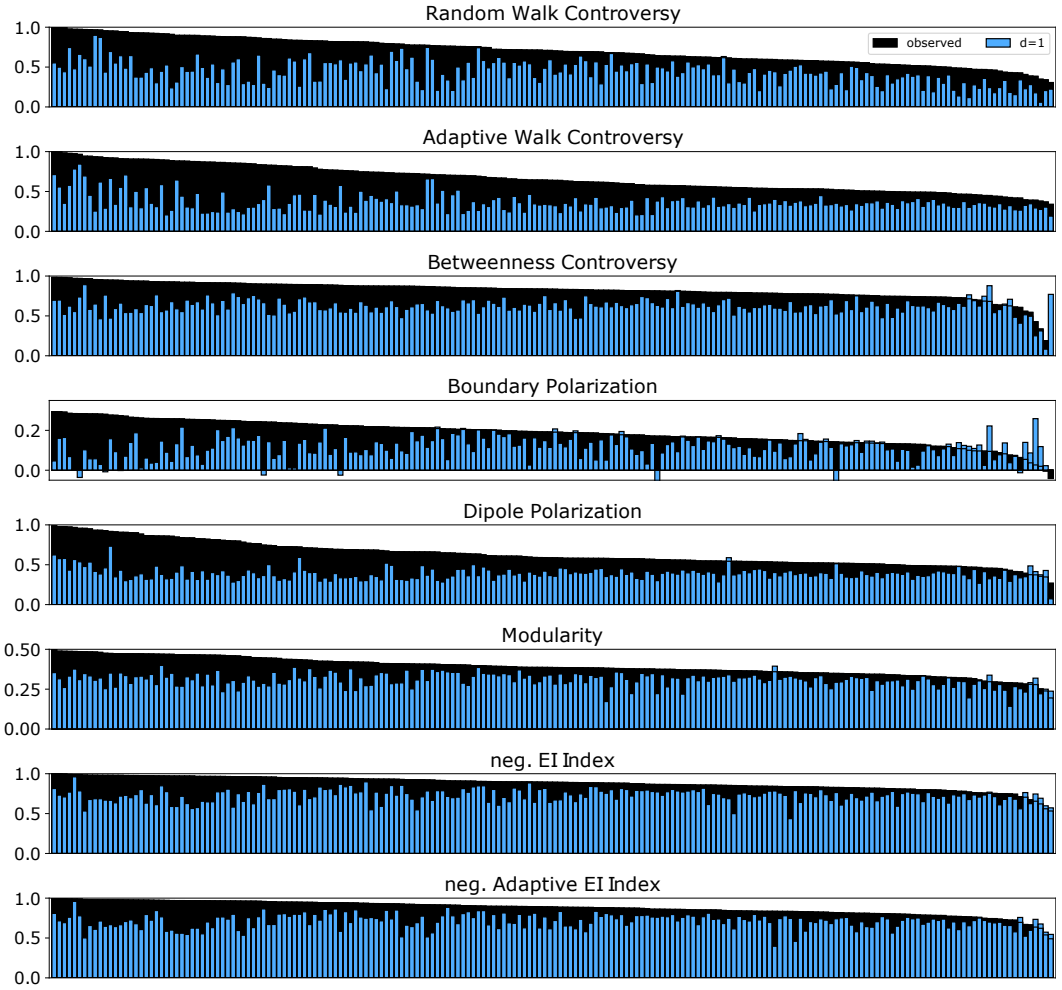
Fig. 21. Polarization scores for the 203 observed networks and their shuffled versions. Each bar corresponds a score, and scores for a network and its randomized versions are on top of each other: observed network is represented with black bars and scores competed for random networks where degree sequence is preserved ($d = 1$) are shown in blue. An interpretation for the figure is that, the amount of blue that is shown tells how much of the total bar height (the score value) is explained by the degree distribution and the amount of black that is shown is not explained by it. Note that in some cases, the randomized networks produce higher scores than the original network and in this case the black bar is fully covered by the blue bar. In this case we draw a black horizontal line on top of the blue bar indicating the height of the black bar. The difference of this figure to Fig. 4 in the main text is that groups in this figure are fine tuned with modularity optimization and only the null model for degree sequence is shown.
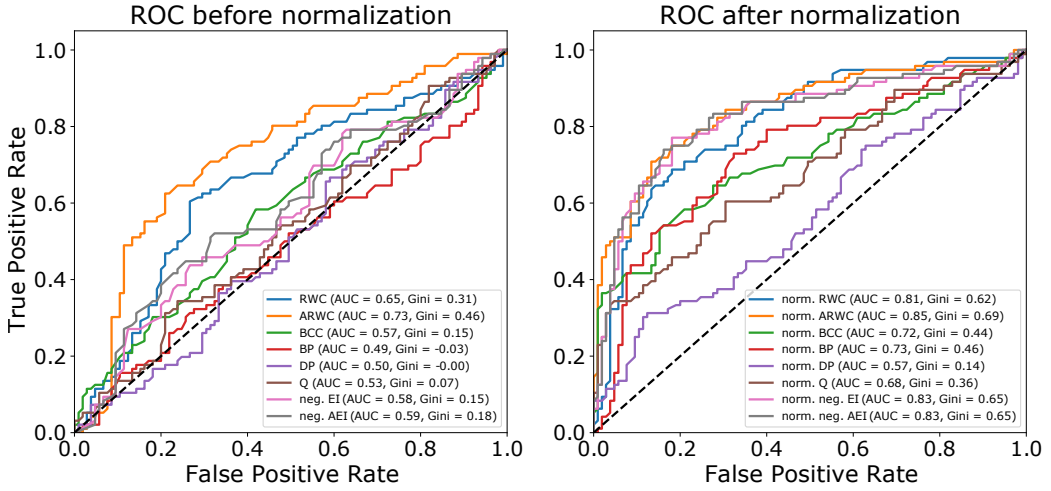
Fig. 22. ROC curves, Gini coefficient values, and AUC values for the task of predicting manually curated labeling of polarized and non-polarized networks. The results shown (left) for the score values before the normalization and (right) after the normalization with denoised scores $\hat{\Phi}(G)$ (see main text). Figures are based on all the 203 empirical networks described in section 3.1. The difference of this figure to Fig. 9 in the main text is that groups in this figure are produced with spectral clustering.
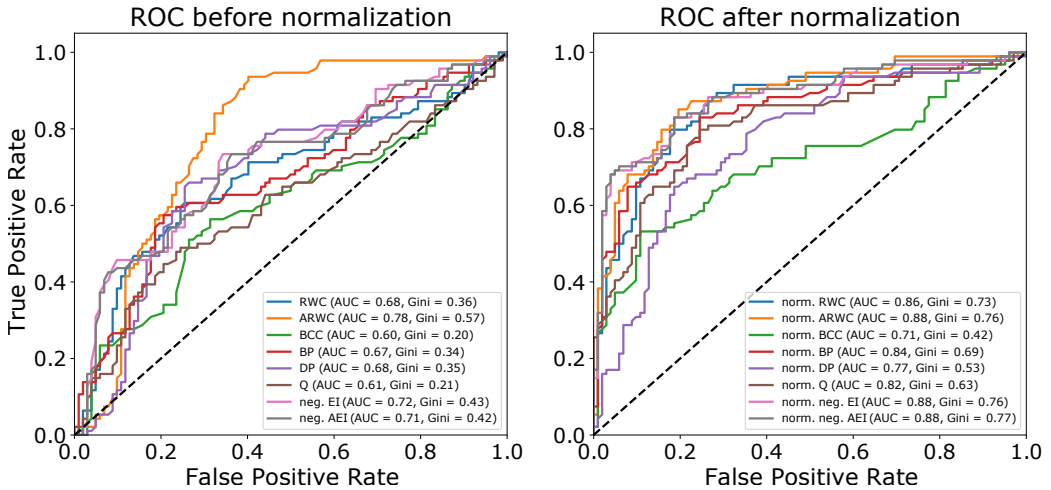

Fig. 23. ROC curves, Gini coefficient values, and AUC values for the task of predicting manually curated labeling of polarized and non-polarized networks. The results shown (left) for the score values before the normalization and (right) after the normalization with denoised scores $\hat{\Phi}(G)$ (see main text). Figures are based on all the 203 empirical networks described in section 3.1. The difference of this figure to Fig. 9 in the main text is that groups in this figure are fine tuned with modularity optimisation.