Supplementary Material — A general model for how attributes can reduce polarization in social groups

Piotr J Górski, Curtis Atkisson, Janusz A Hołyst

May 18, 2023

1 Introduction

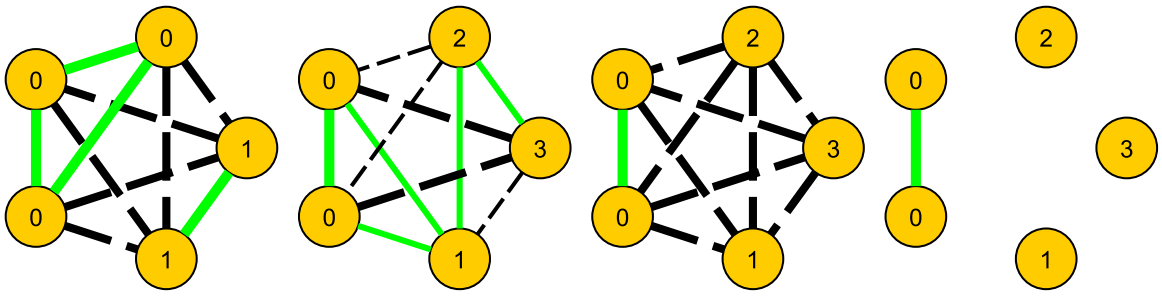
In this document, we present additional analysis that either confirms some of the statements from the main paper or may enhance understanding of the proposed framework.

Section 2 presents an overview how the attribute layer for the considered attributes may look like. Section 3 visualizes the three real-world network structures that were analyzed in the main paper. Section 4 introduces additional measures that allow to assess the influence of attributes on structural balance.

Section 5 presents additional results. Section 5.1 describes the derivation of Eq. (8) from the main paper and analyzes the destabilization issue from the point of view of reaching a structurally unbalanced state. Such an approach agrees with the line of research in some other papers (cited in the main text) where reaching a balanced state was studied in a system with agents possessing attributes. Section 5.2 provides the equation used to calculate exact, analytical local polarization values for the case of very high coupling γ for binary attributes. Section 5.3 gives numerical results for other measures listed above in section 4. Section 5.4 presents additional results related to analysis of attribute impact in the case of non-complete network structures.

2 Attributes layers for considered attribute types

Fig. 1 shows example attribute layers for binary attributes (BAs), ordered attributes (OAs), negative unordered attributes (NUAs) and positive unordered attributes (PUAs) in the case of agents possessing one attribute ($G = 1$).



(a) BA (b) OA, $v = 4$ (c) NUA, $v = 4$ (d) PUA, $v = 4$

Figure 1: Examples of weights $g_{ij} = h_{ij}$ in the attribute layer for considered types of attributes with $G = 1$. The solid green line and the dashed black line indicate positive and negative weights, respectively. All edges have maximal weight values (± 1) for BA and NUA. The edges with a smaller line thickness (for OA) symbolize a lower weight in terms of the absolute value. No edge (for PUA) means a weight of 0.

3 Considered real-world network structures

Network structures of real-world datasets are visualized in Fig. 2. In the high school network, one cannot distinguish separate groups. With a few exceptions, the structure resembles a complete graph. Two distinct groups are perfectly visible for Zachary karate club (ZKC) network. There are not many edges in total and inter-group connections are very rare. Using Fruchterman-Reingold force-directed algorithm for network visualization, groups of old members and newcomers seem not to be so well separated for the Windsurfers network. The reason is that there are more connections across the groups.

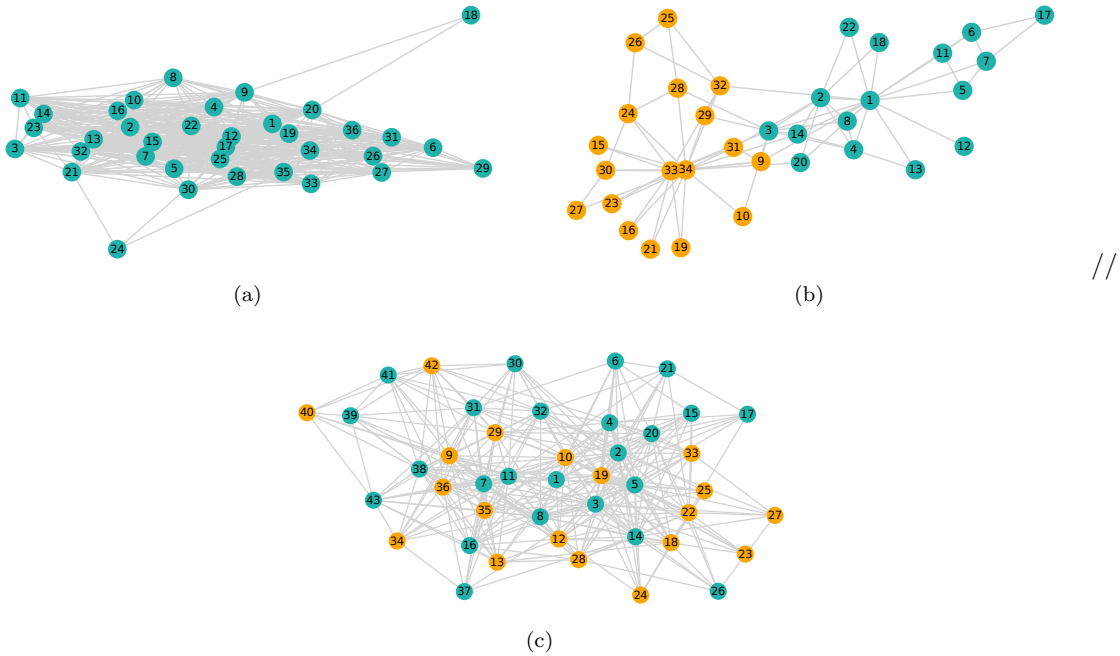


Figure 2: Visualizations of non-complete network structures used in simulations. Panels show the networks of (a) class 2BIO1 of the Highschool13 dataset, (b) Zachary karate club dataset, and (c) Windsurfers dataset. Nodes represent people, and links—registered relations or interactions. Node colors in panels (b) and (c) represent observed groups.

4 Assessing system's fate. Measures of structural balance and polarization

The specified system (with chosen attribute types, number of attributes v and attribute strength γ) evolves from the initial conditions until it reaches a stable point. In such a point (almost) all relations become $x_{ij} = \pm 1$. The rest of the relations (if there are any) reach such values that $\frac{1}{M_{ij}} \sum_{k \in CN_{ij}} x_{ik} x_{kj} + \gamma g_{ij}(\mathbf{A}) = 0$. It is more likely to have such relations when the links participate in an even number of triads (a probable possibility for incomplete graphs or when a complete graph has an even number of agents). Stable values of relations different from ± 1 might also be an artifact of the specific value of coupling strength γ. For instance, when for agents i and j, the influence of their common neighbors is $\frac{1}{M_{ij}} \sum_{k \in CN_{ij}} x_{ik} x_{kj} = +1$, their similarity influence $g_{ij} = -1/3$, then having coupling strength $\gamma = 3$ exactly would make attainable for $\left(\frac{1}{M_{ij}} \sum_{k \in CN_{ij}} x_{ik} x_{kj} + \gamma g_{ij}(\mathbf{A}) \right)$ to become 0. In such a case it might be beneficial to consider not the integer values of coupling strength but real values very close to desirable ones or γ values could be drawn from some distribution. The latter method was not used in this paper.

Apart from estimating the level of polarization the system is in, we were interested in determining the level of system's structural balance (SB). Therefore for each set of parameters we estimated following measures:

- probability of reaching structural balance P_{SB}. This is the probability of reaching a paradise state (all links positive) or a state where it is possible to divide the network into two enemy groups ($K = 2$). All links within a group are positive and all links connecting agents from different groups are negative. Such definition is, in fact, a strong structural balance definition.
- probability of reaching weak structural balance P_{WSB}. As mentioned in the main text, in some data sets the structural balance notion "enemy of my enemy is my friend" is questioned to be true. Removing this assumption we obtain a weak version of structural balance. From a system point of view a weakly balanced network can be divided into K enemy groups (e.g. $K = 3$).
- probability of paradise P_P.

In the following part of this text, structural balance is understood primarily as a balanced system in the strong sense. Each time the discussion concerns a weak understanding of SB, it is additionally emphasized. Having above definitions we can reformulate the definition of global polarization. The global polarization probability P_{GP} is:

$$P_{GP} = P_{WSB} - P_P, \tag{1}$$

5 Additional results

5.1 Destabilization of the balanced state

Eq. (8) in the main text can be obtained by deriving the Jacobian matrix of system dynamics. In the proximity of stable (i.e., balanced) states the matrix is diagonal with following elements:

$$\frac{\partial \dot{x}_{ij}}{\partial x_{ij}} = (1 - 2x_{ij})\Big(\frac{1}{N-2}\sum_{k=1}^{N} x_{ik}x_{kj} + \gamma g_{ij}(\mathbf{A})\Big). \tag{2}$$

The average influence of neighbors in relation layer, i.e., $\overline{\Sigma}_{ij} \equiv \frac{1}{N-2}\sum_{k=1}^{N} x_{ik}x_{kj}$ in the balanced state is equal exactly $+1$ or -1. In the balanced case we have a division into two groups (one group can be empty). When the link x_{ij} is positive, then the influence $\overline{\Sigma}_{ij} = +1$ because if node k is (un)friendly with node i, it is also (un)friendly with node j, thus $x_{ik}x_{jk} = 1$. Similarly, for the negative link x_{ij} the influence $\overline{\Sigma}_{ij} = -1$. Thus, one can write:

$$\dot{x}_{ij}^{\pm} \propto \pm 1 + \gamma g_{ij}(\mathbf{A}), \tag{3}$$

where the sign is marked with an appropriate index. Therefore, one can easily obtain the condition for destabilizing a link [Eq. (8) in the main paper].

The destabilized edge in RL will take the opposite sign after some time. As a result of changing one link, the network will be destabilized. However, this single change may induce further edge changes so that a different balanced state is reached. In order to ensure that the end state of the network will not be in structural balance, there must be a triad whose edges will form a state with one or three negative links. In the case of the $N = 3$ network, the sufficient condition to obtain an unbalanced state is the fulfillment of the following inequalities:

$$\begin{cases} \gamma^3 g_{12}g_{23}g_{13} < 0 \\ |\gamma g_{12}| > 1 \\ |\gamma g_{23}| > 1 \\ |\gamma g_{13}| > 1 \end{cases} \tag{4}$$

The first of the conditions means that the triad in AL is unbalanced (for $\gamma > 0$), and the next ones mean that the influence of each similarity edge is stronger than the influence of other relations in the system. For

an arbitrarily large network to become unbalanced, it is enough for there to be at least one triad for which the inequalities are satisfied. The described conditions are sufficient. However, for a larger network, they are not necessary. When the first condition is met, the following conditions will be met with a sufficiently high strength γ of attributes. However, in a special case, the destabilization of one link may lead to a situation that no other edge will change so that the entire system will become unbalanced.

5.2 Very high strength of attribute layer with binary attributes

Let us assume that the influence of the attribute layer on the relation layer is very strong, i.e., $\gamma \to \infty$. Assuming the weights in the attribute layer are nonzero, the attribute layer will set all the signs of weights in the relation layer such that $sgn(x_{ij}) = sgn(g_{ij})$. Thus in such a case, the analysis of properties of the attribute layer weights gives insight into the properties of the obtained relation layer. Having a system of agents possessing G binary attribute each, one can calculate the probability of drawing different triad types using combinatorics. Let us denote N_k as the number of possible triads having k negative links.

Assuming that the attributes of one of the agents are set, the number of possible triads with three negative links is as follows:

$$N_3 = \sum_{k=1}^{(G-1)/2} \binom{G}{(G-1)/2+k} \sum_{j=k}^{(G-1)/2} \sum_{i=0}^{\lfloor j/2-k/2 \rfloor} \binom{(G+1)/2-k}{i} \binom{(G-1)/2+k}{j-i} \tag{5}$$

Thus, the probability of drawing such a triad with random attributes is:

$$n_3 = \frac{N_3}{2^{2G}} \tag{6}$$

It is important to prove the following relation: $N_1 = 3N_3$, meaning that there are 3 times more triads with one negative link than triads with three negative links. Or in other words, it is 3 times more likely to draw a triad with one negative link. The proof is as follows. Let us assume that there are 3 agents (x, y, z) forming a triad with 3 sets of attributes: $\{a_x^1, ..., a_x^G\}$, $\{a_y^1, ..., a_y^G\}$ and $\{a_z^1, ..., a_z^G\}$; and all the similarities connecting the agents are negative. This means that each pair of agents have more attributes different than the same. Now, let us consider an *alter ego* of agent x, denoted as x', such that all attributes of x and x' are different. In other words the similarity $g_{xx'} = -1$. When odd number of attributes is considered, the signs of similarities $g_{yx'}$ and $g_{zx'}$ are positive, because more than half of the attributes are the same. Therefore, the triad $(x'yz)$ is of the type with one negative link. Similarly, one can consider alter egos of agents y and z. Therefore, for each triad of type with three negative links, there are three triads of type with one link negative. This concludes the proof.

Similar relation can be proven for triads with 0 and 2 negative links: $N_2 = 3N_0$. Thus, obtaining the number N_3 is enough to calculate the numbers (and corresponding probabilities) of all different triad types. Therefore, one can calculate the local polarization measure of attribute layer:

$$P_{LP}(\gamma \to \infty | BA) = \frac{3}{4} - 2n_3 \tag{7}$$

Above equations were used to generate $\gamma \to \infty$ curve for binary attributes in Fig. 3 in the main text. This exact result was verified by comparing the probabilities to the result obtained from Monte Carlo analysis (not shown here).

5.3 Impact of various types of attributes on destabilization and preventing from forming balanced or polarized states

Analysis of the influence of the number of attributes G.

A structurally balanced network may consist of triads having 0 or 2 negative links only. It can be calculated [Eq. (8)] that on average for a complete network of $N = 9$, density of triads with 2 negative links is 0.75. That is the level of local polarization in the case of no attributes' influence in the results from the main text ($\gamma < 1$ in scenario A and $\gamma \approx 0$ in scenario B).

$$n_2 = \frac{1}{\binom{N}{3}^{-1} 2^N} \sum_k \binom{N}{k} \left[\binom{k}{2}(N-k) + \binom{N-k}{2} k \right] \tag{8}$$

Fig. 3 shows the results of numerical calculations for the network $N = 9$ in the case of an increase in the number of attributes G of a given type.

Global polarization probability, shown in Fig. 3c-d, gives similar results to the local polarization metric presented in the main text. The initial network in scenario A or the outcome of the system without the attributes in scenario B is (almost) always a polarized system. Adding some attributes of any type is very efficient in destabilizing (assuming sufficient coupling strength $\gamma > 1$) or preventing polarization, that is the state with perfectly antagonistic groups is not usually observed. However, with more attributes added the mean value of attribute type similarity function $\mathrm{E}[h]$ starts to play a role as well. For scenario A (Fig. 3c) we observe the different outcome when number of attributes G grows depending whether the coupling γ is below or above the threshold $\hat{\gamma}_{th}$. When $1 < \gamma < \hat{\gamma}_{th}$, then after initial drop polarization grows to the level as without the attributes. Such a return does not occur if $\gamma > \hat{\gamma}_{th}$. As explained in the main text, such behavior is not observed in scenario B.

We clearly see that destabilization of the system with NUAs is related to reaching the state of many antagonistic groups. This is visible due to the lack of structural balance in Fig. 3a and the probability of a weak SB that is growing to 1 in Fig. 3b. However, as shown in Fig. 3c-d few negative unordered attributes can considerably lower global polarization probability. For this attribute type in the plot for scenario A, the effect of $\hat{\gamma}_{th}$ is not pronounced so well. The reason is that the consequence of increasing the number of attributes for both cases (either $\gamma < \hat{\gamma}_{th}$ or $\gamma > \hat{\gamma}_{th}$ is the increase of global polarization probability. The reasons are, however, different. In the first case, it is because of vanishing influence of attribute layer, therefore initial state is not destabilized. In the second case, the state of hell is reached.

Analysis of the influence of the number of categories v.

The dependence of the analyzed measures on the number of categories is shown in Fig. 4. The same values of measures are reached for ordered or negative unordered attributes with $v = 2$ which confirms that they are the same type, i.e., binary attributes. For an OA with plenty of categories we obtain an approximately continuous attribute (CA). The exact number of categories necessary for an OA not to be different from a CA may be dependent on network size N. For $N = 9$, $v = 100$ is sufficient. Therefore, the applied, in the main text, number of $v = 1000$ categories is reasonable.

Fig. 4 allows an easy comparison between binary and continuous attributes. Observing the curves for ordered attributes there is a transition between binary and approximately continuous cases. We observe that binary attributes are better at destabilizing a balanced state than continuous attributes (see panel a). Such observation can be also made in other figures in this document, e.g., Fig. 3a, Fig. 5a and Fig. 6a-b.

For NUA and PUA, with a large number of categories, the variance decreases to 0 (see Table 2 in the main text). Then all similarity weights become the same value, equal to $\mathrm{E}[h]$. For a continuous measure, which P_{LP} is, this results in a quick achievement of a constant level (state of hell for NUA and the same level, as for the state without attributes for PUA). For discrete measures (i.e., probabilities of structural balance or global polarization) in the case of PUA, the convergence is slower. The global polarization values can be determined precisely in scenario A for a large number of categories. For example, for a system ($N = 9$, $G = 5$, $v = 1000$) the attributes are irrelevant in about 83.5% of cases[1]. In other cases, there is a pair of agents that have one identical attribute (remaining cases have a negligible chance of occurrence). For such a pair, the weight in AL is 0.2 so $\gamma = 6$ can destabilize the negative edge. This edge is negative with a probability of 0.5. Thus, destabilization will occur with a probability of around 0.0825. Then one link changes its sign without causing further propagation of changes. This leads to an unbalanced system with at least one triad with one negative link (because when changing the sign of the negative edge, at least one of the triads with two negative links is affected). This reasoning agrees with the value obtained in Fig. 4d: 0.914 (as $1 - 0.914 = 0.086$). This value is also the same for probability P_{SB}, which is not shown here.

Analysis of the strength γ of the influence of the attribute layer.

The influence of the strength of attribute layer is shown in Fig. 5. The destabilization and preventing of a structurally balanced system occurs most effectively (in the order according to the lowest required strength) for NUA, BA, and OA (panel a). CA slightly lowers the probability P_{SB} and PUA does not lower at all for wide range of coupling γ. In terms of system polarization, NUA can reduce the system polarization (see Fig. 5b), especially for not too strong coupling in the range of attribute and relation layer weights' interplay. Such

[1] Calculated from $\left(\frac{\binom{1000}{9}9!}{1000^9}\right)^5$.

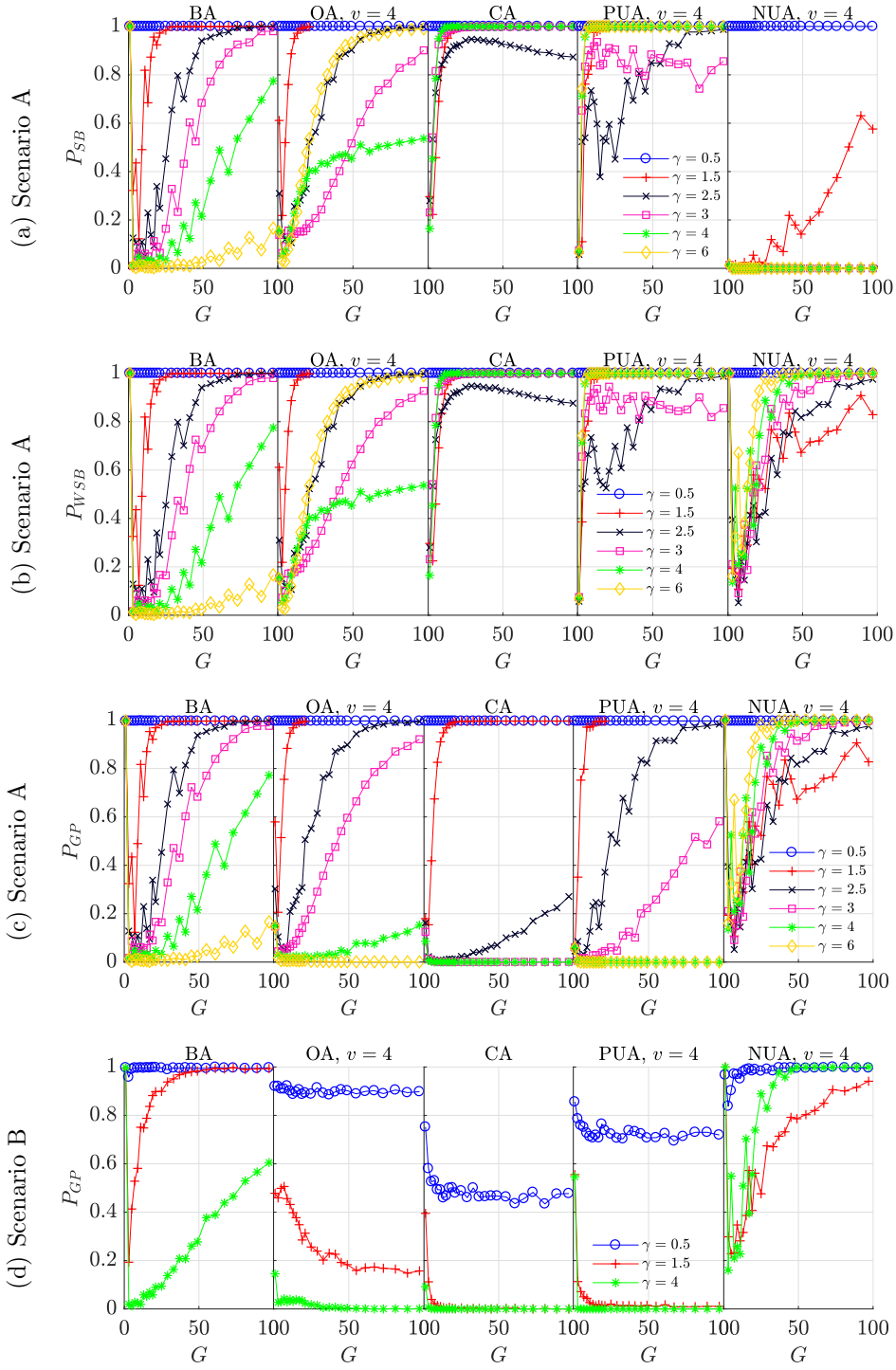


Figure 3: Impact of the growing number of attributes G on the destabilization (scenario A) and preventing (scenario B) from forming balanced and polarized states. The panels show the probabilities of (a) structural balance P_{SB}, (b) weak structural balance P_{WSB}, and (c-d) global polarization measure P_{GP} for complete graph networks of size $N = 9$ for different types of attributes and γ coupling strengths. Destabilization of polarized state is analyzed in panels (a-c) and prevention is analyzed in panel (d).

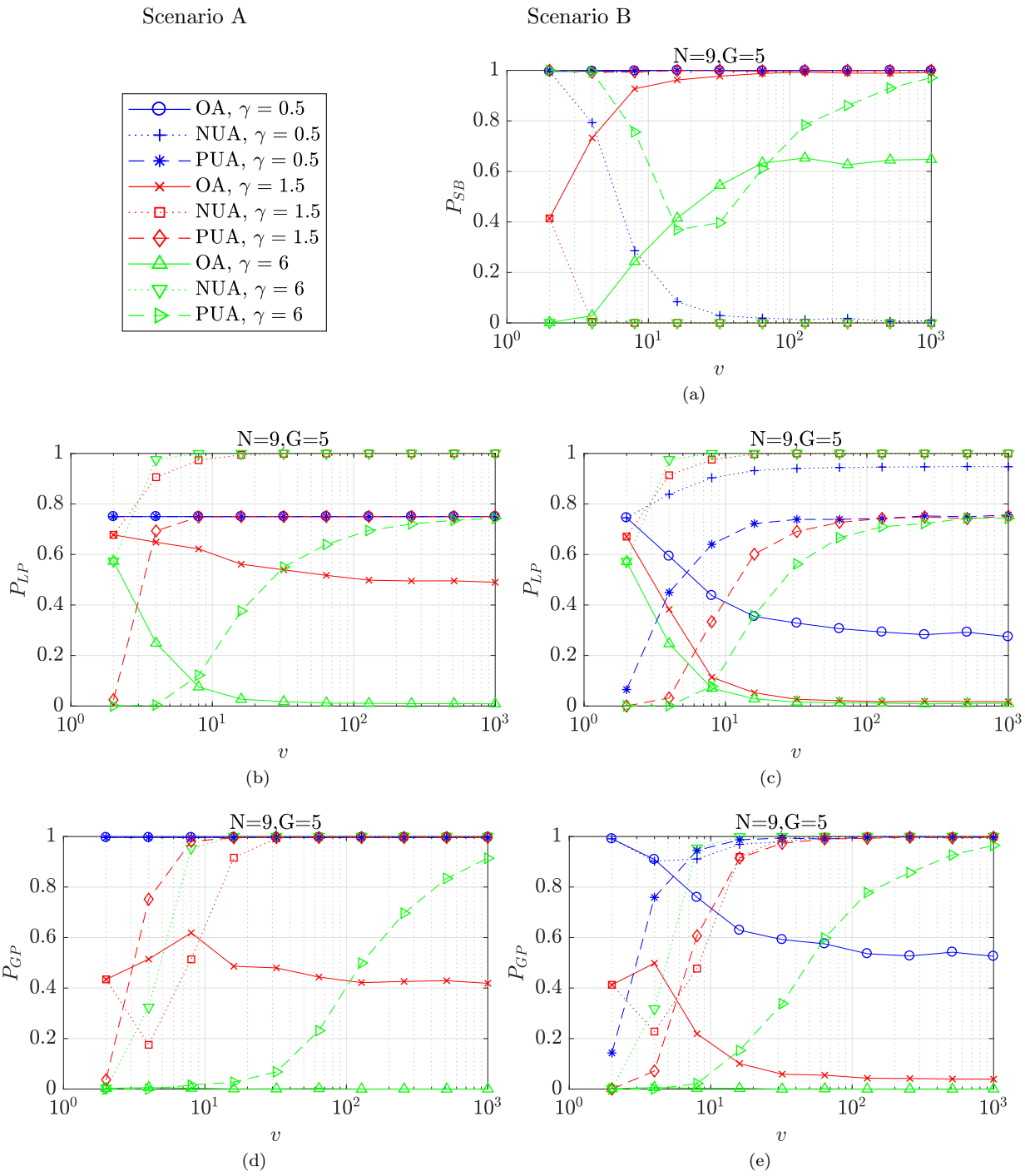


Figure 4: Impact of an increasing number of categories on destabilization (left column) and preventing (right column) from forming balanced and polarized states. The following rows show the probability of reaching structural balance P_{SB} and the local P_{LP} and global P_{GP} polarization measures for complete graph networks of size $N = 9$ for different types of attributes and γ coupling strengths.

interplay is also visible for CA and PUA in Fig. 5a.

By looking at Fig. 5b (and Fig. 4 in the main text) it may seem that having 5 continuous attributes one reaches a paradise state. Fig. 5a shows that it is not true because the probability of structural balance decreases with more influential attribute layer. The reason is that in such a network negative similarity links are still likely. When the coupling γ is high enough, then relation and attribute layer signs become almost the same. Some negative links may occur in relation layer which leads to the formation of triads with one negative link. Such triads are unbalanced and that is why probability P_{SB} is smaller than 1. This is not the case for positive unordered attributes, for which all the similarities are non-negative and the triads with one negative link are not formed.

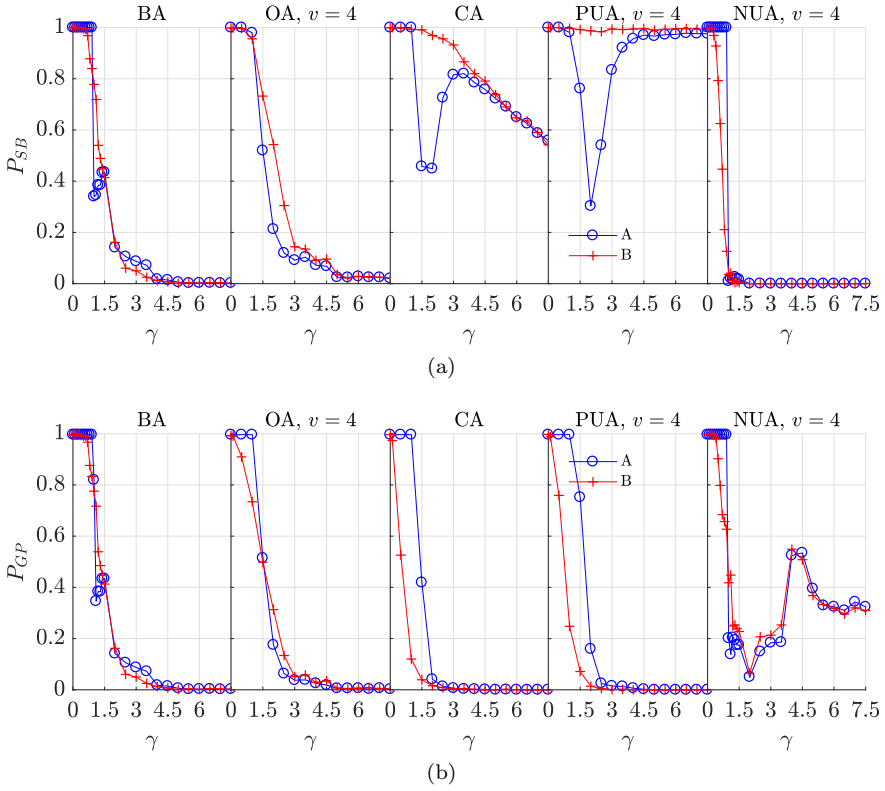


Figure 5: Influence of increasing strength of the attribute layer on destabilization and preventing (respectively A and B in the legends) from forming balanced and polarized states. The panels show (a) the probability of reaching structural balance P_{SB}, and (b) the global polarization measure P_{GP} for complete graphs of size $N = 9$ for $G = 5$ for different types of attributes.

Analysis of the influence of the number of agents N.

Fig. 6 shows the dependence of the influence of five attributes of a given type on complete graph systems with an increasing number of agents. The larger the system, the smaller the chance of achieving paradise in a network without attributes, which can be seen by the increase in global polarization for small γ in Fig. 6c-d. The curves on the BA charts (i.e. OA for $v = 2$), OA for $v = 4$ charts and CA charts (i.e. OA for $v = 1000$) change monotonically with the increasing number of categories. The larger v, the more P_{SB} grows and the smaller the polarization of the system.

5.4 Influence of model parameters on destabilization and preventing in the case of real-world network structures

In this section, we present additional results regarding simulations on real-world network structures. Fig. 7 shows how the increase of the number of attributes affects the change of the local polarization for different types and strengths of attributes when the underlying network structure is the high school network.

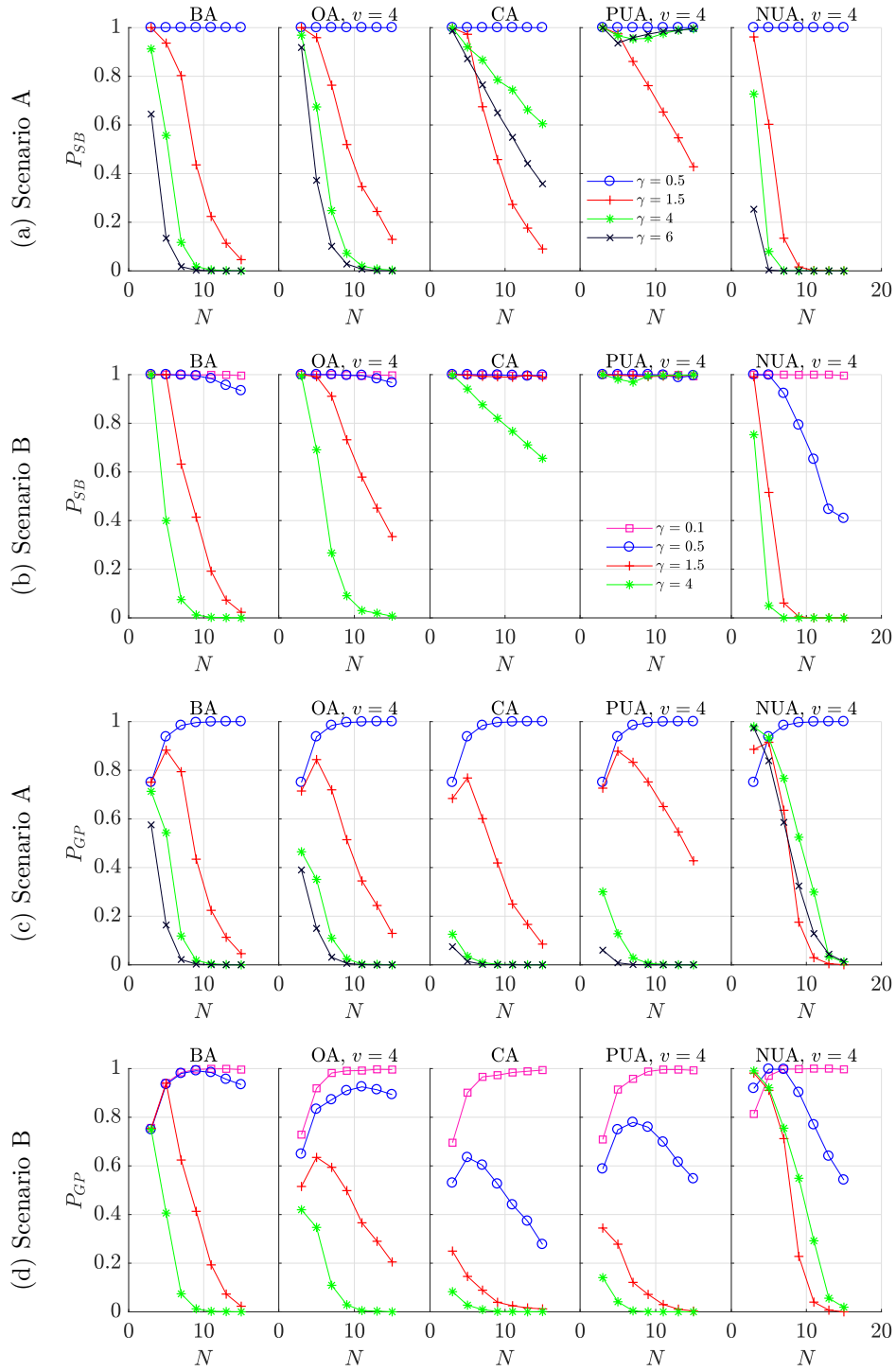


Figure 6: Influence of $G = 5$ attributes of the same type on destabilization and preventing in complete graph systems with an increasing number of nodes N. The panels show (a-b) the probability of reaching structural balance P_{SB}, and (c-d) the global polarization measure P_{GP}. The legends for panels (c-d) are presented in panels (a-b), respectively.

Comparison between this figure and the one for the complete graph structure (Fig. 3 in the main text) reveals no differences. Thus, we conclude that a complete graph is a good approximation of a structure of relations existing in the high school class network.

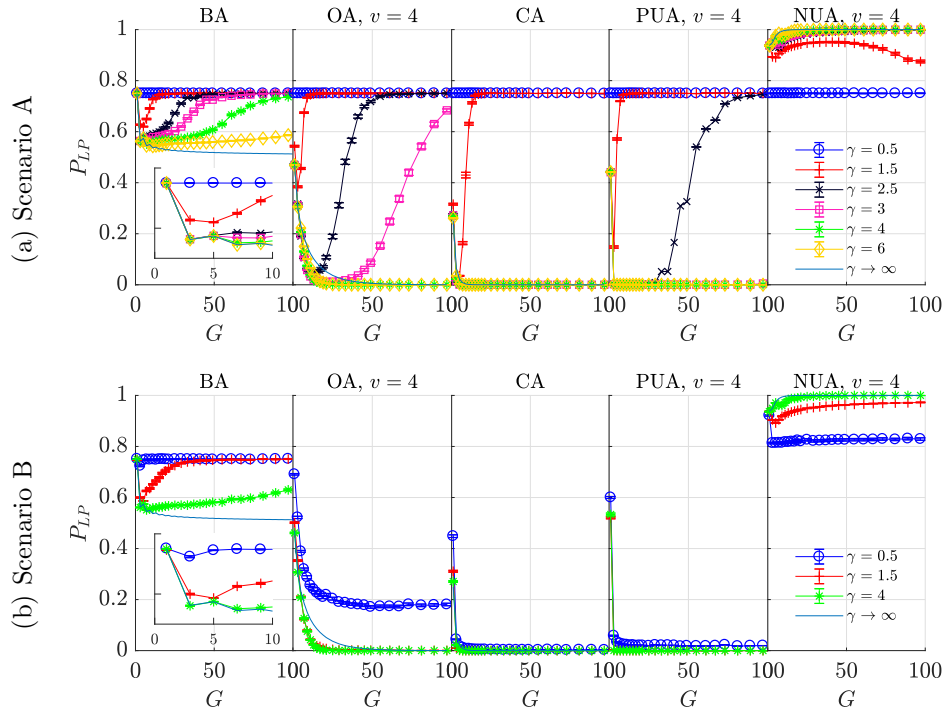
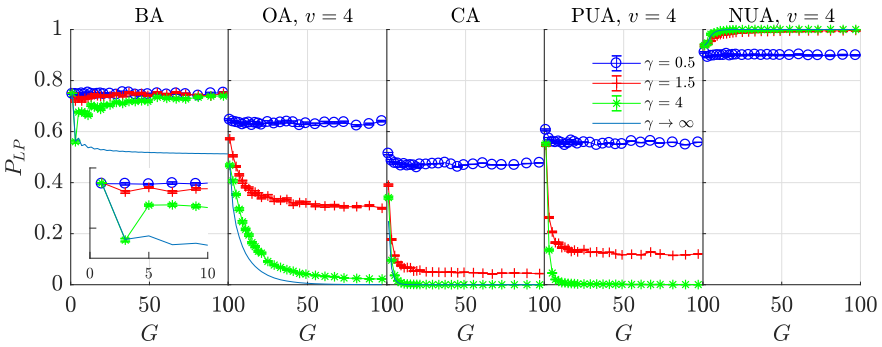


Figure 7: Impact of growing number of attributes G on destabilization (scenario A) and preventing (scenario B) from forming balanced and polarized states in the case of high school class network structure. The panels show the local polarization measure for different types and strengths γ of the attributes. There are almost no differences between this result and analogous results for the complete graph shown in Fig. 3 in the main paper.

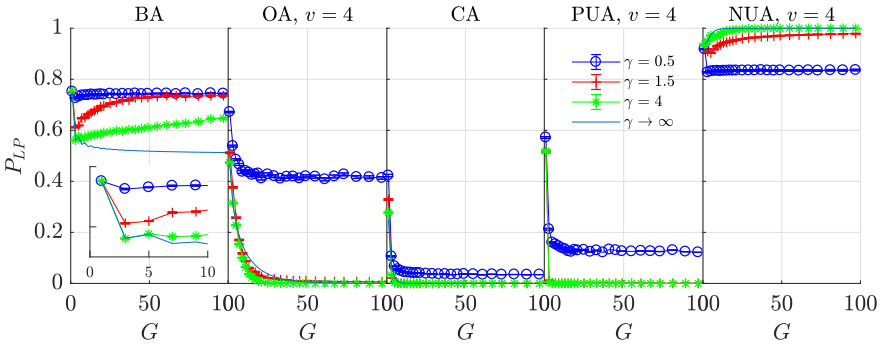
For the scenario of destabilization, community structure significantly affects the results of local polarization. For the scenario of preventing, such differences are not observed. Fig. 7b for the high school network and Fig. 8 for ZKC and Windsurfers networks are similar.

Fig. 9 shows how different types of attributes affect global polarization P_{GP} in the case of destabilization scenario for Windsurfers network. Obtained plots fit to other results shown in the main paper. Attributes are able to destroy the globally polarized state if there are not too many of them or if their strength is significant enough. Community structure of the network did not change that.

Fig. 10 allows comparison of two scenarios for networks with communities in the case of increasing strength of attribute layer. When networks are initially polarized according to distinct communities, then the attributes introduce local tensions which lead to an increase of the local polarization metric.



(a) Zachary karate club network structure



(b) Windsurfers network structure

Figure 8: Preventing polarization from forming (scenario B) for networks with community structure. Panels show local polarization metric P_{LP} as a function of the number of attributes G for different attribute types and strengths γ. For preventing, there are no significant differences between structures with and without communities.

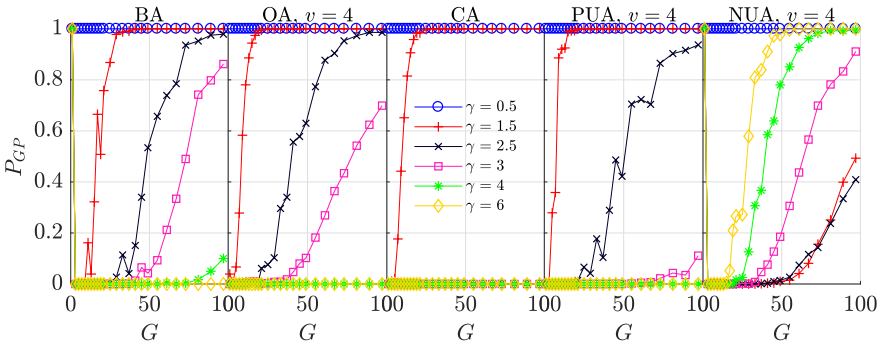
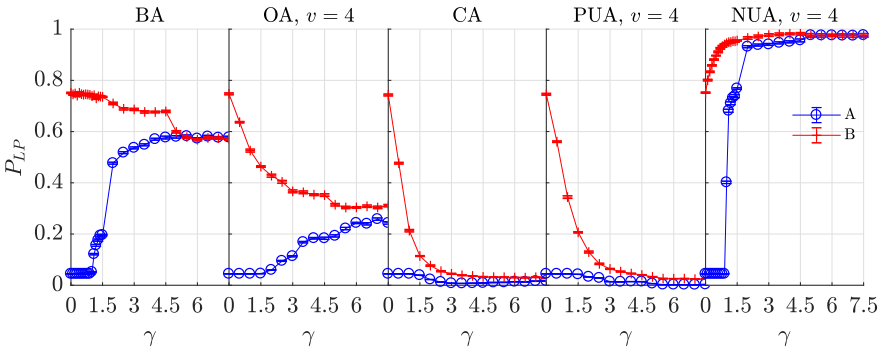
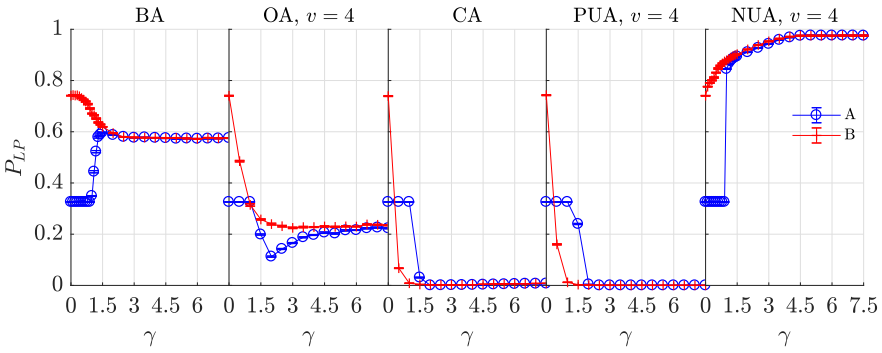


Figure 9: Impact of growing number of attributes G on global polarization metric P_{GP} in the case of Windsurfers network structure. Plots show results for the destabilization scenario.



(a) Zachary karate club network structure



(b) Windsurfers network structure

Figure 10: Influence of attributes in scenarios of (A) destabilization and (B) preventing from forming a polarized state for network structures with communities. Plots show local polarization metric P_{LP} as a function of attribute layer strength γ for different attribute types. In the case of destabilization, attributes introduce local tensions in existing communities.