Prejudice, group cohesion and the dynamics of disagreement

By Daniel C. Opolot*

Draft: April 29, 2019

This paper studies how individual prejudice—a set of preconceived and inflexible opinions—and group cohesion generate everlasting public disagreement in models of learning by averaging. We consider an endogenous model of opinion formation where agents compromise between respecting their own personal prejudice and conforming to the opinions held by others with whom they share close ties. We quantify the extent of equilibrium disagreement and show that its magnitude increases with the intensities of prejudice and group cohesion. Similarly, the speed of learning is a logarithmic function of the intensities of prejudice and group cohesion.

JEL: D83, D85, Z13, J15

Keywords: Prejudice, social learning, networks, group cohesion, disagreement

An active line of research in economics and sociology examines how people, connected through a social network, form opinions through a form of repeated averaging.[1] This line of research builds on the canonical model of DeGroot (1974)—the DeGroot model hereafter—and focuses on establishing a set of general conditions under which a group of agents converge to a state of *consensus*. However, ranging from financial markets, political institutions and various organizational settings, a consensus is rarely reached. The predominance of disagreement is even more pronounced across groups with differing attributes, interests and ideologies (Mutz, 2002; Huckfeldt, Johnson and Sprague, 2004; Grönlund, Herne and Setälä, 2015).[2] Despite this empirical evidence, there is limited theoretical literature on how the extent of cohesion within subgroups impacts public disagreement.

This paper aims to: (a) define a quantitative measure for the extent of disagreement in

* University of Cape Town, Private Bag X3, Rondebosch, 7701 Cape Town, daniel.opolot@uct.ac.za. This research was conducted during the time I spent at the University of Cape Town as a URC Postdoctoral Research Fellow. This paper has benefited from comments by Suraj Shekhar, Co-Pierre Georg and Michael Rose. I also acknowledge comments from the seminar participants at the School of Economics and AIFMRM – University of Cape Town. For all the help and advice that I have received from others, the usual disclaimer applies.

[1] For example DeMarzo, Vayanos and Zwiebel (2003), Golub and Jackson (2010), Acemoglu, Ozdaglar and ParandehGheibi (2010), and Jackson (2010) for a survey of the literature.

[2] For example, in 2009, over 97% of climate scientists, 82% of the overall scientific community and 58% of the American public believed that human activity contributed to global temperature changes (Doran and Zimmerman, 2009); and in March 2006, 41% of Republicans and 7% Democrats believed that Iraq had weapons of mass destruction just before the 2003 (World Public Opinion, 2006). This kind of heterogeneity in opinions between subgroups is predominant across a range of factual issues.

a population and the *intensity of group cohesion*; (*b*) examine how group cohesion impacts the magnitude of disagreement and the rate at which it decays/grows over time in models of learning by averaging. To do so, we revisit a sociological model of learning by Friedkin and Johnsen (1990)—Friedkin-Johnsen model hereafter—that is known to generate disagreement in equilibrium. We first demonstrate that the Friedkin-Johnsen model coincides with an economic model of best-response dynamics in which agents compromise between respecting their own personal *prejudice*—generally understood as a set of preconceived and inflexible opinions—and forming beliefs in line with those held by others with whom they share close ties. This interpretation is also consistent with research in psychology showing that some attitudes are resistant to change. This component of attitudes may in part be heritable and inculcated through culturalization (Tesser, 1993); stored in memory and activated automatically with little conscious control (Fazio, 1986; Houston and Fazio, 1989).[3]

To capture the notion of cohesive subgroups, we consider a learning process in which agents interact through a *social network*. The network captures relationships such as family and friendship ties, media-audience relations and ties based on political and religious orientation, social class and racial relations. Real-world networks are known to exhibit segregated patterns (McPherson, Smith-Lovin and Cook, 2001; McPherson, Smith-Lovin and Brashears, 2006; DiPrete et al., 2011; Del Vicario et al., 2016). We define cohesive subgroups based on such segregation patterns. Specifically, a subgroup of agents is cohesive if every member of the group has more than half of her interactions with other members of the group. We define the *intensity of cohesion* of a given subgroup as the relative total weight of interactions *among* group members, to the total weight of *all interactions* of group members. The overall intensity of group cohesion for a given network is then the sum of intensities of cohesion of all identifiable disjoint cohesive subgroups. Thus, a network with a high overall intensity of group cohesion is highly segregated.

To quantify the extent of disagreement, we define the magnitude of disagreement as the distance between the vector of equilibrium opinions and the vector of opinions describing the expected consensus. Recognizing that the primary cause of disagreement in the model is prejudice, the expected consensus vector is the vector of equilibrium opinions in an identical model of learning but without prejudice. This measure of the magnitude of disagreement is suitable for policy makers or planners who aim to implement policies to reduce or eradicate disagreement on specific issues (e.g. on public programs such as vaccination, and on environmental, social, political and economic policies). Such policies would directly or indirectly reduce the extent of prejudice, which ensures that the vector of equilibrium opinions is very close to the expected

[3] Further empirical and experimental evidence in support of this framework includes Kahan et al. (2012) and Kahan et al. (2017); and Wilson, Lindsey and Schooler (2000) provides a survey of the related literature in psychology. Kahan et al. (2012) and Kahan et al. (2017) specifically show that disagreement arises from the conflict of interest between conforming ones beliefs to their political outlook, and to the beliefs held by others with whom they share close ties.

equilibrium consensus vector. In this context, the magnitude of disagreement measures the costs associated with implementing policies that steer the population close to a consensus.

We derive bounds for the magnitude of disagreement in terms of the overall intensity of group cohesion and the intensity of prejudice. The bounds are increasing functions of the intensities of prejudice and group cohesion. Group cohesion, however, only acts to reinforce the effects of prejudice and is not the primary source of disagreement. That is, even in the absence of cohesive subgroups, disagreement can persist in equilibrium provided the intensity of prejudice is non-zero. From a policy perspective however, any policies that increase intergroup interactions can still help reduce the extent of disagreement across groups. Moreover, there is empirical evidence suggesting that such policies also have indirect positive effects on the extent of prejudice; that is, they reduce the extent of prejudice (Masson and Verkuyten, 1993; Pettigrew and Tropp, 2006). Thus, policies that directly increase intergroup contact have a larger impact in reducing disagreement compared to educational policies that do so by directly targeting individual prejudice (Hogan and Mallott, 2005; Kulik and Roberson, 2008).

Our findings offer an alternative explanation to the recent debate on political polarization in the American public. The Pew Research Centre for example documents ideological polarization along party lines (i.e. Republicans and Democrats), which is also reflected by the polarization in the U.S. Congress (Neal, 2018).[4] Our findings suggest that such polarization, which is equivalent to an increase in the magnitude of disagreement over time, may be a result of opposing subgroups becoming more cohesive, leading to lesser direct interaction/dialogue between Democrats and Republicans. Indeed, there is empirical evidence of increased levels of intensity of within subgroup cohesion for both Republicans and Democrats. The Pew Research Centre survey finds that 63% of consistent conservatives and 49% of consistent liberals say most of their close friends share their political views; and that 50% and 35% of people on the right and left respectively say it is important to them to live in a place where most people share their political views. These segregation patterns based on political views are even stronger in online social networks (Garrett, 2009; Del Vicario et al., 2016).

We also derive bounds for the speed of learning. We define the *convergence time* as the time it takes the learning process to converge to equilibrium. We show that the bounds for the convergence time decrease logarithmically with the intensity of prejudice. That is, when agents place more weight to their preconceived beliefs, there is lesser opinion exchange and the learning process settles to equilibrium fast. The bounds for convergence time however increase logarithmically with one minus the intensity of group cohesion. Here, higher intensity of group

[4] The Pew Research Centre surveys find that "the overall share of Americans who express consistently conservative or consistently liberal opinions has doubled over the past two decades from 10% to 21%. And ideological thinking is now much more closely aligned with partisanship than in the past. As a result, ideological overlap between the two parties has diminished: Today, 92% of Republicans are to the right of the median Democrat, and 94% of Democrats are to the left of the median Republican."

cohesion acts as a bottleneck for the exchange of opinions across subgroups, leading to slow convergence.

There are two implications of these findings. The first concerns the optimal persuasion-airtime allocation: for example, airtime allocation in political campaigns, court trials, and public programs campaigns. Persuasion-airtime is costly. If the objective of a political or public program campaign is to bring about a consensus, it is intuitive to think that the more airtime allocated, the better the outcome in terms of the proportion of the population that gets persuaded. Our findings suggest that in highly prejudiced groups, people make up their minds quickly (i.e. equilibrium is reached fast) and no amount of extra persuasion can help change their decisions; unless of course the extra persuasion is meant to change their prejudices. Thus, extended persuasion-airtime in such situations becomes a waste of resources. Second, our findings highlight the difference in convergence rates across different models of learning by averaging. These differences can thus be used in lab and field experiments to distinguish between various models.

Our paper contributes to a small but growing literature on disagreement (Friedkin and Johnsen, 1990; Krause, 2000; Hegselmann, Krause et al., 2002; Acemoğlu et al., 2013; Bindel, Kleinberg and Oren, 2015; Melguizo, 2016). Krause (2000) and Hegselmann, Krause et al. (2002) study variations of the Friedkin-Johnsen model and focus on establishing convergence results. They find that disagreement persists in equilibrium just as in the Friedkin-Johnsen model. Bindel, Kleinberg and Oren (2015) study the price of disagreement (i.e. the ratio of the total payoff in equilibrium to the total payoff at the social optimum) in the Friedkin-Johnsen model and examine the effects of different network structures. Melguizo (2016) studies a network formation model where agents rewire their links based on how close their opinions and types are. This model generates an equilibrium network structure that is segregated based on differences in opinions. Acemoğlu et al. (2013) shows that stubborn agents can be a source of disagreement in a random matching model. Except for Melguizo (2016), all the aforementioned papers thus study variations of the Friedkin-Johnsen model and focus on examining the structure of equilibrium opinions. The present paper instead focuses on characterising the magnitude and dynamics of disagreement and how they depend on group cohesion.

A more closely related paper is Golub and Jackson (2012) who examine the effect of *homophily*—the tendency of individuals to associate disproportionately with others who are similar to themselves—on the speed of learning in the DeGroot model. Our measure of the intensity of group cohesion however differs from the measure of *spectral homophily*—the second largest eigenvalue of the interaction matrix—adopted in Golub and Jackson (2012). As we will show, the two measures, the second largest eigenvalue of the network and intensity of group cohesion, are directly related only under some restrictions on the network structure. More importantly, however, our measure of intensity of group cohesion is very intuitive and easily computable from data, making it easily applicable to empirical analysis. Another fundamental difference with our paper is that we focus on finite deterministic networks and Golub and Jackson (2012) instead focus on random infinite networks. Since eigenvalue spectra of matrices, and hence spectral homophily, are generally very sensitive to matrix operations, the results for infinite random networks are not directly generalizable to finite deterministic networks.

The remainder of the paper is organized as follows. Section I outlines the model of opinion formation with prejudices. Section II quantifies disagreement and derives bounds for the magnitude of disagreement. Section III characterizes the speed of learning and Section IV offers concluding remarks. Technical proofs are relegated to the Appendix.

I. A model of opinion formation with prejudices

A. *Agents and interactions*

We consider a society of finite size denoted by a set $N = \{1, 2, \cdots, n\}$, where individuals interact through a social network. For each agent, a *neighbourhood* is the set of other agents (e.g. family, friends, colleagues, e.t.c) they interact with and regularly exchange ideas. The interactions are summarized by an $n \times n$ non-negative interaction matrix W, whereby $w_{ij} > 0$ indicates the weight that i attaches to j' opinions. Interactions are directed, so that $w_{ij} > 0$ need not imply $w_{ji} > 0$, or equality between w_{ij} and w_{ji}. We assume that W is row-stochastic, which means that an agent's interactions with her neighbours are normalized to sum to one. Relaxing this assumption does not impact the results qualitatively but affects whether the learning process converges. We discuss, in Section II, additional restrictions on the network structure that are sufficient for the validity of our results. We write N_i for the set of neighbours of i and d_i for its cardinality.

B. *Prejudice and the cost of miscoordination*

We model opinion formation as a learning process where agents repeatedly minimize the cost of miscoordination. When interacting with others, an agent's behaviour is driven by two competing motives. While an agent wants to agree with her personal *long-held opinions*, her utility depends on the degree to which her opinion coordinates with those held by others with whom she shares close ties. We interpret long-held opinions as *prejudice*, defined as a set of preconceived and *inflexible* opinions or beliefs about individual attributes, group behaviour or government policies and public programs. We interpret inflexibility of prejudices to imply that although they may change, they change more slowly compared to the rate at which overall opinions change due to learning from others. There is evidence in psychology literature showing that some attitudes/beliefs are resistant to change and may in part be heritable (Tesser, 1993). That is, inculcated through culturalization. This component of attitudes is stored in memory and activated automatically with little conscious control (Fazio, 1986; Houston and Fazio, 1989). As individuals adjust their attitudes through learning, the inflexible beliefs are not completely erased (Wilson, Lindsey and Schooler, 2000).

The desire to coordinate ones opinion and behaviour with that of ones neighbours is driven by individual desire for social conformism. Since the work of Asch and Guetzkow (1951) on social pressure, individual desire for social conformism is by now a well-studied phenomenon. More recently, Salganik, Dodds and Watts (2006) find evidence of conformity in individual taste in music, reflecting individual opinions about what constitutes good music. In economics, social conformism has been observed in work habits and effort exerted (e.g. Falk and Ichino (2006), Chen et al. (2010), Zafar (2011) and Abeler et al. (2011)), and participation in public good provision (Carpenter, 2004).

Our model is particularly in line with the empirical evidence in Kahan et al. (2012) and Kahan et al. (2017), showing that the observed patterns of disagreement regarding the role of human activity on climate change, results from the conflict of interest between conforming ones beliefs to their political outlook (which is representative of ones prejudices), and to the beliefs held by others with whom they share close ties. We then suppose that when adjusting their opinions, agents compromise between respecting their own personal prejudice, and conforming their opinions to those held by others with whom they share close ties.

To formalize these ideas, let p_i and $\bar{p}_i$ be the opinion and the extent of prejudice of agent i respectively. We assume that $p_i \in \mathbb{R}_+$, that is, *unidimensional* and represented by a positive real number. This assumption is consistent with empirical evidence suggesting that while individuals have opinions on many issues, spanning domains such as politics, lifestyle and the economy, an individuals opinions on all dimensions can be described using a *unidimensional spectrum*. Poole and Daniels (1985) and Ansolabehere, Rodden and Snyder (2008) find that the voting behaviour of both legislators and individual voters can be explained by a single liberal-conservative dimension. Following in a similar line of argument, we assume that personal prejudices can also be described using a unidimensional spectrum, and more specifically, it takes values in the range $[0, 1]$. For any issue, say individual attribute, group behaviour, or government policy and public program, a value of $\bar{p}_i = 1$ means that agent i is fully prejudiced towards an issue, and a value of $\bar{p}_i = 0$ means i is not prejudiced. A value between zero and one then means that an agent is partially prejudiced. In situations where agents hold prejudices on many related issues, $\bar{p}_i$ is then a unidimensional parameter capturing the overall or average prejudice of an agent. The assumption regarding the range of values of $\bar{p}_i$ does not affect our results in a qualitative sense. Thus, in situations where it is suitable, it is feasible to assume

$\bar{p}_i \in [-1, 1]$ with -1 representing a fully negatively prejudiced agent, 1 for a fully positively prejudiced agent, zero for no prejudice, and any other value in-between means an agent is only partial prejudiced.

Let $\mathbf{p}$ and $\bar{\mathbf{p}}$ denote the vectors of opinions and prejudices respectively, and write $\mathbf{p}_{-i}$ as the vector of opinions with i's opinion excluded. Each agent minimizes the cost $C_i(p_i, \mathbf{p}_{-i}, \bar{p}_i)$ of mis-coordinating her opinion against her personal prejudice and the opinions of her neighbours. That is

$$C_i(p_i, \mathbf{p}_{-i}, \bar{p}_i) = -\lambda_i \sum_{j=1}^{n} w_{ij}(p_j - p_i)^2 - (1 - \lambda_i)(p_i - \bar{p}_i)^2 \tag{1}$$

where λ_i is the intensity with which i conforms to the opinions of her neighbours; as such, we interpret $1 - \lambda_i$ as the intensity with which i conforms to her prejudice, or simply the *intensity of prejudice*. The first order condition to (1) is

$$p_i = -\lambda_i \sum_{j=1}^{n} w_{ij} p_j + (1 - \lambda_i)\bar{p}_i \tag{2}$$

Although we focus on a model of opinion formation, the cost function in (1) admits a game theoretic interpretation where p_i is i's action or effort and $\bar{p}_i$ is i's preference. Under this interpretation, an agent minimizes the cost of mis-coordinating her actions against her personal preference and that against the actions of her neighbours. Our model and results can thus be interpreted as a model of evolution of behaviour in the presence of relatively static preferences.

C. *Evolution of opinions*

By definition, prejudices are inflexible or change very slowly compared to the rate at which overall opinions are updated. That is, the process of learning by averaging reaches equilibrium before prejudices are updated. Here, we also assume that the network structure does not change at the same rate as that at which opinions are updated.

We can thus think of the overall dynamics of learning to be unfolding on two discrete time scales demonstrated in Figure 1. The first time frame $t \in \{0, 1, 2, \cdots, T\}$, represents the time intervals at which agents update their opinions due to learning from neighbours. Within this time frame, the network structure, vector of prejudices and intensities of prejudice are all either constant or change at negligible rates. The final period T in this time frame represents the time at which the learning process reaches equilibrium; hence, T varies depending on the network structure and intensities of prejudice—later in the paper, we define T as a convergence time and examine how it depends on the network structure (intensities of group cohesion) and prejudices

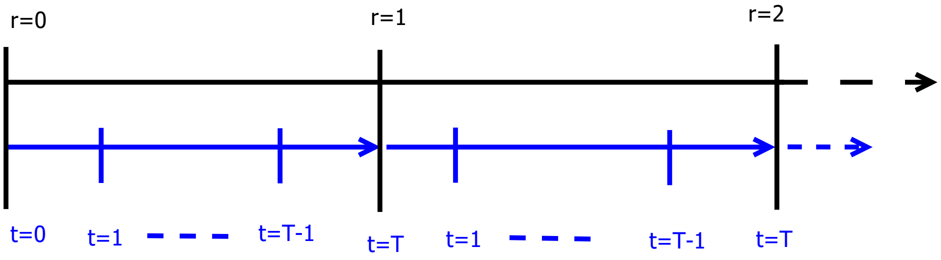


Figure 1. : Time scales for evolution of opinions and network structure.

(intensities of prejudice). The second time scale is $r \in \{0, 1, 2, \cdots, R\}$, which is the time frame at which the network structure and/or prejudices can significantly change. Once either network structure or prejudices change, a new equilibrium vector of opinions will be reached. Since the length of time, T, at which the evolutionary process converges within the time frame $\{0, 1, 2, \cdots, T\}$ depends on the structures of the network and prejudices, it must be a function of $r \in \{0, 1, 2, \cdots, R\}$; that is, $T := T(r)$.

Within $t \in \{0, 1, 2, \cdots, T(r)\}$, each agent i updates her opinion myopically in accordance with the optimal value of p_i in (2). For a given $r \in \{0, 1, 2, \cdots, R\}$, let $p_i(t; r)$ and $\mathbf{p}(t; r)$ denote p_i and $\mathbf{p}$ respective at time t. Then from (2), agent i's opinion at t is given by

$$(3) \qquad p_i(t;r) = -\lambda_i(r) \sum_{j=1}^{n} w_{ij}(r) p_j(t-1;r) + (1 - \lambda_i(r))\bar{p}_i(r) \quad \text{for } i = 1, \cdots, n.$$

Let $\Lambda(r)$ be an $n \times n$ diagonal matrix with entries $\lambda_{ii}(r) = \lambda_i(r)$ and zero otherwise. Let also I denote an $n \times n$ identity matrix. Then the system (3) can be expressed in matrix form as

$$(4) \qquad \mathbf{p}(t;r) = \Lambda(r) W(r) \mathbf{p}(t-1;r) + (I - \Lambda(r))\bar{\mathbf{p}}(r)$$

The system of equations in (4) describe a simultaneous evolution of individual opinions where agents take weighted averages of neighbours' opinions and add them to their static prejudices. The model of evolution of opinions through the system described by (4) coincides with the Friedkin-Johnsen model of opinion formation. Friedkin and Johnsen (1990) start their analysis from (3), and interpret $\bar{p}_i$ as an initial opinion. They then interpret the second term on the right hand side of (3) as an exogenous influence on an agents opinion by the conditions that have formed her initial opinions. Friedkin and Johnsen (1990) then use this framework to show that disagreement persists in the long-run. We have instead motivated (4) as resulting from an endogenous best-response dynamic process.

To examine the dynamics of disagreement and the convergence rates for the evolutionary process (4), we first fix $r \in \{0, 1, 2, \cdots, R\}$ and derive the expressions for the magnitude of disagreement and convergence rate over the time frame $\{0, 1, 2, \cdots, T(r)\}$. Thus, we drop argument r from the variables W, Λ and $\bar{\mathbf{p}}$ so that (4) becomes.

$$\mathbf{p}(t) = \Lambda W \mathbf{p}(t-1) + (I - \Lambda)\bar{\mathbf{p}} \tag{5}$$

We then perform comparative statics as a way of examining the effects of varying W, Λ and $\bar{\mathbf{p}}$ over the time frame $\{0, 1, 2, \cdots, R\}$.

D. *Convergence and equilibrium opinions*

The dynamic system in (5) is convergent and stable if for any vector of prejudice $\bar{\mathbf{p}}$ and initial opinions $\mathbf{p}(0)$, there exist a unique vector of opinions $\mathbf{p}^* = \lim_{t\to\infty} \mathbf{p}(t)$, whereby each p_i^* for all i is finite. The vector $\mathbf{p}^*$ represents equilibrium opinions. Note that $\mathbf{p}^*$ is defined over an infinite limit of t. But since the population size is finite, and as we show in Section III, if (5) is convergent, then it converges to equilibrium opinions within a finite time, $T(r)$. Thus, given $r \in \{0, 1, 2, \cdots, R\}$, we say that (5) is convergent and stable if

$$\mathbf{p}^* = \lim_{t\to\infty} \mathbf{p}(t) \equiv \lim_{t\to T(r)} \mathbf{p}(t)$$

From Parsegov et al. (2017, Theorem 1 & Corollary 1), the sufficient conditions for (5) to be convergent and stable are for either $0 \leq \lambda_i < 1$ for all i, or $\Lambda \neq 1$ (i.e. there exists at least one $i \in N$ for whom $\lambda_i < 1$) and W is *strongly connected*. An interaction matrix W, and hence the associated network structure, is strongly connected if for every pair of agents i and j, there exists a path of links connecting i to j and vice versa. Thus, by definition, if W is strongly connected, then it is also irreducible. For (5) to be convergent and stable, it is then sufficient either for all agents to be prejudiced or a few agents to be prejudiced and W to be strongly connected. When either of these conditions are satisfied, the equilibrium vector $\mathbf{p}^*$ is given by[5]

$$\mathbf{p}^* = (I - \Lambda W)^{-1} (I - \Lambda)\bar{\mathbf{p}} \tag{6}$$

Based on the above sufficient conditions for convergence, we make the following three assumptions for the remainder of the paper:

(i) All agents are prejudiced and $\lambda_i = \lambda$ for all $i \in N$.

[5] From (5) we see that the equilibrium vector $\mathbf{p}^*$ is given by $\mathbf{p}^* = \Lambda W \mathbf{p}^* + (I - \Lambda)\bar{\mathbf{p}}$, which yields $\mathbf{p}^* = (I - \Lambda W)^{-1} (I - \Lambda)\bar{\mathbf{p}}$.

(ii) The network structure is strongly connected so that W is irreducible; we also assume that W is aperiodic.

(iii) The vector of prejudice $\bar{\mathbf{p}} = \mathbf{p}(0)$.

Note that the first part of assumption (i) (i.e. requiring all agents to be prejudiced) is sufficient for (5) to be convergent. The second part of assumption (i) (i.e. requiring the intensity of prejudice to be identical for all agents) is for analytical simplicity. It ensures that we can derive intuitive analytical results on the effect of the networks structure and the intensity of prejudice on the extent of disagreement. Assuming heterogeneous prejudices complicates the analysis without much added value with regards to qualitative findings. Assumption (ii) is also for analytical simplicity and assuming otherwise does not change the results qualitatively. Strong connectivity of the network, and hence irreducibility, together with aperiodicity of W puts restrictions on the structure of the eigenvalue spectrum of W.[6] It ensures that the leading eigenvalue of W is one and all other eigenvalues are less than one. Both conditions simplify analysis in the next sections.

Assumption (iii) states that the vector of prejudices is equivalent to the vector of initial opinions. This assumption is reasonable because individual prejudices are formed by some exogenous conditions that influenced an individual or a group of individuals in the past. Since the vectors of prejudices and initial opinions store information about an individual's or a group's history, it is reasonable to assume equivalence between them.

II. Quantifying disagreement

To define an appropriate measure for the extent of disagreement, it helps to first examine the primary source of equilibrium disagreement in model (5). Disagreement is said to exist in equilibrium if there exists at least one pair of agents i and j for whom $p_i^* \neq p_j^*$. A consensus, per contra, occurs if $p_i^* = p_j^*$ for all pairs $i, j \in N$.[7] A close examination of the structure of the dynamic system (5) and equilibrium opinions in (6) reveals that the primary source of disagreement is prejudice. Specifically, in the absence of prejudice, (i.e. when $\lambda = 1$), (5) reduces to the DeGroot model of opinion formation, which is known to converge to a consensus (DeGroot, 1974).[8]

Example 1: As a demonstration, Figure 2 plots evolution of opinions of four agents whose

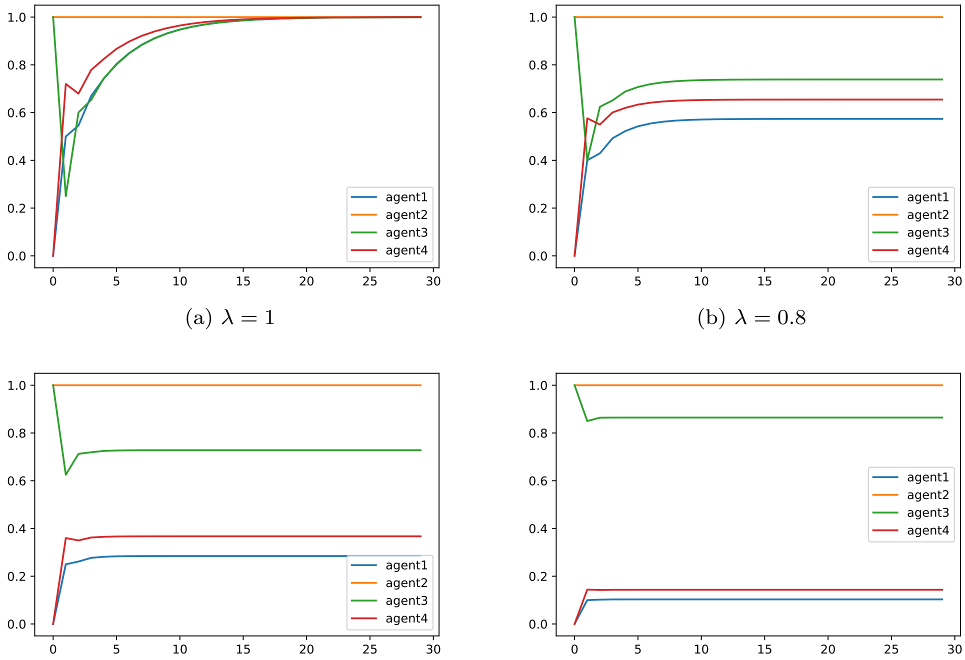
[6] Aperiodicity ensures that no cycles exist within a strongly connected subgroup of agents. For example, if a cycle exists among a group of agents $\{i, j, k\}$, then the interactions are of the form $i \to j \to k \to i$. In such situations, agents within a cycle alternate in adopting each others opinions. The process may thus fail to converge, especially in the case where $\lambda = 1$.

[7] Note that these definitions can be extended to subgroups of agents; in which case, a subgroup of agents $C \subset N$ reaches a consensus if $p_i^* = p_j^*$ for all pairs $i, j \in C$, and disagreement if $p_i^* \neq p_j^*$. Here, we stick to defining disagreement at the level of the entire population.

[8] Specifically, the DeGroot model converges to a consensus provided the network does not consist of cycles that lead agents to alternate in adopting each other's opinions. In the presence of prejudices, disagreement persists in equilibrium regardless of the network structure.

interactions are depicted in (7). Starting from Figure 2 (a), where $\lambda = 1$, to Figure 2 (d) where $\lambda = 0.2$, the equilibrium behaviour evolves from a consensus, in which all agents adopt agent 2's initial opinion (prejudice), to a more heterogeneous behaviour with different equilibrium opinions. The smaller λ, the larger the extent of disagreement among agents. The primary source of disagreement in model (5) is thus prejudice and the network structure only acts to reinforce the effects of prejudice.

$$W = \begin{bmatrix} 0.2 & 0.14 & 0.36 & 0.3 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.35 & 0.1 & 0.15 & 0.4 \\ 0.18 & 0.45 & 0.27 & 0.1 \end{bmatrix} \tag{7}$$



(a) $\lambda = 1$

(b) $\lambda = 0.8$

(c) $\lambda = 0.5$

(d) $\lambda = 0.2$

Figure 2. : Each subfigure, from (a) to (d), graphs the evolution of opinions over time, where in Figure (a), $\lambda = 1.0$; Figure (b), $\lambda = 0.8$; Figure (c), $\lambda = 0.5$; and Figure (d), $\lambda = 0.2$. The initial vector of opinions in all cases is $\mathbf{p}(0) = (0, 1, 1, 0)$.

We exploit this relationship between disagreement and prejudice in defining the magnitude of disagreement. Specifically, given the interaction matrix W, we define the magnitude of disagreement at a given level of intensity of prejudice λ as the distance between the vector of equilibrium opinions at λ, and the *expected consensus vector*. For a given interaction matrix W, the expected consensus vector corresponds to the vector of equilibrium opinions when $\lambda = 1$.

Let $\mathbf{c}$ be the expected consensus vector. Then the magnitude of disagreement $D(W;\lambda)$ is defined as follows.

DEFINITION 1: *Given an equilibrium vector of opinions* $\mathbf{p}^*$ *in the model of opinion formation with prejudice, the magnitude of equilibrium disagreement* $D(W;\lambda)$ *under interaction matrix* W *and intensity of prejudice* λ *is defined for the time frame* $\{0,1,2,\cdots,T(r)\}$ *as*

$$D(W;\lambda) = \sup_{\bar{\mathbf{p}}\in[0,1]^n} \|\mathbf{p}^* - \mathbf{c}\|_{\mathbf{v}} = \sup_{\bar{\mathbf{p}}\in[0,1]^n} \left(\sum_{i=1}^{n} v_i \left(p_i^* - c_i\right)^2 \right)^{\frac{1}{2}} \tag{8}$$

where $\mathbf{v}$ *is a normalized vector, that is* $\sum_{i\in N} v_i = 1$*, representing the relative importance of each agent's contribution to group disagreement.*

Note that $D(W;\lambda)$ is already a function of the network structure and intensity of prejudice, which are assumed to be constant within a given $r \in \{0,1,2,\cdots,R\}$. Thus, it is not necessary to include r in the argument of $D(W;\lambda)$. To examine the dynamics of $D(W;\lambda)$ over the time frame $\{0,1,2,\cdots,R\}$, we use comparative statics to examine how $D(W;\lambda)$ varies with W and λ.

Depending on the context, there are alternative ways of quantifying disagreement. For example, if the focus is on the disutility that individuals incur from disagreeing with their neighbours, then the suitable measure should take into account the deviations of individual equilibrium opinions from that of the neighbours.[9] We are instead interested in quantifying disagreement from the point of view of a planner or policy maker aiming to reduce the extent of disagreement among group members and society in general. Under such circumstances, the target vector of equilibrium opinions for the policy maker is the consensus vector. That is, the result of any policies that directly or indirectly reduce the intensity of prejudice is to shift the vector of equilibrium opinions close to $\mathbf{c}$. The magnitude of disagreement, $D(W;\lambda)$, thus measures the costs associated with implementation of policies that steer the population to a consensus.

We highlight two aspects of the expressions in (8). The first concerns the supremum over all possible vectors of prejudices. The vector of equilibrium opinions $\mathbf{p}^*$ and the consensus vector $\mathbf{c}$ are both functions of the vector of prejudices. This implies that the vector of prejudices can be suitably chosen so that $\mathbf{p}^*$ is close or identical to $\mathbf{c}$. We aim to derive bounds for the worst possible scenario; that is, we consider a choice of the vector of prejudices that produces the largest values of the magnitude of disagreement given the network structure and intensity of prejudice. The direct way of achieving this is to take the supremum over all possible vectors of prejudices.

[9] For example, disagreement could be defined as $D(W;\lambda) = \sum_{i\in N}\sum_{j\in N_i}(p_j^* - p_i^*)^2$, where each $(p_j^* - p_i^*)^2$ is the disutility that i incurs from having an equilibrium opinion that is not aligned with j's equilibrium opinion.

The second aspect concerns the weighting vector $\mathbf{v}$. This vector ensures that each agent's contribution to overall disagreement is weighted by v_i. A natural candidate for the choice of $\mathbf{v}$ is the vector of influences that agents command in the network, and the influence they exert on each others equilibrium opinions. For example, when $\lambda > 0$, we see from (6) that the vector of influence, generally referred to as *Bonacich centrality* (Bonacich, 1987),[10] is given by

$$\mathbf{b}(\lambda) = (1-\lambda)\left[(I-\lambda W)^{-1}\right]\mathbf{e} = (1-\lambda)\left[\sum_{\tau=0}^{+\infty}(\lambda W)^{\tau}\right]\mathbf{e} \tag{9}$$

where $\mathbf{e}$ is a column vector of ones. The Bonacich centrality accords each player a level of influence that is proportional to the number of connections she has, and the number of connections her neighbours have, and so on. When $\lambda = 1$, the vector of Bonacich centralities reduces to *eigenvector centralities* π, so called because it is equivalent to the left eigenvector that corresponds to the leading eigenvalue of W.[11] Thus, when $\lambda = 1$, then $\mathbf{p}^* = \mathbf{c}$ so that equilibrium opinions of all agents are identical and given by $c_i = \sum_{j=1}^{n} \pi_j p_j(0)$. The vector of eigenvector centralities thus describes the influence each agent commands in the network (since it excludes the contribution of individual prejudice). For the remainder of the paper, we assume, without loss of generality, that $\mathbf{v} = \pi$.

We aim to establish bounds for the magnitude of disagreement in terms of intensity of prejudice and parameters related to the network structure. We identify *group cohesion* as the main property of the network that amplifies the magnitude of disagreement. A subgroup of agents is *cohesive* if every member of the subgroup has more than half of her interactions with other members of the subgroup. That is, let $L_k \subset N$ be a subset of N; then L_k is said to be cohesive if for every $i \in L_k$,

$$\sum_{j \in L_k} w_{ij} > \frac{1}{2}\sum_{l \in N} w_{il} \tag{10}$$

Every network structure consists of at least one cohesive subgroup $L_k \subseteq N$ for $k = 1, 2, \cdots, K$, where $L_k \cup L_l = \emptyset$ for all pairs of subgroups. A complete network, where each agent interacts with every other, is one of the special cases with only one cohesive subgroup. Most other network structure however contain at least two disjoint cohesive subgroups. For each W, let $\mathcal{L}(W)$, or simply $\mathcal{L}$ where no confusion arises, be the set of all its cohesive subgroups; and let

[10] The Bonacich centrality is the main network measure that determines the equilibrium behaviour of most dynamic processes on networks. For example, Ballester, Calvó-Armengol and Zenou (2006) show that equilibrium behaviour in network games depends on Bonacich centralities.

[11] To see why, first note that when $\lambda = 1$ (4) reduces to $\mathbf{p}(t) = W^t\mathbf{p}(0)$ so that the vector of equilibrium opinions becomes $\mathbf{p}^* = \lim_{t\to\infty} W^t\mathbf{p}(0)$. Since this model converges to a consensus, the limit $\lim_{t\to\infty} W^t = \Pi = \mathbf{e}\pi^T$, where π is the vector of equilibrium influence. Thus, $\mathbf{p}^* = \mathbf{e}\pi^T\mathbf{p}(0)$, which implies that $p_i^* = \sum_{j=1}^{n} \pi_j p_j(0)$. Since Π is derived by infinitely iterating W^t, then $\Pi^t = \Pi$ and $\Pi = \Pi W$. This implies that for each influence vector π, $\pi^T = \pi^T W$, and hence π is a left eigenvector of W corresponding to the leading eigenvalue $\lambda_1 = 1$.

n_k be the cardinality of each $L_k \in \mathcal{L}$. We define the *intensity of cohesion* $\iota(L_k)$ of subgroup L_k as the relative total weight of interactions among members of L_k to the total weight of all interactions of members of L_k; that is,

$$\iota(L_k) = \frac{\sum_{i\in L_k}\sum_{j\in L_k} w_{ij}}{\sum_{i\in L_k}\sum_{j\in N} w_{ij}} = \frac{1}{n_k}\sum_{i\in L_k}\sum_{j\in L_k} w_{ij} \tag{11}$$

where the second equality on the right hand side of (11) follows from the fact that W is a row stochastic matrix so that $\sum_{j\in N} w_{ij} = 1$. For each L_k, $\iota(L_k)$ is in the interval $(\frac{1}{2}, 1)$. The larger $\iota(L_k)$, the larger the total weight of interactions among group members compared to interactions with non-group members, and hence, the higher the level of cohesion among group members. Given all cohesive subgroups of W, we define the overall intensity of group cohesion $\iota(W)$ of W as the sum of all intensities of cohesion of all subgroups:

$$\iota(W) = \sum_{L_k\in\mathcal{L}(W)} \iota(L_k).$$

The following additional definitions and notations are used in the analysis that follows. We define $\bar{W}$ as an $n \times K$ matrix with each element $\bar{w}_{il}$ of $\bar{W}$ as the total weight that agent i attaches to subgroup L_l; that is, $\bar{w}_{il} = \sum_{j\in L_l} w_{ij}$. We also define a matrix $\tilde{W}$ with each element $\tilde{w}_{kl}$ of $\tilde{W}$ as the total weight that agents in subgroup L_k attach to subgroup L_l; that is, $\tilde{w}_{kl} = \frac{1}{n_k}\sum_{i\in L_k}\sum_{j\in L_l} w_{ij}$. When compared to (11), we see that the diagonal elements of $\tilde{W}$ are the respective intensities of subgroup cohesion, and the trace of $\tilde{W}$ (i.e. the sum of diagonal elements of $\tilde{W}$) is equivalent to $\iota(W)$.

PROPOSITION 1: *Let W be strongly connected and aperiodic. In addition, let W contain at least two cohesive subgroups where $\bar{w}_{il} = \bar{w}_{jl}$ for all pairs $i, j \in L_k$ and for each subgroup L_l; and let $\tilde{W}$ be symmetric, that is, $\tilde{w}_{kl} = \tilde{w}_{lk}$ for each pair $L_k, L_l \in \mathcal{L}$. Then for an evolutionary process described by* (5)*, the magnitude of disagreement is bounded by*

$$\left(\frac{(1-\lambda)\,K}{K-\lambda\,(2\iota(W)-K)}\right) \le D(W;\lambda) \le \pi_{\min}^{-\frac{1}{2}}\left(\frac{1-\lambda}{1-\lambda\,(2\iota(W)-1)}\right) \tag{12}$$

where $\pi_{\min} = \min_{i\in N}\pi_i$.

PROOF:

See Appendix A.A1

The proof of Proposition 1 proceeds in three steps, which in turn determine the stated conditions for the validity of (12). The first step involves establishing the bounds for the magnitude of disagreement in terms of the second largest eigenvalue of the interaction matrix. The corresponding conditions for the validity of results at this stage of the proof are, a strongly connected and aperiodic interaction matrix. As discussed in Section I.D, these two assumptions ensure that the eigenvalues of the interaction matrix are distinct, and hence, the second largest and smallest eigenvalues of W are non-identical and less than one. This also ensures that the dynamics process in (4) converges.

The second stage of the proof establishes the relationship between the second largest eigenvalue of the interaction matrix W and that of the inter-group interaction matrix $\tilde{W}$. Recall that the trace of $\tilde{W}$ defines the overall intensity of cohesion $\iota(W)$ of subgroups. We can thus establish the relationship between the magnitude of disagreement and intensity of group cohesion by exploiting the relationship between the eigenvalue spectrum of $\tilde{W}$ and its trace. We show that the sufficient condition for the equivalence between the second largest eigenvalues of W and $\tilde{W}$ is for $\bar{w}_{il} = \bar{w}_{jl}$ for all pairs $i, j \in L_k$ and for each subgroup $L_l \in \mathcal{L}$. For any pair of distinct subgroups L_l and L_m, however, $\bar{w}_{iL_l}$ need not equal $\bar{w}_{iL_m}$. That is, all agents belonging to the same cohesive group attach the same weight to a given cohesive subgroup, but these weights need not be identical across subgroups. Put differently, this condition requires agents that share attributes to attach a similar weight to the opinion/behaviour of agents with specific attributes. This is a reasonable assumption in situations where cohesive subgroups consist of agents with similar attributes such as political orientation, religious views, social or income class and ethnic groups.

The eigenvalue spectra of matrices are generally sensitive to matrix operations. The condition discussed in the preceding paragraph ensures that the interaction matrix W can be collapsed to $\tilde{W}$ without changing the eigenvalue spectrum. We perform a sensitivity analysis in Section A.A2 and show that provided the deviation from the stated condition is not large, the deviation of the second largest eigenvalue of $\tilde{W}$ from that of W is small. If the deviations are large, then the second largest eigenvalues of W and $\tilde{W}$ need not be identical or close to identical. Consequently, Proposition 1 need not hold.

The third stage of the proof leads to the third condition in Proposition 1: $\tilde{W}$ must be symmetric; that is, $\tilde{w}_{kl} = \tilde{w}_{lk}$ for each pair $L_k, L_l \in \mathcal{L}$. This condition enables us to establish the relationship between the second largest eigenvalue of $\tilde{W}$ and the overall intensity of group cohesion $\iota(W)$. We specifically use the results on bounds for eigenvalues using traces (Wolkowicz and Styan, 1980). There are several successive papers since Wolkowicz and Styan (1980) (e.g. Merikoski and Virtanen (1997)) that have established tighter bounds for eigenvalues in term of a matrix trace but the respective expressions are cumbersome and do not improve the results of Proposition 1 in a qualitative sense. A symmetric $\tilde{W}$ matrix implies that the links between cohesive subgroups are undirected so that the weights that a pair of cohesive subgroups attach to each other's opinions are identical. It is possible to relax this assumption, but it comes at the cost of tighter bounds for the second largest eigenvalue of $\tilde{W}$. Moreover, just as in the second condition discussed above, it is a reasonable assumption in situations where cohesive subgroups represent groups of agents with similar attributes.

From (12), we see that the lower and upper bounds for the magnitude of disagreement increase with the intensities of prejudice $(1-\lambda)$ and subgroups cohesion $\iota(W)$. Consistent with Example 1 above, the primary source of disagreement is prejudice, whereby, the magnitude of disagreement increases from zero, when the intensity of prejudice is zero, to the maximum possible value of one when the intensity of prejudice is also one. The interaction structure only acts to reinforce the effects of intensity of prejudice. That is, even in complete networks, where every agent attaches the same weight to every other agent and the overall intensity of group cohesion is zero, disagreement can persist in equilibrium provided the intensity of prejudice is non-zero. Figure 3 captures this scenario. It plots the evolution of opinions for a complete network and clearly depicts persistence of disagreement in equilibrium.

$$W = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

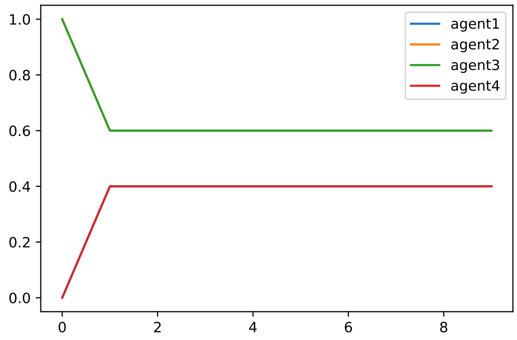
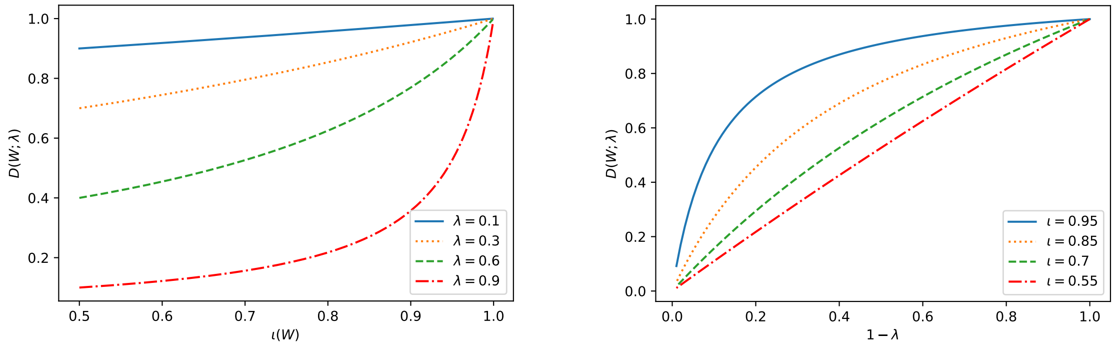


Figure 3. : An example of an interaction structure in which every agent attaches the same weight to every other. The figure on the right hand side plots the evolution of opinions for $\lambda = 0.8$ and the vector of prejudice/initial opinions $\mathbf{p}(0) = (0, 1, 1, 0)$. Disagreement persists in the long-run where agents 2 and 3 converge to the same opinion of 0.6 and agents 1 and 4 converge to the same opinion of 0.4.

The rates at which the bounds for the magnitude of disagreement increase with the intensities of prejudice and group cohesion are not constant. Figure 4 demonstrate this relationship. Clearly, the magnitude of disagreement is very sensitive to changes in intensity of group cohesion at low levels of intensity of prejudice (i.e. when λ is large). And, the magnitude of disagreement is very sensitive to changes in intensity of prejudice at high levels of intensity of group cohesion.

We highlight two implications of Proposition 1. First, Proposition 1 provides an alternative explanation to the recent debate on political polarization in the American public. Polarization is defined as the divergence of opinions over time. The Pew Research Centre for example documents ideological polarization along party lines (i.e. Republicans and Democrats). Specifically, they find that "the overall share of Americans who express consistently conservative or consistently liberal opinions has doubled over the past two decades from 10% to 21%. And ideological



(a) Magnitude of disagreement versus intensity of group cohesion

(b) Magnitude of disagreement versus intensity of prejudice

Figure 4. : The figure plots the magnitude of disagreement versus intensities of prejudice and group cohesion using the expression of the lower bound in (12).

thinking is now much more closely aligned with partisanship than in the past. As a result, ideological overlap between the two parties has diminished: Today, 92% of Republicans are to the right of the median Democrat, and 94% of Democrats are to the left of the median Republican." Neal (2018) also documents polarization patterns in the U.S Congress. Polarization in the context of our model is equivalent to the growth of the magnitude of disagreement over the time frame $\{0, 1, \cdots, R\}$.

According to Proposition 1, as also depicted in Figure 4, increasing the intensity of group cohesion over time, increases the magnitude of disagreement, and hence, leading to polarization of opinions. Several surveys and empirical analysis document evidence of high levels of group cohesion among Democrats and Republicans in recent years. For examples, The Pew Research Centre survey finds that 63% of consistent conservatives and 49% of consistent liberals say most of their close friends share their political views; and that 50% and 35% of people on the right and left respectively say it is important to them to live in a place where most people share their political views. Such segregation patterns are even stronger in online social networks and appear in form of echo chambers" (Garrett, 2009; Del Vicario et al., 2016). The observed political polarization could thus be a result of the opposing groups becoming more cohesive.

Second, the results of Proposition 1 offer insights for policy makers. There are two ways in which a policy maker can reduce the extent of disagreement in the society: by reducing the intensities of prejudice and group cohesion. According to several studies, individual prejudice can be reduced through diversity education programs. For example, Hogan and Mallott (2005) show that diversity courses in higher education were effective in improving students' intergroup tolerance (see Kulik and Roberson (2008) for a review of the related literature). Besides educational programs, research in social psychology and political science shows that policies that encourage contact across cohesive subgroups also tend to indirectly reduce the level of prejudice in the society (Masson and Verkuyten, 1993; Pettigrew and Tropp, 2006; Mutz, 2002; Grönlund, Herne and Setälä, 2015). Thus, policies and public programs aimed at fostering social, economic and cultural integration, have a double positive impact in reducing the magnitude of disagreement.

The results in Proposition 1 are closely related to Golub and Jackson (2012), who show that *homophily*—the tendency for agents to interact with those with whom they share attributes—influences the speed of convergence in the DeGroot model. Golub and Jackson (2012) define *spectral homophily* as the second largest eigenvalue of the interaction matrix describing inter-subgroup interactions. To derive their results, Golub and Jackson (2012) consider a family of networks formed through a random Bernoulli process so that every neighbour who is listened to is weighted equally. Using mean-field approximation techniques, Golub and Jackson (2012) show that as population size tends to infinity, the second largest eigenvalue of the entire network coincides with spectral homophily.

As stated above, the eigenvalue spectra of matrices, and hence spectral homophily, are sensitive to matrix operations. Thus, the results in Golub and Jackson (2012) need not directly generalize beyond infinitely large random networks. In contrast, we define the intensity of group cohesion, for finite networks, as the total weight that agents attach to fellow group members. This measure is intuitive and easily computable from empirical data compared to spectral homophily. We also focus on finite deterministic networks, making our results directly applicable to empirical analysis.

III. The speed of learning

For any model of learning, examining the convergence rate (speed of learning) is just as relevant as examining the properties of equilibrium behaviour. It is very important to understand whether the predicted equilibrium behaviour can be reached at the time scales of economic relevance. In this section, in addition to this necessity, we demonstrate that the speed of learning can also be used to distinguish between models of learning by averaging. Since disagreement in the society about factual issues occurs more often than not, some papers have studied variations of the DeGroot model so as to generate disagreement as an equilibrium behaviour (Acemoğlu et al., 2013; Melguizo, 2016). It then remains to be empirically demonstrated which among the existing models best fits reality.

From Example 1 above, given an interaction structure, the rate at which opinions converge (i.e. the decay rate of disagreement) increases with the intensity of prejudice. Specifically, from Figure 2, when $\lambda = 0.2$ opinions get close to equilibrium values in less than five steps of iteration. When $\lambda = 0.8$, it takes at least 10 steps of iteration, and when $\lambda = 1$, it takes at least 20 steps. The convergence rate also varies with the magnitude of the second largest eigenvalue of W, and hence with the overall intensity of group cohesion. Consider, for example, the interaction structures in Figures 2, 3 and 5 with the second largest eigenvalues of $\mu_2 = 0.766$, $\mu_2 = 0$ and $\mu_2 = 0.2$ respectively. Fixing $\lambda = 0.8$, the convergence rate decreases with the second largest eigenvalue. We thus aim to derive the lower and upper bounds for the rate of decay of disagreement as a function of the intensities of group cohesion and prejudice.

$$W = \begin{bmatrix} 0.2 & 0.4 & 0. & 0.4 \\ 0.4 & 0.2 & 0.4 & 0. \\ 0. & 0.4 & 0.2 & 0.4 \\ 0.4 & 0. & 0.4 & 0.2 \end{bmatrix}$$

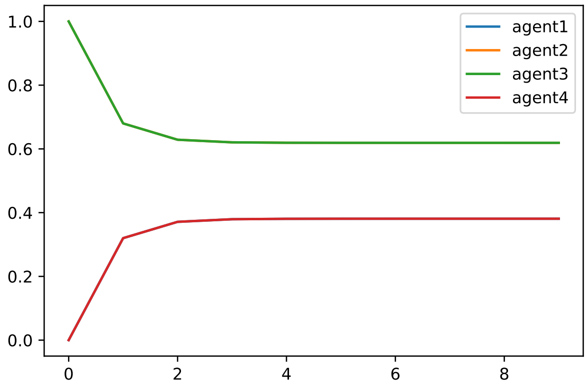


Figure 5. : The evolution of opinions for a cyclic interaction structure with the second largest eigenvalue of $\mu_2 = 0.2$. We take $\lambda = 0.8$ and the vector of prejudice/initial opinions to be $\mathbf{p}(0) = (0, 1, 1, 0)$. Agents 2 and 3 converge to the same opinion of 0.62 and agents 1 and 4 converge to the same opinion of 0.38.

We define the speed of convergence as the time it takes the learning process to get close to equilibrium within the time frame $\{0, 1, \cdots, T(r)\}$. For a fixed $r \in \{0, 1, \cdots, R\}$, and hence a fixed W, λ and $\bar{\mathbf{p}} = \bar{\mathbf{p}}(r)$, define the distance between the vector of opinions $\mathbf{p}(t)$ at t and equilibrium vector of opinions $\mathbf{p}^*$ as $DE(W; \lambda; \bar{\mathbf{p}}; t) = \|\mathbf{p}(t) - \mathbf{p}^*\|_{\mathbf{v}} = \left(\sum_{i=1}^n v_i \left(p_i(t) - p_i^*\right)^2\right)^{\frac{1}{2}}$, where as before, we take $\mathbf{v} = \pi$. Since $DE(W; \lambda; \bar{\mathbf{p}}; t)$ is a function of the vector of prejudice $\bar{\mathbf{p}}$, $\bar{\mathbf{p}}$ can be chosen in such a way that $\mathbf{p}(1)$, and hence $\mathbf{p}(t)$ for $t \geq 2$, is close to $\mathbf{p}^*$; under this scenario, learning stops after a few steps. We consider the worst possible scenario by taking a supremum over all possible vectors of prejudice/initial opinions, that is, $DE(W; \lambda; t) = \sup_{\bar{\mathbf{p}} \in [0,1]^n} DE(W; \lambda; \bar{\mathbf{p}}; t)$. We then define the convergence time $CT(W; \lambda; \varepsilon)$, for some small real number $\varepsilon > 0$, as the time it takes for the distance $DE(W; \lambda; t)$ to get below ε.

DEFINITION 2: *The convergence time $CT(W; \lambda; \varepsilon)$ to $\varepsilon > 0$ under interaction matrix W is*

$$CT(W; \lambda; \varepsilon) = \min\{t : \; DE(W; \lambda; t) < \varepsilon\} \tag{13}$$

Note that the convergence time captures the decay rate of disagreement in that as the distance between $\mathbf{p}(t)$ and $\mathbf{p}^*$ tends to zero, the distance between $\mathbf{p}(t)$ and the expected consensus vector $\mathbf{c}$ also tends to $D(W; \lambda)$—the long-run/equilibrium magnitude of disagreement.

In analogy to time frame $t \in \{0, 1, 2, \cdots, T(r)\}$, the convergence time is closely similar to $T(r)$. Specifically, the magnitude of the difference between $CT(W; \lambda; \varepsilon)$ and $T(r)$ decreases with ε, and the two measures are exactly identical when $\varepsilon = 0$. The definition of convergence time derives from the notion of mixing time, which is widely studied in the literature of Markov chains.

As in the case for the dynamics of disagreement in Section II, we examine the dynamics of the speed of learning over the time frame $\{0, 1, \cdots, R\}$ through comparative statics; that is, by varying λ and W. The following proposition establishes the lower and upper bounds for the convergence time.

PROPOSITION 2: *Let W be strongly connected and aperiodic. The convergence time in model* (5) *is bounded by*

(14)

$$\ln\left(\frac{\varepsilon\,(1-\lambda \mid \mu_2 \mid)}{\lambda\,(1- \mid \mu_2 \mid)}\right) \Big/ \ln\left(\lambda \mid \mu_2 \mid\right) \leq CT(W;\lambda;\varepsilon) \leq \ln\left(\frac{\varepsilon\pi_{\text{min}}^{\frac{1}{2}}\,(1-\lambda \mid \mu_2 \mid)}{2-\lambda\,(1+ \mid \mu_2 \mid)}\right) \Big/ \ln\left(\lambda \mid \mu_2 \mid\right)$$

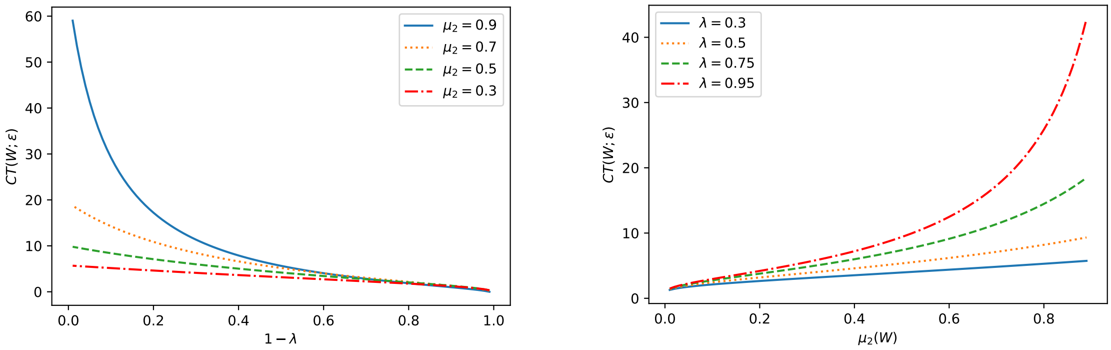
PROOF:

See Appendix A.A3

Proposition 2 shows that the convergence time decreases logarithmically with the intensity of prejudice. This is consistent with the examples presented above, whereby, when the intensity of prejudice is one (i.e. $(1-\lambda) = 1$), the learning process converges after one step of iteration, and the number of iterations increase with λ. The reason being that as the intensity of prejudice increases, agents place less and lesser weight on neighbours' opinions. When $(1-\lambda) = 1$, agents place zero weight on neighbours opinions, and hence, learning does not occur. The rate of decrease is depicted in Figure 6 (a), which plots the lower bound in (14) against the intensity of prejudice. As expected of logarithmic functions, the rate of decrease is stronger for lower values of the intensity of prejudice.

These results have implications for optimal persuasion-airtime allocations. By persuasion-airtime we mean, for example, airtime in political campaigns, court trials, and public programs campaigns. Persuasion-airtime is costly. If the objective of a political or public program campaign is to bring about a consensus, it is intuitive to think that the more airtime allocated, the better the outcome in terms of the proportion of the population that gets persuaded. Proposition 2 shows that in highly prejudiced groups, people make up their minds quickly and no amount of extra persuasion can help change their decisions; unless of course the extra persuasion is meant to change their prejudices.

Proposition 2 also states that the convergence time increases with the second largest eigenvalue of the interaction matrix. The rate of increase is logarithmic in $1 - \mu_2(W)$, so that the rate is highest at higher values of $\mu_2(W)$. This relationship is depicted in Figure 6 (b), that plots the convergence time against the second largest eigenvalue of W. A corollary to Proposition



(a) Convergence time versus intensity of group cohesion.

(b) Convergence time versus the second largest eigenvalue of W.

Figure 6. : The figure plots convergence time against the intensities of prejudice and the second largest eigenvalue of the interaction matrix using the expression of the lower and upper bounds in (14) respectively. We take $\varepsilon = 0.001$.

2 is that under the conditions stated in Proposition 1, the lower and upper bounds for the convergence time are increasing functions of the intensity of group cohesion. Specifically, if W consists of at least two cohesive subgroups whereby, $\bar{w}_{il} = \bar{w}_{jl}$ for all pairs $i, j \in L_k$, and that $\tilde{W}$ is symmetric, then $\mu_2(W) = \mu_2(\tilde{W})$ such that $\left(\frac{2\iota(W)}{K} - 1\right) \leq \mu_2(\tilde{W}) \leq 2\iota(W) - 1$ (see Lemmas 4 and 3). The lower and upper bounds for the convergence time can thus be expressed in terms of the overall intensity of group cohesion. And just as in the case of the second largest eigenvalue of W, the convergence time increases with the overall intensity of group cohesion.

As a final remark, we can conclude from Proposition 2 that in the absence of prejudice (i.e. $\lambda = 1$), which corresponds to the DeGroot learning model, the convergence time depends only on the second largest eigenvalue of the interaction matrix. The differences between the expressions for the convergence times can thus be used to differentiate between the DeGroot and Freidkin-Johnsen models of learning, and other related variations. For example, in a variation of the DeGroot model, according to which agents linearly combine their personal experiences with the opinions of their neighbours, Jadbabaie, Molavi and Tahbaz-Salehi (2013) show that the convergence time depends on agents' eigenvector centralities. It is therefore feasible to differentiate these two frameworks of learning but theoretically computing the rate of decay of disagreement (i.e. $\mathbf{p}(t-1) - \mathbf{p}(t)$) and fitting it to experimental data.

IV. Concluding remarks

The question of how peoples' attitudes and behaviours evolve is at the centre of social and behavioural sciences. The models of learning by averaging others' opinions or behaviours, commonly known as naïve learning, have been instrumental in explaining the processes of opinion formation and behavioural change. Empirical studies also suggest that people indeed tend to follow naïve rules of learning. The prediction of the canonical models of naïve learning is that the society will converge to a consensus in the long-run. This prediction is however not consistent with the observed persistence of disagreement in the society across factual issues. To reconcile the discrepancies between theoretical predictions and empirical observation, some papers in the literature have come up with modifications of the canonical models of naive learning that generate disagreement in equilibrium.

The objective of this paper has been to quantity the extent of disagreement and the speed of learning in the Friedkin-Johnson model of opinion formation that is known to generate disagreement in equilibrium. We start by demonstrating that the Friedkin-Johnson model coincides with an endogenous model of opinion formation where agents compromise between respecting their own personal prejudice and conforming their opinions to those held by others with whom they share close ties. We showed that the intensity of prejudice and of group cohesion interactively drive the extent of disagreement. We also demonstrate how the speed of learning can be used as a mechanism for distinguishing different models of learning by averaging. One aspect that is largely missing in the study of opinion and behaviour formation is empirical studies that attempt to establish which among the existing models best describes reality. Finally, we discussed policy implications of our findings.

REFERENCES

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference points and effort provision." *American Economic Review*, 101(2): 470–92.

Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. "Spread of (mis) information in social networks." *Games and Economic Behavior*, 70(2): 194–227.

Acemoğlu, Daron, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar. 2013. "Opinion fluctuations and disagreement in social networks." *Mathematics of Operations Research*, 38(1): 1–27.

Ansolabehere, Stephen, Jonathan Rodden, and James M Snyder. 2008. "The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting." *American Political Science Review*, 102(2): 215–232.

Asch, Solomon E, and H Guetzkow. 1951. "Effects of group pressure upon the modification and distortion of judgments." *Groups, leadership, and men*, 222–236.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou. 2006. "Who's who in networks. Wanted: The key player." *Econometrica*, 74(5): 1403–1417.

Barr, DR, and MU Thomas. 1977. "An eigenvector condition for Markov chain lumpability." *Operations Research*, 25(6): 1028–1031.

Bhatia, Rajendra. 2001. "Linear algebra to quantum cohomology: the story of Alfred Horn's inequalities." *The American Mathematical Monthly*, 108(4): 289–318.

Bindel, David, Jon Kleinberg, and Sigal Oren. 2015. "How bad is forming your own opinion?" *Games and Economic Behavior*, 92: 248–265.

Bonacich, Phillip. 1987. "Power and centrality: A family of measures." *American journal of sociology*, 92(5): 1170–1182.

Buchholz, Peter. 1994. "Exact and ordinary lumpability in finite Markov chains." *Journal of applied probability*, 31(1): 59–75.

Carpenter, Jeffrey P. 2004. "When in Rome: conformity and the provision of public goods." *The Journal of Socio-Economics*, 33(4): 395–408.

Chen, Yan, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. "Social comparisons and contributions to online communities: A field experiment on movielens." *The American economic review*, 100(4): 1358–1398.

DeGroot, Morris H. 1974. "Reaching a consensus." *Journal of the American Statistical Association*, 69(345): 118–121.

Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. "The spreading of misinformation online." *Proceedings of the National Academy of Sciences*, 113(3): 554–559.

DeMarzo, Peter M, Dimitri Vayanos, and Jeffrey Zwiebel. 2003. "Persuasion bias, social influence, and unidimensional opinions." *The Quarterly Journal of Economics*, 118(3): 909–968.

DiPrete, Thomas A, Andrew Gelman, Tyler McCormick, Julien Teitler, and Tian Zheng. 2011. "Segregation in social networks based on acquaintanceship and trust." *American Journal of Sociology*, 116(4): 1234–83.

Doran, Peter T, and Maggie Kendall Zimmerman. 2009. "Examining the scientific consensus on climate change." *Eos, Transactions American Geophysical Union*, 90(3): 22–23.

Falk, Armin, and Andrea Ichino. 2006. "Clean evidence on peer effects." *Journal of labor economics*, 24(1): 39–57.

Fazio, Russell H. 1986. “How do attitudes guide behavior.” *Handbook of motivation and cognition: Foundations of social behavior*, 1: 204–243.

Friedkin, Noah E, and Eugene C Johnsen. 1990. “Social influence and opinions.” *Journal of Mathematical Sociology*, 15(3-4): 193–206.

Garrett, R Kelly. 2009. “Echo chambers online?: Politically motivated selective exposure among Internet news users.” *Journal of Computer-Mediated Communication*, 14(2): 265–285.

Golub, Benjamin, and Matthew O. Jackson. 2010. “Nave Learning in Social Networks and the Wisdom of Crowds.” *American Economic Journal: Microeconomics*, 2(1): 112–49.

Golub, Benjamin, and Matthew O Jackson. 2012. “How homophily affects the speed of learning and best-response dynamics.” *The Quarterly Journal of Economics*, 127(3): 1287–1338.

Grönlund, Kimmo, Kaisa Herne, and Maija Setälä. 2015. “Does enclave deliberation polarize opinions?” *Political Behavior*, 37(4): 995–1020.

Hegselmann, Rainer, Ulrich Krause, et al. 2002. “Opinion dynamics and bounded confidence models, analysis, and simulation.” *Journal of artificial societies and social simulation*, 5(3).

Hogan, David E, and Michael Mallott. 2005. “Changing racial prejudice through diversity education.” *Journal of College Student Development*, 46(2): 115–125.

Horn, Roger A, and Charles R Johnson. 1990. *Matrix analysis.* Cambridge university press.

Houston, David A, and Russell H Fazio. 1989. “Biased processing as a function of attitude accessibility: Making objective judgments subjectively.” *Social cognition*, 7(1): 51–66.

Huckfeldt, Robert, Paul E Johnson, and John Sprague. 2004. *Political disagreement: The survival of diverse opinions within communication networks.* Cambridge University Press.

Jackson, Matthew O. 2010. *Social and economic networks.* Princeton university press.

Jadbabaie, Ali, Pooya Molavi, and Alireza Tahbaz-Salehi. 2013. “Information Heterogeneity and the Speed of Learning in Social Networks.” *Columbia Business School Research Paper 13-28.*

Kahan, Dan M, Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. “Motivated numeracy and enlightened self-government.” *Behavioural Public Policy*, 1(1): 54–86.

Kahan, Dan M, Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. 2012. "The polarizing impact of science literacy and numeracy on perceived climate change risks." *Nature climate change*, 2(10): 732–735.

Krause, Ulrich. 2000. "A discrete nonlinear and non-autonomous model of consensus formation." *Communications in difference equations*, 227–236.

Kulik, Carol T, and Loriann Roberson. 2008. "Common goals and golden opportunities: Evaluations of diversity education in academic and organizational settings." *Academy of Management Learning & Education*, 7(3): 309–331.

Levin, David Asher, Yuval Peres, and Elizabeth Lee Wilmer. 2009. *Markov chains and mixing times.* American Mathematical Soc.

Masson, Cees N, and Maykel Verkuyten. 1993. "Prejudice, ethnic identity, contact and ethnic group preferences among Dutch young adolescents." *Journal of Applied Social Psychology*, 23(2): 156–168.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual review of sociology*, 27(1): 415–444.

McPherson, Miller, Lynn Smith-Lovin, and Matthew E Brashears. 2006. "Social isolation in America: Changes in core discussion networks over two decades." *American sociological review*, 71(3): 353–375.

Melguizo, Isabel. 2016. "Endogenous homophily and the persistence of disagreement." *Mimeo.*

Merikoski, Jorma Kaarlo, and Ari Virtanen. 1997. "Bounds for eigenvalues using the trace and determinant." *Linear algebra and its applications*, 264: 101–108.

Mutz, Diana C. 2002. "Cross-cutting social networks: Testing democratic theory in practice." *American Political Science Review*, 96(1): 111–126.

Neal, Zachary P. 2018. "A sign of the times? Weak and strong polarization in the US Congress, 1973–2016." *Social Networks.*

Parsegov, Sergey E, Anton V Proskurnikov, Roberto Tempo, and Noah E Friedkin. 2017. "Novel multidimensional models of opinion dynamics in social networks." *IEEE Transactions on Automatic Control*, 62(5): 2270–2285.

Pettigrew, Thomas F, and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology*, 90(5): 751.

Poole, Keith T, and R Steven Daniels. 1985. "Ideology, party, and voting in the US Congress, 1959–1980." *American Political Science Review*, 79(2): 373–399.

Salganik, Matthew J, Peter Sheridan Dodds, and Duncan J Watts. 2006. "Experimental study of inequality and unpredictability in an artificial cultural market." *science*, 311(5762): 854–856.

Tesser, Abraham. 1993. "The importance of heritability in psychological research: The case of attitudes." *PSYCHOLOGICAL REVIEW-NEW YORK-*, 100: 129–129.

Wang, Bo-Ying, and Bo-Yan Xi. 1997. "Some inequalities for singular values of matrix products." *Linear algebra and its applications*, 264: 109–115.

Wilson, Timothy D, Samuel Lindsey, and Tonya Y Schooler. 2000. "A model of dual attitudes." *Psychological review*, 107(1): 101.

Wolkowicz, Henry, and George PH Styan. 1980. "Bounds for eigenvalues using traces." *Linear algebra and its applications*, 29: 471–506.

Zafar, Basit. 2011. "An experimental investigation of why individuals conform." *European Economic Review*, 55(6): 774–798.

Mathematical Appendix

A1. Proof of Proposition 1

The proof of Proposition 1 is split into lemmas. Let $1 = \mu_1(W) \geq \mu_2(W) \geq, \cdots, \geq \mu_n(W)$ represent the eigenvalues of W, and let $\mu_*(W) = \max\{\mid \mu_2(W) \mid, \mid \mu_n(W) \mid\}$. We start with Lemmas 1 and 2 below that relate $D(W;\lambda)$ to λ and $\mid \mu_2(W) \mid$.

LEMMA 1: *Let W be strongly connected and aperiodic. Then $\sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\|W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right\|_\pi$ is bounded by*

$$\text{(A1)} \qquad \mid \mu_2(W) \mid^t \leq \sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\|W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right\|_\pi \leq \pi_{\min}^{-\frac{1}{2}} \mid \mu_2(W) \mid^t$$

where $\pi_{\min} = \min_{i\in N} \pi_i$.

PROOF:

The proof relies on results from Levin, Peres and Wilmer (2009, Lemmas 12.1 & 12.2). From Section I.D, we restrict the interaction matrix W to be strongly connected (i.e irreducible) and aperiodic. Thus, from Levin, Peres and Wilmer (2009, Lemma 12.1):

(i) $| \mu_i(W) | \leq 1$ for all $i = 1, \cdots, n$, and -1 is not an eigenvalue of W;

(ii) The vector space of eigenvectors corresponding to the eigenvalue $\mu_1(W) = 1$ is the one-dimensional space generated by the column vector $\mathbf{e} := (1, 1, ..., 1)^T$.

Let π be the unique stationary distribution of the Markov chain $\mathbf{p}(t) = W^t\mathbf{p}(0)$, which also means that π is the left eigenvector of W corresponding to the eigenvalue $\mu_1(W) = 1$. By the assumptions on W above (i.e. irreducibility and aperiodicity), π is unique; we can in turn assume that W is reversible with respect to π. Define the inner product $\langle \mathbf{x}, \mathbf{r} \rangle_\pi$ for any two vectors $\mathbf{x}$ and $\mathbf{r}$ in the vector space $\mathbb{R}^n$ as $\langle \mathbf{x}, \mathbf{r} \rangle_\pi = \sum_{i=1}^n v_i r_i \pi_i$; and where no confuses arises, we write μ_i for $\mu_i(W)$. It then follows from Levin, Peres and Wilmer (2009, Lemma 12.2) that:

(a) The inner product space $(\mathbb{R}^n, \langle .,. \rangle_\pi)$ has an orthonormal basis of real-valued eigenfunctions $\{f_i\}_{i=1}^n$ corresponding to real eigenvalues $\{\mu_i\}$.

(b) The interaction matrix W can be decomposed as

$$w^t_{jk} = \pi_k + \sum_{i=2}^n \mu_i^t \pi_k f_i(j) f_i(k) \tag{A2}$$

Statement (a) above implies that for any eigenfunctions f_i and f_j, we have $\langle f_i, f_i \rangle_\pi = 1$ and $\langle f_i, f_j \rangle_\pi = 0$. Statement (b) on the other hand leads to the following relations.

$$\frac{w^t_{jk}}{\pi_k} - 1 = \sum_{i=2}^n \mu_i^t f_i(j) f_i(k) \tag{A3}$$

$$W^t\bar{\mathbf{p}} = \Pi\bar{\mathbf{p}} + \sum_{i=2}^n \mu_i^t f_i \sum_{k=1}^n \pi_k \bar{p}_k f_i(k) = \Pi\bar{\mathbf{p}} + \sum_{i=2}^n \mu_i^t \langle f_i, \bar{\mathbf{p}} \rangle_\pi f_i \tag{A4}$$

Using these definitions and concepts, we can now derive the upper and lower bounds for $\sup_{\bar{\mathbf{p}} \in [0,1]^n} \left\| W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}} \right\|_\pi$. For the upper bound, we have

$$\left\| W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}} \right\|_\pi^2 = \left\| \sum_{i=2}^n \mu_i^t \langle f_i, \bar{\mathbf{p}} \rangle_\pi f_i \right\|_\pi^2 = \sum_{i=2}^n \left\| \mu_i^t \langle f_i, \bar{\mathbf{p}} \rangle_\pi f_i \right\|_\pi^2 = \sum_{i=2}^n \mu_i^{2t} \langle f_i, \bar{\mathbf{p}} \rangle_\pi^2 \left\| f_i \right\|_\pi^2 \tag{A5}$$

where the second equality of (A5) follows from the application of the Pythagorean law to orthogonal vectors. That is, if $\{x_1, \cdots, x_k\}$ is an orthogonal set, then the Pythagorean Law states that

$$\|x_1 + \cdots + x_k\|^2 = \|x_1\|^2 + \cdots + \|x_k\|^2$$

Note however that

$$\|f_i\|_\pi^2 = \sum_{j=1}^n \pi_j f_i^2(j) = \langle f_i, f_i\rangle_\pi = 1$$

And

$$\text{(A6)} \qquad \langle f_i, \bar{\mathbf{p}}\rangle_\pi^2 = \left[\sum_{k=1}^n \pi_k \bar{p}_k f_i(k)\right]^2 \le \|f_i\|_\infty^2 \left[\sum_{k=1}^n \pi_k \bar{p}_k\right]^2$$

where $\|f_i\|_\infty = \max_k \mid f_i(k) \mid$ is the infinity norm of f_i. Taking a supremum over $\bar{\mathbf{p}} \in [0,1]^n$ yields

$$\sup_{\bar{\mathbf{p}}\in[0,1]^n} \langle f_i, \bar{\mathbf{p}}\rangle_\pi^2 \le \|f_i\|_\infty^2 \left[\sum_{k=1}^n \pi_k\right]^2 = \|f_i\|_\infty^2$$

Substituting into (A5) yields

$$\text{(A7)} \qquad \sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\|W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right\|_\pi^2 \le \sum_{i=2}^n \mu_i^{2t} \|f_i\|_\infty^2 \le \mu_*^{2t} \sum_{i=2}^n \|f_i\|_\infty^2$$

To derive expression for the summation on the right hand side of (A7), note that if $\delta_j(k)$ is defined as $\delta_j(k) = 1$ when $j = k$ and zero otherwise, then δ_j can be written via basis decomposition as

$$\delta_j = \sum_{i=1}^n \langle \delta_j, f_i\rangle_\pi f_i = \sum_{i=1}^n f_i(j)\pi_j f_i$$

We can in turn write π_j as

$$\pi_j = \langle \delta_j, \delta_j\rangle_\pi = \left\langle \sum_{i=1}^n f_i(j)\pi_j f_i\,, \sum_{i=1}^n f_i(j)\pi_j f_i \right\rangle_\pi = \pi_j^2 \sum_{i=1}^n f_i(j)^2$$

where the second equality follows from the fact that $\langle f_i, f_i\rangle_\pi = 1$. Hence, $\sum_{i=1}^n f_i(j)^2 = \pi_j^{-1}$ and $\sum_{i=2}^n f_i(j)^2 \le \pi_j^{-1}$; consequently, $\sum_{i=2}^n \|f_i\|_\infty^2 \le \pi_{\min}^{-1}$, where $\pi_{\min} = \min_j \pi_j$. Substituting into (A7) and taking the square root of both sides yields the upper bounds

$$\text{(A8)} \qquad \sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\|W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right\|_\pi \le \pi_{\min}^{-\frac{1}{2}} \mid \mu_* \mid^t$$

But since $\mid \mu_2(W) \mid \ge \mid \mu_n(W) \mid$ for most strongly connected and aperiodic networks, it follows that $\sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\|W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right\|_\pi \le \pi_{\min}^{-\frac{1}{2}} \mid \mu_2 \mid^t$.

To derive the lower bound, first note that substituting the first equality of (A6) into (A5) yields

$$\left\|W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right\|_\pi^2 = \sum_{i=2}^{n} \mu_i^{2t} \langle\, f_i, \bar{\mathbf{p}}\rangle_\pi^2 \tag{A9}$$

Without loss of generality, we can choose $\bar{\mathbf{p}}$ to be equal or parallel to f_2 so that $\langle\, f_2, \bar{\mathbf{p}}\rangle_\pi = 1$, but for all $f_i \neq f_2$, we have $\langle\, f_i, \bar{\mathbf{p}}\rangle_\pi = 0$. This way, (A9) reduces to

$$\sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\|W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right\|_\pi \geq\mid \mu_2 \mid^t \tag{A10}$$

This completes the proof of Lemma 1.

LEMMA 2: *Let W be strongly connected and aperiodic. Then $D(W;\lambda)$ is bounded by*

$$\left(\frac{1-\lambda}{1-\lambda\mid \mu_2\mid}\right) \leq D(W;\lambda) \leq \pi_{\min}^{-\frac{1}{2}}\left(\frac{1-\lambda}{1-\lambda\mid \mu_2\mid}\right) \tag{A11}$$

PROOF:

Recall from (6) that when $\lambda_i = \lambda$ for all $i \in N$, then

$$\mathbf{p}^* = (1-\lambda)\left[(I-\lambda W)^{-1}\right]\bar{\mathbf{p}} = (1-\lambda)\left[\sum_{\tau=0}^{+\infty}(\lambda W)^\tau\right]\bar{\mathbf{p}} = (1-\lambda)\sum_{\tau=0}^{+\infty}\lambda^\tau W^\tau \bar{\mathbf{p}} \tag{A12}$$

The consensus vector can also be written in the form of (A12) above. Recall that $\mathbf{c} = \Pi\bar{\mathbf{p}}$, where we use the assumption of equivalence between $\mathbf{p}(0)$ and $\bar{\mathbf{p}}$. Note that the matrix $\Pi = \mathbf{e}\pi^T$ is derived by infinitely iterating W^t, so that $\Pi^t = \Pi$ for any $t = 1, 2, \cdots$. The following relation then holds.

$$\Pi = (1-\lambda)\sum_{\tau=0}^{\infty}(\lambda\Pi)^\tau = (1-\lambda)\Pi\sum_{\tau=0}^{\infty}(\lambda)^\tau = \Pi \tag{A13}$$

where the last equality is because $\sum_{\tau=0}^{\infty}\lambda^\tau = \frac{1}{1-\lambda}$. The magnitude of disagreement can thus be rewritten as

$$\begin{aligned} DP(W;\lambda) &= \sup_{\bar{\mathbf{p}}\in[0,1]^n} \|\mathbf{p}^* - \mathbf{c}\|_\pi \\ &= \sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\|(1-\lambda)\sum_{\tau=0}^{\infty}\lambda^\tau W^\tau\bar{\mathbf{p}} - (1-\lambda)\sum_{\tau=0}^{\infty}\lambda^\tau \Pi^\tau \bar{\mathbf{p}}\right\|_\pi \\ &= (1-\lambda)\sup_{\bar{\mathbf{p}}\in[0,1]^n}\left\|\sum_{\tau=0}^{\infty}\lambda^\tau\left(W^\tau\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}}\right)\right\|_\pi \end{aligned} \tag{A14}$$

Applying the triangular inequality and the results of Lemma 1 yields

$$\begin{aligned} DP(W;\lambda) &\leq (1-\lambda)\sum_{\tau=0}^{\infty}\lambda^{\tau}\left(\sup_{\bar{\mathbf{p}}\in[0,1]^n}\|W^{\tau}\bar{\mathbf{p}}-\Pi\bar{\mathbf{p}}\|_{\pi}\right) \\ &\leq (1-\lambda)\pi_{\min}^{-\frac{1}{2}}\sum_{\tau=0}^{\infty}(\lambda\mid\mu_2\mid)^{\tau} \\ &= \pi_{\min}^{-\frac{1}{2}}\left(\frac{1-\lambda}{1-\lambda\mid\mu_2\mid}\right) \end{aligned} \tag{A15}$$

where the last equality is because $\sum_{\tau=0}^{\infty}(\lambda\mid\mu_2\mid)^{\tau}=\frac{1}{1-\lambda|\mu_2|}$.

For the lower bound, recall the assumption made in deriving the lower bound for the $\sup_{\bar{\mathbf{p}}\in[0,1]^n}\left\|W^t\bar{\mathbf{p}}-\Pi\bar{\mathbf{p}}\right\|_{\pi}$ in (A10); that is, $\bar{\mathbf{p}}=f_i$. This assumption implies that $W^t\bar{\mathbf{p}}-\Pi\bar{\mathbf{p}}=\mu_2^tf_2$ so that the only variable terms in the second equality of (A14) are λ^{τ} and μ_2^{τ}. Using this assumption, the magnitude of disagreement then becomes

$$DP(W;\lambda)\geq(1-\lambda)\left\|\sum_{\tau=0}^{\infty}\lambda^{\tau}\mid\mu_2\mid^{\tau}f_2\right\|_{\pi}=(1-\lambda)\sum_{\tau=0}^{\infty}\lambda^{\tau}\mu_2^{\tau}\left\|f_2\right\|_{\pi} \tag{A16}$$

Substituting for $\|f_i\|_{\pi}=(\langle f_i,f_i\rangle_{\pi})^{\frac{1}{2}}=1$ and $\sum_{\tau=0}^{\infty}(\lambda\mu_2)^{\tau}=\frac{1}{1-\lambda\mu_2}$ yields

$$DP(W;\lambda)\geq\left(\frac{1-\lambda}{1-\lambda\mid\mu_2\mid}\right) \tag{A17}$$

This completes the proof of Lemma 2.

The next step of the proof establishes the relationship between the spectra of matrices W and $\tilde{W}$.

LEMMA 3: *Let $\mu_2(W)$ and $\mu_2(\tilde{W})$ be the respective second largest eigenvalues of W and $\tilde{W}$. If for all pairs $i,j\in L_k$, $\bar{w}_{il}=\bar{w}_{jl}$ for each subgroup L_l, then $\mu_2(W)=\mu_2(\tilde{W})$.*

PROOF:

Recall that $\mathcal{L}=\{L_1,L_2,\cdots,L_k\}$ is the set of disjoint cohesive subgroups of W, with n_l as the respective cardinality of L_l. Given the partition $\mathcal{L}$, let V be an $n\times K$ *collector matrix* with elements $v_{il}=1$ if $i\in L_l$, and zero otherwise. It follows from the definition of $\bar{W}$ that $\bar{W}=WV$; that is, the element in the ith row and lth column of WV is given by $\bar{w}_{il}=\sum_{j\in L_l}w_{ij}$.

By definition of V above, its column vectors are linearly independent. That is, if V_i is the ith column of V, then for some scalars a_i for $i=1,\cdots,K$,

$$a_1V_1+a_2V_2+\cdots a_KV_K=0$$

if and only if $a_1 = a_2 = \cdots = a_K = 0$. Linear independence of the columns of V then implies that V has a pseudo inverse defined as

$$V^{-1} = \left(VV^T\right)^{-1} V^T$$

The matrix V^{-1} is of size $K \times n$, with elements $\frac{1}{n_l}$ if i belongs to subgroup L_l and zero otherwise. Given V^{-1} the matrix $\tilde{W}$ can thus be rewritten as

$$\tilde{W} = V^{-1}WV = V^{-1}\bar{W}$$

That is, the element in the kth row and lth column of $V^{-1}\bar{W}$ is given by $\frac{1}{n_k}\sum_{i\in L_k}\sum_{j\in L_l} w_{ij} = \tilde{w}_{kl}$.

The conditions stated in Lemma 3 (i.e. for all pairs $i, j \in L_k$, $\bar{w}_{il} = \bar{w}_{jl}$ for each subgroup L_l) are also the necessary condition for *strong lumpability*. Let e_i be the ith basis vector, that is, a row vector of zeroes except a one in the ith coordinate. Then the ith row $\bar{w}_i$ of $\bar{W}$ can be expressed as $\bar{w}_i = e_i WV$. A matrix W is then said to be strongly lumpable under $\mathcal{L}$ into another matrix $\tilde{W}$ if $(e_i - e_j)WV = \mathbf{0}$ for all $i, j \in L_l \in \mathcal{L}$ (Buchholz, 1994); where $\mathbf{0}$ is a row vector of zeroes. Thus, strong lumpability is equivalent to requiring $\bar{w}_{il} = \bar{w}_{jl}$ for all pairs $i, j \in L_k$ and for each subgroup $L_l \in \mathcal{L}$.

The equality $\mu_2(W) = \mu_2(\tilde{W})$ then follows from Barr and Thomas (1977, Theorem 1), who show that if W is lumpable to $\tilde{W}$, then the eigenvalues of $\tilde{W}$ are eigenvalues of W, so that $\mu_2(W) = \mu_2(\tilde{W})$.

The final step of the proof establishes the relationship between the intensity of cohesion $\iota(W)$ of subgroups in W, which is the trace of matrix $\tilde{W}$, and the second largest eigenvalue of $\tilde{W}$.

LEMMA 4: *If $\tilde{W}$ is symmetric, that is, $\tilde{w}_{kl} = \tilde{w}_{lk}$ for each pair $L_k, L_l \in \mathcal{L}$, then its second largest eigenvalue $\mu_2(\tilde{W})$ is bounded by*

$$\left(\frac{2\iota(W)}{K} - 1\right) \leq \mu_2(\tilde{W}) \leq 2\iota(W) - 1 \tag{A18}$$

PROOF:

Let $\mathrm{Tr}(\tilde{W})$ be the trace of $\tilde{W}$ so that $\mathrm{Tr}(\tilde{W}) = \iota(W)$. If $\tilde{W}$ is symmetric, that is, $\tilde{w}_{kl} = \tilde{w}_{lk}$ for each pair $L_k, L_l \in \mathcal{L}$, then from Wolkowicz and Styan (1980, Theorem 2.2)

$$\frac{\iota(W)}{K} \leq \frac{\mu_1(\tilde{W}) + \mu_2(\tilde{W})}{2} \leq \frac{\iota(W)}{K} + \left(\frac{K-2}{2}\right)^{\frac{1}{2}} \left(\frac{\iota^2(W)}{K} - \left(\frac{\iota(W)}{K}\right)^2\right)^{\frac{1}{2}} \tag{A19}$$

The right hand side inequality of (A19) simplifies to

$$\begin{aligned}\frac{\iota(W)}{K}+\left(\frac{K-2}{2}\right)^{\frac{1}{2}}\left(\frac{\iota^2(W)}{K}-\left(\frac{\iota(W)}{K}\right)^2\right)^{\frac{1}{2}} &= \iota(W)\left(\frac{1}{K}+\left(\frac{K-2}{2}\right)^{\frac{1}{2}}\left(\frac{K-1}{K^2}\right)^{\frac{1}{2}}\right)\\ &\leq \iota(W)\left(\frac{1}{K}+\left(\frac{(K-1)^2}{K^2}\right)^{\frac{1}{2}}\right)\\ &= \iota(W)\left(\frac{1}{K}+\frac{(K-1)}{K}\right)=\iota(W)\end{aligned}$$

Substituting into (A19) yields

$$\text{(A20)}\qquad \frac{\iota(W)}{K}\leq\frac{\mu_1(\tilde{W})+\mu_2(\tilde{W})}{2}\leq\iota(W)$$

Note however that, just like W, $\tilde{W}$ is also row stochastic. To see why, recall that the kth row and lth column of $\tilde{W}$ is given by $\tilde{w}_{kl}=\frac{1}{n_k}\sum_{i\in L_k}\sum_{j\in L_l}w_{ij}$. The sum of all elements in the kth column of $\tilde{W}$ is then given by

$$\sum_{l=1}^{K}\tilde{w}_{kl}=\sum_{l=1}^{K}\frac{1}{n_k}\sum_{i\in L_k}\sum_{j\in L_l}w_{ij}=\frac{1}{n_k}\sum_{i\in L_k}\sum_{l=1}^{K}\sum_{j\in L_l}w_{ij}=1$$

where the last equality follows because $\sum_{l=1}^{K}\sum_{j\in L_l}w_{ij}=\sum_{j\in N}w_{ij}=1$, and $\frac{1}{n_k}\sum_{i\in L_k}=1$. Row stochasticity of $\tilde{W}$ then implies that $\mu_1(\tilde{W})=1$, such that

$$\text{(A21)}\qquad \left(\frac{2\iota(W)}{K}-1\right)\leq\mu_2(\tilde{W})\leq 2\iota(W)-1$$

A2. The relationship between $\mu_2(W)$ and $\mu_2(\tilde{W})$

This section aims to show that the second largest eigenvalues of W and $\tilde{W}$ are close to identical if the structure of W does not deviate by much from the conditions of strong lumpability stated in Lemma 3: that is, $\bar{w}_{il}=\bar{w}_{jl}$ for all pairs $i,j\in L_k$ and for each subgroup L_l. Recall that this conditions can be equivalently stated as

$$\text{(A22)}\qquad (e_i-e_j)WV=0\quad\text{for all } i,j\in L_l\in\mathcal{L}.$$

where e_i is a basis vector, that is, a vector of zeroes except a one in the ith coordinate.

To relax condition (A22), we consider small deviations from strong lumpability and define a related notion of *near lumpability*. For some small $\varepsilon>0$, an interaction matrix W is near lumpable if it can be expressed as $W = P + \varepsilon R$, where P is strongly lumpable and R is an arbitrary matrix. In analogy to (A22), W is near lumpability if for some small $\varepsilon > 0$,

$$(e_i - e_j)WV \leq \varepsilon \mathbf{e} \quad \text{for all } i, j \in L_l \in \mathcal{L}. \tag{A23}$$

PROPOSITION 3: *Let W be symmetric and nearly lumpable with respect to a partition $\mathcal{L}$ to $\tilde{W}$ so that $W = P + \varepsilon R$, where P is strongly lumpable and R is some arbitrary matrix. Let $R_{\max} = \max_{i \in N} \sum_{j=1}^{n} r_{ij}$ be the maximum row-sum of R. Then*

$$\mid \mu_2(W) - \mu_2(\tilde{W}) \mid \leq \varepsilon R_{\max} \left(1 + \frac{1}{\mid \mu_n(R) \mid}\right) \tag{A24}$$

where $\mu_n(R)$ is the magnitude of the smallest eigenvalue of R.

Proposition 3 demonstrates that if W is nearly lumpable, then its second largest eigenvalue does not deviate by much from that of $\tilde{W}$. Under this condition, ε and/or $R_{\max}$ are sufficiently small that the right hand side of (A24) is very small. The result in Proposition 3 is valid for the case when W is symmetric. That is, for any pair of agents i and j, $w_{ij} = w_{ji}$. Symmetry of W is a mild assumption that makes the derivation of the relation in (A24) significantly less cumbersome. Relaxing this assumption leads to highly complex expressions with less added value qualitatively.

Proof of Proposition 3

The definition of near lumpability in (A23) implies that W can be rewritten as $W = P + \varepsilon R$, where P is strongly lumpable and R is some arbitrary transition matrix. Let $U = V^{-1}$. Multiplying W with U on the left and V on the right hand sides yields

$$\tilde{W} = UWV = UPV + \varepsilon URV = \tilde{P} + \varepsilon URV \tag{A25}$$

We apply the following inequalities on eigenvalues to establish the relationship between $\mu_2(W)$ and $\mu_2(\tilde{W})$. Let A and B be $n \times n$ non-negative symmetric matrices with eigenvalues $a_1 \geq \cdots \geq a_n$ and $b_1 \geq \cdots \geq b_n$ respectively. Then the eigenvalues $c_1 \geq \cdots \geq c_n$ of $C = A + B$ have the following bounds (Bhatia, 2001).

$$c_{i+j-1} \leq a_i + b_j \quad \text{whenever } 0 < i, j, i + j < n, \text{ and} \quad c_i \geq a_i + b_n \quad \text{for } 0 \leq i \leq n. \tag{A26}$$

Since W is non-negative and symmetric, P, R, $\tilde{P}$ and URV must all be symmetric. Letting $i = 1$ and $j = 2$, it follows from the first inequality of (A26) that

$$\mu_2(W) \leq \mu_2(P) + \varepsilon\mu_1(R) \tag{A27}$$

and from the second inequality

$$\mu_2(\tilde{W}) \geq \mu_2(\tilde{P}) + \varepsilon\mu_k(URV) \tag{A28}$$

Since $\mu_2(P) = \mu_2(\tilde{P})$ by virtue of lumpability of P, it follows that

$$\mid \mu_2(W) - \mu_2(\tilde{W}) \mid \leq \mid \varepsilon\mu_1(R) - \varepsilon\mu_k(URV) \mid \leq \varepsilon \mid \mu_1(R) \mid + \varepsilon \mid \mu_k(URV) \mid \tag{A29}$$

To establish the relationship between $\mu_1(R)$ and $\mu_k(URV)$, we first examine the eigenvalues of VUR, and then later apply the relation $\mu_k(URV) = \mu_k(VUR)$. This relation follows from Horn and Johnson (1990, Theorem 1.3.20); that is, Suppose that $A \in \mathbb{M}^{m,n}$ and $B \in \mathbb{M}^{n,m}$, with $m \leq n$. Then the n eigenvalues of BA are the m eigenvalues of AB together with $n - m$ zeroes. We then use the following lemma for bounding $\mu_k(VUR)$.

LEMMA 5: *(Wang and Xi, 1997, Lemma 2) Let $G, H \in \mathbb{C}^{n\times n}$ be positive definite Hermitian, and let $1 \leq i_1, \cdots, i_l \leq n$. Then*

$$\prod_{\tau=1}^{l} \mu_{i_\tau}(GH) \leq \prod_{\tau=1}^{l} \mu_{i_\tau}(G)\mu_{i_\tau}(H) \tag{A30}$$

$$\prod_{\tau=1}^{l} \mu_{i_\tau}(GH) \geq \prod_{\tau=1}^{l} \mu_{i_\tau}(G)\mu_{n-\tau+1}(H) \tag{A31}$$

The inequality (A30) implies that

(A32)

$$\mu_1(VUR).\mu_2(VUR).\cdots.\mu_k(VUR) \leq \left[\mu_1(VU).\mu_2(VU).\cdots.\mu_k(VU)\right]\left[\mu_1(R).\mu_2(R).\cdots.\mu_k(R)\right]$$

The matrix VU is of size n and consists of k diagonal block matrices; it thus consists of the first k eigenvalues equal to one. That is $\mu_1(VU) = \mu_2(VU) = \cdots = \mu_k(VU) = 1$, so that

$$\mu_1(VUR).\mu_2(VUR).\cdots.\mu_k(VUR) \leq \mu_1(R).\mu_2(R).\cdots.\mu_k(R) \tag{A33}$$

and that

$$\mu_k(VUR) \leq \frac{\mu_1(R).\mu_2(R).\cdots.\mu_k(R)}{\mu_1(VUR).\mu_2(VUR).\cdots.\mu_{k-1}(VUR)} \tag{A34}$$

Using (A31), $\mu_{k-1}(VUR)$ is bounded from below by

$$\begin{aligned}&\mu_1(VUR).\mu_2(VUR).\cdots.\mu_{k-1}(VUR)\\ &\qquad \geq \left[\mu_1(VU).\mu_2(VU).\cdots.\mu_{k-1}(VU)\right]\left[\mu_{n-k+2}(R).\mu_{n-k+3}(R).\cdots.\mu_n(R)\right]\\ &\qquad = \mu_{n-k+2}(R).\mu_{n-k+3}(R).\cdots.\mu_n(R)\end{aligned}$$

so that

$$\mu_{k-1}(VUR) \geq \frac{\mu_{n-k+2}(R).\mu_{n-k+3}(R).\cdots.\mu_n(R)}{\mu_1(VUR).\mu_2(VUR).\cdots.\mu_{k-2}(VUR)}$$

Substituting into (A34) yields

$$\mu_k(VUR) \leq \frac{\mu_1(R).\mu_2(R).\cdots.\mu_k(R)}{\mu_{n-k+2}(R).\mu_{n-k+3}(R).\cdots.\mu_n(R)} \leq \frac{k\mu_1(R)}{(k-1)\mu_n(R)} \leq \frac{\mu_1(R)}{\mu_n(R)} \tag{A35}$$

And hence

$$\mid \mu_2(W) - \mu_2(\tilde{W}) \mid \leq \varepsilon \mid \mu_1(R) \mid \left(1 + \frac{1}{\mid \mu_n(R) \mid}\right) \tag{A36}$$

Since R is non-negative, it follows from Gershgorin circle theorem that $\mid \mu_1(R) \mid \leq R_{\max} = \max_{i\in N} \sum_{j=1}^n r_{ij}$.

A3. *Proof of Proposition 2*

We start by showing that

$$\mathbf{p}(t) = (1-\lambda)\sum_{\tau=0}^{t-1}(\lambda W)^\tau \bar{\mathbf{p}} + (\lambda W)^t \bar{\mathbf{p}} \tag{A37}$$

The expression on the right hand side of (A37) is a generalization of an iterative process. That is,

$$\mathbf{p}(1) = \lambda W\bar{\mathbf{p}} + (1-\lambda)\bar{\mathbf{p}} = \big(\lambda W + I - \lambda I\big)\bar{\mathbf{p}}$$

$$\mathbf{p}(2) = \lambda W\mathbf{p}(1) + (1-\lambda)\bar{\mathbf{p}} = \big((\lambda W)^2 + \lambda W + I - \lambda^2 W - \lambda I\big)\bar{\mathbf{p}}$$

$$\mathbf{p}(3) = \lambda W\mathbf{p}(2) + (1-\lambda)\bar{\mathbf{p}} = \big((\lambda W)^3 + (\lambda W)^2 + \lambda W + I - \lambda^3 W^2 - \lambda^2 W - \lambda I\big)\bar{\mathbf{p}}$$

$$\mathbf{p}(4) = \lambda W\mathbf{p}(3) + (1-\lambda)\bar{\mathbf{p}} = \big((\lambda W)^4 + (\lambda W)^3 + (\lambda W)^2 + \lambda W + I - \lambda^4 W^3 - \lambda^3 W^2 - \lambda^2 W - \lambda I\big)\bar{\mathbf{p}}$$

$$-----------------------------------$$

$$\mathbf{p}(t) = \big((\lambda W)^t + (\lambda W)^{t-1} + \cdots + \lambda W + I - \lambda^t W^{t-1} - \lambda^{t-1} W^{t-2} - \cdots - \lambda^2 W - \lambda I\big)\bar{\mathbf{p}} = (1-\lambda)\sum_{\tau=0}^{t-1}(\lambda W)^\tau\bar{\mathbf{p}} + (\lambda W)^t\bar{\mathbf{p}}$$

We next derive the lower and upper bounds for the distance $DE(W;t)$ between the vectors $\mathbf{p}(t)$ and $\mathbf{p}^*$, where as before

(A38)
$$\mathbf{p}^* = (1-\lambda)\sum_{\tau=0}^{\infty}(\lambda W)^\tau\bar{\mathbf{p}}$$

LEMMA 6: *The distance $DE(W;\lambda;t)$ to the long-run distribution after t iterations is bounded by*

(A39)
$$\frac{\lambda\,(1-\mid\mu_2\mid)}{1-\lambda\mid\mu_2\mid}\,(\lambda\mid\mu_2\mid)^t \le DE(W;\lambda;t) \le \pi_{\text{min}}^{-\frac{1}{2}}\left[\frac{2-\lambda\,(1+\mid\mu_2\mid)}{1-\lambda\mid\mu_2\mid}\right](\lambda\mid\mu_2\mid)^t$$

PROOF:

As a starting point, note that Π can be rewritten as

(A40)
$$\Pi = \frac{1-\lambda}{\lambda^t}\sum_{\tau=t}^{\infty}(\lambda\Pi)^\tau = \frac{1-\lambda}{\lambda^t}\Pi\sum_{\tau=t}^{\infty}(\lambda)^\tau = \frac{1-\lambda}{\lambda^t}.\frac{\lambda^t}{1-\lambda}\Pi = \Pi$$

Thus, $\lambda^t\Pi^t = \lambda^t\Pi = (1-\lambda)\sum_{\tau=t}^{\infty}(\lambda\Pi)^\tau$. We can then express $\mathbf{p}^* - \mathbf{p}(t)$ as

$$\begin{aligned}\mathbf{p}^* - \mathbf{p}(t) &= (1-\lambda)\sum_{\tau=t}^{\infty}(\lambda W)^\tau\bar{\mathbf{p}} - (\lambda W)^t\bar{\mathbf{p}}\\ &= (1-\lambda)\sum_{\tau=t}^{\infty}(\lambda W)^\tau\bar{\mathbf{p}} - (1-\lambda)\sum_{\tau=t}^{\infty}(\lambda\Pi)^\tau\bar{\mathbf{p}} + \lambda^t\Pi^t\bar{\mathbf{p}} - (\lambda W)^t\bar{\mathbf{p}}\\ &= (1-\lambda)\sum_{\tau=t}^{\infty}\lambda^\tau\left(W^\tau - \Pi\right)\bar{\mathbf{p}} - \lambda^t\left(W^t - \Pi\right)\bar{\mathbf{p}}\end{aligned}$$

The upper bound for the distance to equilibrium $DE(W;\lambda;t)$ is then given by

$$\begin{aligned} DE(W;\lambda;t) &= \sup_{\bar{\mathbf{p}}\in[0,1]^n} \left\| (1-\lambda)\sum_{\tau=t}^{\infty}\lambda^{\tau}\left(W^{\tau}-\Pi\right)\bar{\mathbf{p}} - \lambda^{t}\left(W^{t}-\Pi\right)\bar{\mathbf{p}} \right\|_{\pi} \\ &\leq (1-\lambda)\sum_{\tau=t}^{\infty}\lambda^{\tau}\left(\sup_{\bar{\mathbf{p}}\in[0,1]^n}\left\|W^{\tau}\bar{\mathbf{p}}-\Pi\bar{\mathbf{p}}\right\|_{\pi}\right) + \lambda^{t}\sup_{\bar{\mathbf{p}}\in[0,1]^n}\left\|W^{t}\bar{\mathbf{p}}-\Pi\bar{\mathbf{p}}\right\|_{\pi} \\ &\leq (1-\lambda)\pi_{\min}^{-\frac{1}{2}}\left[\sum_{\tau=t}^{\infty}\left(\lambda\mid\mu_2\mid\right)^{\tau} + \frac{\lambda^{t}}{1-\lambda}\mid\mu_2\mid^{t}\right] \\ &= (1-\lambda)\pi_{\min}^{-\frac{1}{2}}\left[\frac{\left(\lambda\mid\mu_2\mid\right)^{t}}{1-\lambda\mid\mu_2\mid} + \frac{\left(\lambda\mid\mu_2\mid\right)^{t}}{1-\lambda}\right] \\ &= \pi_{\min}^{-\frac{1}{2}}\left[\frac{2-\lambda\left(1+\mid\mu_2\mid\right)}{1-\lambda\mid\mu_2\mid}\right]\left(\lambda\mid\mu_2\mid\right)^{t} \end{aligned} \tag{A41}$$

where we wrote μ_2 for $\mu_2(W)$, the second inequality follows from triangular inequalities for sums, the third inequality follows from Lemma 1.

The derivation of the lower bound follows similar steps as in the proof of the lower bound for the magnitude of disagreement. Specifically, let $\bar{\mathbf{p}} = f_i$ without loss of generality so that $W^t\bar{\mathbf{p}} - \Pi\bar{\mathbf{p}} = \mu_2^t f_2$. The only variable terms in (A41) are λ^τ and μ_2^τ, and hence,

$$\begin{aligned} DE(W;\lambda;t) &= \left\| (1-\lambda)\sum_{\tau=t}^{\infty}\lambda^{\tau}\mid\mu_2\mid^{\tau} f_2 - \lambda^{t}\mid\mu_2\mid^{t} f_2 \right\|_{\pi} \\ &= \left\| \left[(1-\lambda)\sum_{\tau=t}^{\infty}\lambda^{\tau}\mid\mu_2\mid^{\tau} - \lambda^{t}\mid\mu_2\mid^{t}\right] f_2 \right\|_{\pi} \\ &= \left| (1-\lambda)\sum_{\tau=t}^{\infty}\lambda^{\tau}\mid\mu_2\mid^{\tau} - \lambda^{t}\mid\mu_2\mid^{t} \right| \left\|f_2\right\|_{\pi} \\ &= \left| \frac{1-\lambda}{1-\lambda\mid\mu_2\mid}\left(\lambda\mid\mu_2\mid\right)^{t} - \left(\lambda\mid\mu_2\mid\right)^{t} \right| \\ &= \left| \frac{-\lambda\left(1-\mid\mu_2\mid\right)}{1-\lambda\mid\mu_2\mid} \right| \left(\lambda\mid\mu_2\mid\right)^{t} \\ &= \frac{\lambda\left(1-\mid\mu_2\mid\right)}{1-\lambda\mid\mu_2\mid}\left(\lambda\mid\mu_2\mid\right)^{t} \end{aligned} \tag{A42}$$

where the fourth equality os because $\left\|f_2\right\|_{\pi} = 1$, and the last equality follows from the fact that $\mid\mu_2\mid\leq 1$.

To derive the upper bound for the convergence time, note that when $DE(W;\lambda;t) \leq \varepsilon$,

$$t \geq \left(\ln\left(\varepsilon\right) - \ln\left(\pi_{\min}^{-\frac{1}{2}}\left[\frac{2-\lambda\left(1+\mid\mu_2\mid\right)}{1-\lambda\mid\mu_2\mid}\right]\right)\right)\Big/\ln\left(\lambda\mid\mu_2\mid\right) = \ln\left(\frac{\varepsilon\pi_{\min}^{\frac{1}{2}}\left(1-\lambda\mid\mu_2\mid\right)}{2-\lambda\left(1+\mid\mu_2\mid\right)}\right)\Big/\ln\left(\lambda\mid\mu_2\mid\right)$$

From the definition of $CT(W;\lambda;\varepsilon)$ as the minimum t at which $DE(W;\lambda;t) \leq \varepsilon$, it follows that

$$CT(W;\lambda;\varepsilon) \leq \ln\left(\frac{\varepsilon\pi_{\min}^{\frac{1}{2}}\left(1-\lambda\mid\mu_2\mid\right)}{2-\lambda\left(1+\mid\mu_2\mid\right)}\right)\Big/\ln\left(\lambda\mid\mu_2\mid\right)$$

Similarly, when $DE(W;\lambda;t) \geq \varepsilon$, then

$$t \leq \left(\ln\left(\varepsilon\right)-\ln\left(\left[\frac{\lambda\left(1-\mid\mu_2\mid\right)}{1-\lambda\mid\mu_2\mid}\right]\right)\right)\Big/\ln\left(\lambda\mid\mu_2\mid\right) = \ln\left(\frac{\varepsilon\left(1-\lambda\mid\mu_2\mid\right)}{\lambda\left(1-\mid\mu_2\mid\right)}\right)\Big/\ln\left(\lambda\mid\mu_2\mid\right)$$

Thus, by definition of $CT(W;\varepsilon)$, we have

$$CT(W;\lambda;\varepsilon) \geq \ln\left(\frac{\varepsilon\left(1-\lambda\mid\mu_2\mid\right)}{\lambda\left(1-\mid\mu_2\mid\right)}\right)\Big/\ln\left(\lambda\mid\mu_2\mid\right)$$