Essays on social learning and social dynamics

Inaugural-Dissertation
zur Erlangung des Doktorgrades
des Fachbereichs Wirtschaftswissenschaften
der Johann Wolfgang Goethe-Universität
Frankfurt am Main

vorgelegt von
Steffen Eger
aus
Oberndorf am Neckar

Frankfurt am Main, 2013

Erstgutachter: Professor Matthias Blonski

Zweitgutachter:

Tag der Promotion:

Contents

List of Figures

List of Tables

Chapter 1

Opinion dynamics under opposition

Abstract

We study a DeGroot-like opinion dynamics model in which agents may oppose other agents. As an underlying motivation, in our setup, agents want to adjust their opinions to match those of the agents they follow (their 'ingroup' or those they trust) and, in addition, they want to adjust their opinions to match the 'inverse' of those of the agents they oppose (their 'outgroup' or those they distrust). Our paradigm can account for a variety of phenomena such as consensus, neutrality, disagreement, and (functional) polarization, depending upon network (multigraph) structures and specifications of deviation functions, as we demonstrate, both analytically and by means of simple simulations. Psychologically and socio-economically, we interpret opposition as arising either from rebels; countercultures; rejection of the norms and values of disliked others, as 'negative referents'; or, simply, distrust.

1.1 Introduction

On many issues of everyday life, such as economic, political, social, or religious agendas, disagreement among individuals is pervasive: whether or not Iraq had weapons of mass destructions,[1] the scientific standing of evolution, whether taxes/social subsidies/unemployment benefits/(lower bounds on) wages should be increased or decreased, the right course of government in general, the effectiveness of alternative (or 'standard') medicine such as homeopathy, the effectiveness and appropriateness of death penalty, etc., are all highly debated despite the fact that plenty of data bearing on these issues is available.[2] In fact, in certain contexts such as the political arena, disagreement is 'built in' into and part of the system of opinion exchange. Yet, it has been observed that, contradicting this factual basis, the phenomenon of disagreement is not among the predictions of, in the social and economic context, renown and widely used theoretical models of opinion dynamics, whether they are based on fully rational, *Bayesian*, agents or boundedly rational or *non-Bayesian* actors (see, e.g., the discussions in Acemoglu and Ozdaglar, 2011; Acemoglu, Como, et al., 2012; Yildiz et al., 2012)). Namely, in these models, a standard prediction is that agents tend toward a *consensus opinion*, that is, that all agents eventually hold the same opinion (or belief)[3] about any specific issue. Typically, this applies to both Bayesian frameworks — which is the reason why Acemoglu and Ozdaglar (2011) call them "[no] natural framework[s] for understanding persisent disagreement" (p. 6) — and non-Bayesian setups such as the famous DeGroot model of opinion dynamics (DeGroot, 1974), where consensus obtains as long as the social network wherein agents

[1] See the polling data in "Iraq: The Separate Realities..." (World Public Opinion, 2006).

[2] Some of our examples are taken from Golub and Jackson (2012) and Acemoglu and Ozdaglar (2011). We also note, however, that the 'scope' of disagreement in society is disputed in the relevant literature, see, e.g., Baldassari and Bearman (2007).

[3] Typically, in the literature, the term *belief* is used when there exists a *truth* for an agenda, and the term *opinion* is used when truth is not explicitly modeled, although this may vary from author to author and discipline to discipline. In this work, where we only consider the latter situation, we typically say that agents hold *opinions* on issues, but we take the freedom to occasionally use both terms interchangeably.

communicate with each other is strongly connected (and aperiodic).[4]

Concerning the non-Bayesian DeGroot model, as we consider in this work, a few amendments have more recently been suggested which are capable of producing disagreement among agents. In one strand of literature, models including a *homophily* mechanism, whereby agents limit their communication to individuals whose opinions are not too different from their own, can reproduce patterns of opinion diversity and disagreement (Hegselmann and Krause, 2002; Hegselmann and Krause, 2005; Hegselmann and Krause, 2006; Deffuant et al., 2000).[5] In another strand, Daron Acemoglu and colleagues (cf. Acemoglu and Ozdaglar, 2011; Yildiz et al., 2012) introduce two types of agents, *regular* and *stubborn*, whereby the latter never update their opinions but 'stubbornly' retain their old beliefs, which may be considered an autarky condition; multiple stubborn agents with distinct opinions on a certain agenda may then draw society toward distinct opinion clusters. Such stubborn agents, it is argued, may appear in the form of opinion leaders, (propaganda) media, or political parties that wish to influence others without receiving any feedback from them. Ultimately, the assumption of stubbornness appears problematic, however, since complete autarky in reality probably very rarely obtains (cf. the famous 'no man is an island' condition according to which agents are generally interconnected, even in fragmented societies, cf. Acemoglu and Ozdaglar, 2011).[6]

In this work, we investigate an alternative explanation of disagreement. We consider a non-Bayesian DeGroot-like opinion dynamics model where agents are related with each other via *two types of links*: one link type represents the usual 'weight' that one agent places upon another in DeGroot learning models — these weights, in DeGroot models, typically represent 'trust' between agents, importance, or simply a 'listening/connectedness structure' and are given by real numbers, and, in our model, have the interpretation of strength or intensity of relationship between two agents — and the other link type represents whether or not agents *oppose* each other, whereby opposition is given as a functional relationship ('endomorphism', a mapping from the set of possible opinions to itself) on opinions. In short, in our model, one link type represents *kind of relationship* between agents (opposition or not) and the other represents *intensity of relationship*. The non-opposition case, which we also refer to as *following* ('one agent *follows* another agent's opinion'), is the simple situation where an agent maps another agent's opinion to itself via the identity function and corresponds to the standard operation — although not usually explicated — in DeGroot learning models. The opposition case, which we also refer to as *deviation* or *deviating* ('one agent *deviates from* another agent's opinion') is our model's novel ingredient: in its most abstract form, it simply means that an agent *inverts* another agent's opinions via an endomorphism that is *not the identity function*. Then, after inverting or not, agents take a *weighted arithmetic average*, as in standard DeGroot learning models, of all other agents' possibly inverted opinion signals. This process of inverting or not and subsequent averaging is repeated *ad infinitum* and one of the questions we ask is about the limiting results of the mechanism: e.g., in the limit, will agents tend toward a consensus or will they disagree?

Our model is probably most easily understood in the setup of a 'binary voter' model where only two possible opinions are available (candidate A or B; policy A or B; etc.). Here, the opposition case necessarily means that, if agent i opposes agent j, i will invert j's opinion to B, provided that j holds opinion A, and to A otherwise. Agent i does so for all of his neighbors, leaving the opinions of agents he follows unchanged, and then averages these (possibly inverted) opinions in order to form his next period opinion; of course, in the discrete case, averaging by arithmetic means may not be well-defined and here, we would, e.g., instead consider the operation of i adopting the (weighted) majority opinion of his neighbors' possibly inverted opinion signals. As indicated, we thus allow agents to have both individual

[4] For a recent discussion of the 'problem of consensus', see, e.g., Acemoglu and Ozdaglar (2011); for an early discussion of the problem, see, e.g., Abelson (1964).

[5] However, much depends on the precise modeling of homophily. If homophily means that agents with distinct opinions *never* talk to each other, then disagreement is a likely outcome. However, if homophily is modeled in such a way that agents with distinct beliefs only place low(er) trust weights upon each other, then, again, agreement is a standard prediction, see, e.g., Pan (2010).

[6] As still another explanation of disagreement in DeGroot learning models, it might be argued that even the standard model predicts consensus only as a *limiting* result and that, for all finite intermediate communication stages, disagreement is in fact in accordance with the model. Golub and Jackson (2012) seem to adhere to this interpretation. Problematic about this is that the standard models typically not only imply (full) agreement in the long run but also 'ϵ-agreement' within short periods of time.

neighborhoods (whom they are connected with at all) and individual opposition behavior (whom they follow/deviate from), while, as a first approximation and for simplicity reasons, we do not allow agents to have individual deviation functions, that is, the choice of deviation function is fixed within a population of agents.

Opposition behavior, or deviating, as we have sketched, may be a plausible behavioral assumption from a variety of viewpoints. Firstly, as discussed, in politics, for example, opposition toward members of other parties, most typically the governing party in charge, is so common that opposition may even be considered 'blind' (Jones, 1995; Cohen, 2003), negating whatever opinions competitors hold. Secondly, deviating from an opinion signal may also be plausible when an agent is (suspected of) lying; see, for instance, in the economics context, the abundance of experimental evidence from cheap-talk games (Gneezy, 2005; Rode, 2010; Sutter, 2009). In the following, we discuss four more possible justifications of opposition (that are related both to each other and the justifications brought forth thus far), one based on the concept of *rebels* who derive utility from making different choices than (certain) other agents; one based on the concept of *countercultures* like hippies, punks, etc., that inherently tend to counteract mainstream beliefs, actions, and opinions (in political terms, countercultures may be thought of as playing the opposition parties' roles); one based on the concept of *rejection* of the norms of disliked interaction partners, as has been outlined, e.g., in psychology and sociology, as an important motivation underlying human behavior; and one based on the concept of *distrust*, whereby opposition is thought of as arising from a distrusting stance toward (certain) others, which may include the supposition that certain others are not truthful.

- It has been argued that some agents, e.g., *rebels*, in contrast to *conformists* (see the models of Cao et al., 2011 and Jackson, 2009, p.271), may derive utility simply from the fact of making *different* (opinion) choices than their neighbors. Cao et al. (2011) argue that an attitude of negation, rebelism, may be merely '(intellectually) fashionable', quoting Krugman on his defense of free trade (Krugman, 1996) as saying that some intellectuals attack the concept in question, free trade, merely for the reason that "in a culture that always prizes the avant-garde, attacking that icon [free trade] is seen as a way to seem daring and unconventional." In Zhang et al. (2013), rebels and conformists are interpreted within a 'fashion' context.

- Opposition of opinions and beliefs of others may also arise in the context of the phenomenon of *countercultures*. In fact, counterculture, as defined by Yinger (1977), refers to a group of individuals who hold "a set of norms and values [...] that sharply contradict the dominant norms and values of the society of which that group is a part" (p.833) and who stand "in sharp opposition to the prevailing culture" (p. 834).[7] Accordingly, members of a counterculture define their norms, values, opinions and beliefs *negatively* (or *invertedly*) with respect to the norms, values, opinions, and beliefs held by the 'mainstream culture', at least with respect to certain agendas. This aspect of functional opposition is also emphasized by Davis (1971) who states that "[...] hippies, too, are an instance *par excellence* of a contraculture whose *raison d'être* [...] lies in its members' almost studied inversion of certain key middle class American values and practices." Essentially, thus, countercultures do not simply ignore the opinions of others, but rely on them, as their contrast. It has also been claimed that countercultures are an invariant force in human history (see the discussion in Yinger, 1977 and others), present in ancient and tribal societies as well as throughout the modern era (including, in more recent times, the hippies, the rock experience or Hare Krshna), with prominent relevance, e.g., in modern arts.[8] Finally, countercultures have been said to be the ultimate drivers, via their dialectic opposition of current beliefs, behind change (see the discussion and references in Yinger (1977)).

- Opposition is also closely related to what has, a.o., been termed *rejection* of beliefs, actions, and values of others. According to this concept, agents change their normative systems to become

[7] Yinger (1977) also gives the terms *reversal*, *inversion*, and *opposition* as being definitoric for countercultures, see also Yinger (1960).

[8] In particular, it is contended that countercultures are particularly prominent under conditions of the modern society — rapid economic growth; rapid importation of new ideas, techniques, and goods; sharp increase in life's possibilities; lower participation in intimate and supporting social circles; a loss of meaning in the deepest symbols and rituals of society; etc.

more dissimilar to interaction partners they dislike (cf. Abelson, 1964; Kitts, 2006; Tsuji, 2002; cf. also Groeber, Lorenz, and Schweitzer, 2013) insofar as disliked others may serve as 'negative referents' who inspire contrary behavior (see the discussion in Kitts, 2006). For example, in the simulational study of Fent, Groeber, and Schweitzer (2007), agents maximize utility functions that include positive terms for their *ingroup* members — that is, agents strive to choose norms or traits similar to those of their ingroup — and negative terms for their *outgroup* members — that is, agents, in addition, strive to choose norms or traits dissimilar to those of their outgroup, which entails both attractive and repulsive forces acting upon agents. We note that ingroup favoritism and outgroup 'discrimination' are important and well-established notions in social psychology (see, for instance, Brewer, 1979; Castano et al., 2002) that have also more recently been included in economists' models (cf., e.g., in an experimental context, Charness, Rigotti, and Rustichini, 2007; Fehrler and Kosfeld, 2013, etc.).

We also note that in social network theory, antagonistic relationships between agents are nothing novel, with early work in this context dating back to the 1940's and 1950's already (see Chapter 5 in Beasley and D. Kleinberg, 2010 and references therein). Applications have ranged from international relations (alliances vs. hostile relations) and trust/distrust much in the same way as we indicate below.[9] Often, the concept of *signed networks* (network links have negative or positive 'signs') has been used to model both positive and negative influences. Novel in our context is the application of these notions to the problem of *opinion dynamics*, but see also our discussion in Section 1.2.

- *Distrust.*[10] Opposition, or deviating, may also be thought of as arising from *distrust* between agents, e.g., in the form of *distrusting belief-integrity* (in our situation, believing that the other person does not tell the truth), *institution-based distrust* (believing that appropriate, e.g. legal, structural conditions that are conducive to situational success are not in place), or, generally, a *disposition to disrust*, also referred to as *distrusting stance* or *suspicion to distrust* (a consistent tendency to not be willing to depend on general others across a broad spectrum of situations and persons).[11] In fact, as shown by Mellinger (1956) (see also the typology of Newcomb, 1953), distrust in communication may lead to *aggression* in a sender-receiver setting, that is, to a maximizing of (presumed) disagreement between sender and receiver, which may entail that, e.g., the receiver deviates from the signal sent by the receiver; see also the recent evidence from cheap-talk games under situations of distrust (cf. Rode, 2010), where it is shown that distrust may lead to a larger deviation rate among receivers.

  As concerns the implications of distrust, distrust in communication may be beneficial, in particular, because distrust may prevent harm from distrustors (e.g., preventing them from making the 'wrong' decision in cheap talk games; for a more general setting, see, e.g., Schul, Mayo, and Burnstein, 2008). However, too much distrust may lead to paranoid cognitions, as McKnight and Chervany (2001) emphasize, where "no matter what the other party says or does, their actions and words are interpreted negatively", so that "a balance of trust and distrust is important" (p.45). We also point out that distrust may be a more severe issue in certain institutional settings[12] than in others; in particular, it has apparently become more prevalent in recent times (Deutsch, 1973; Mitchell, 1996; Rotter, 1971; Aupers, 2012).[13]

The outline of this work is as follows. First, to illustrate key concepts and ideas, we start with a 'discrete majority voting DeGroot model' where, in each period, agents adopt their neighbors' weighted

[9] Recent empirical validity of both positive and negative relationships between individuals in social networks is, amongst others, provided in Leskovec, Huttenlocher, and J. M. Kleinberg (2010).

[10] As one particular example, which can also be subsumed under the notion of countercultures, of 'large-scale' distrust, modern conspiracy culture, which distrusts 'conventional' and 'official' explanations of the order of things (such as the assassination of John F. Kennedy, the 9/11 attacks, etc.), may be cited (cf. Aupers, 2012).

[11] Here, we follow the distrust typology of McKnight and Chervany (2001).

[12] For example, in anarchy, dictatorship, etc.; to make a case, Mishler and Rose (1997) call distrust the "predicted legacy of Communist rule", see also Howard (2002).

[13] Trust, or distrust, is clearly also related to income; see, e.g., Ananyev and Guriev (2013) and references therein, and to personal experiences (Nee, Opper, and Holm, 2013).

majority opinion, where, as throughout our paper, we allow agents to invert the opinions of certain other agents. In this discrete model, the set of possible opinions is finite or even binary ('candidate A or B'), and, to our knowledge, the analysis of the repeated weighted majority voter model alone, even without opposition, is a novel setting.[14] Subsequent to the discrete setup, we consider the continuous model where agents hold opinions that lie in a convex subset of the real line and update opinions by taking weighted arithmetic averages of their peers' opinions. In general, the differences between the discrete and the continuous setups are, firstly, that the discrete model is 'more robust' to changes, both in the opinion vectors and the structure of the social networks; this comes as no surprise since, to sketch an example, if 90 neighbors of an agent i hold opinion A and 10 hold opinion B, then i will favor A over B even when a moderate or large quantity of his neighbors change their mind, while, in the continuous case, arbitrarily small changes in neighbors' opinions may always impact i's opinion. Secondly, from a modeler's perspective, the continuous model is simpler to analyze because of the availability of strong mathematical theorems in this case (e.g., results on limits of iterates of continuous functions, continuous fixed-point theorems, spectra of (linear) operators, etc.). Accordingly, in the discrete model, we will content ourselves with results on opinion profiles agents can, or cannot, converge to (fixed-points of the opinion update operators), while in the continuous model, we in addition study actual dynamics. Concerning positive results, both the discrete and the continuous version of our opposition DeGroot model allow the following outcome scenarios.

- **Consensus**: In the discrete model, opposition may have no impact at all as long as the groups of agents that agents oppose are not 'influential enough'; thus, insofar as the non-opposition model can generate consensus profiles as limits of DeGrootian opinion updating, our opposition model may entail the same patterns, in this situation. A similar outcome can be observed in the continuous model. Here, if the groups of agents that agents follow (agents' 'ingroups') *are* 'influential enough', then agents can reach arbitrary consensus profiles that depend on their initial opinions in the same way as in the standard DeGroot model. In particular, we give sufficient conditions under which agents can (and do) reach such consensus profiles and we show that consensus opinions are, in this situation, given by a weighted linear combination of initial opinions where the weights represent the social influence of the agents (Section 1.6.2).

- **Neutrality**: As an important special case of a consensus, we show that both in the discrete and the continuous model, agents can reach a 'neutral' consensus profile. Neutrality means that the opinions in the consensus profile 'admit no opposite' (of course, this depends on the specification of the deviation endomorphism). We think of such opinions as 'undisputable', 'uncontroversial' or, simply, 'neutral'. We also show that both in the discrete and the continuous model, agents can *only* attain neutral consensus profiles as long as agents' outgroups are, again, 'influential enough' in that the weights that agents assign them (a single one suffices) exceed a certain threshold. Moreover, for the continuous model under *affine-linear* deviation functions, we show that our opposition model *typically* leads agents to neutral consensus profiles, as limits of the updating dynamics; we give necessary and sufficient conditions on when this happens.

  To say another word on neutral consensus opinions, we also think of this result as a particular kind of 'withdrawal' of opinion that has empirically, e.g., been observed in situations of distrust in communication (cf. Mellinger, 1956). In fact, if opinions are generally distrusted (opposed/inverted), then it may be safest to utter an opinion neutral enough to admit no opposite (such as 'I don't know', rather than affirmation or negation); at least, it may be an equilibrium in which no one has a unilateral incentive to defect, even though, of course, neutrality may not be desirable from a 'truth perspective', as we discuss in the conclusion.

- **Disagreement**: If opposition is 'hard enough' or if the distribution of deviation endomorphisms satisfies a certain pattern (which we call 'anti-opposition bipartite') agents may disagree forever (cf. Example 1.4.5) and their opinions may even cyclically repeat. Hard opposition may also lead to heavy short-term fluctuations of opinions (cf. Kramer, 1971) as Figure 1.5 illustrates. In the

[14] The binary voter (DeGroot) model considered in Yildiz et al. (2012) is of a much different nature than our approach since it considers agents who randomly adopt one of the neighbors' opinions, rather than by averaging via majority rule.

discrete model, disagreement (or non-consensus) may typically occur both in the non-opposition as well as in the opposition setup, although disagreement likelihood tends to increase with opposition (cf. Figure 1.9).

- **Polarization**: As a special case of disagreement, we show that a certain distribution of deviation endomorphisms (which we call 'opposition bipartite') admits *polarization* as a fixed-point of opinion updating dynamics. By polarization, we mean that agents' opinions cluster in two distinct regimes of the opinion space. For the continuous model and for affine-linear deviation endomorphisms, we derive necessary and sufficient conditions under which opinion dynamics always lead to polarization, no matter the agents' initial opinions. Our models admit, moreover, *functional polarization* in which what the two groups of agents believe are opposites of each other rather than arbitrary, unrelated, disagreeing opinions. Functional polarization would plausibly be the predicted outcome under countercultural opposition, for example, as our above discussion suggests.

As our work's highlight and main theorem, we present, in Theorem 1.6.2, necessary and sufficient conditions on when agents, in our setup, polarize, reach a neutral consenus, and diverge (another special case of disagreement), for arbitrary initial opinions of agents, as limit results of our DeGroot-like 'opposition' opinion dynamics process; the theorem holds for the special case when the deviation endomorphism has a form we call 'soft opposition' (which yields networks, or 'multigraphs', that correspond to the signed networks discussed in the social networks theory literature) and when the network within which agents communicate is symmetric. Our necessary and sufficient conditions are purely in the language of graph theory, which renders them clear and attractive.

The structure of this work is as follows. In Section 1.2, we survey variants of DeGroot learning proposed in recent years. In Section 1.3, we outline our model mathematically. In this context, we also give different economic justifications of our opposition DeGroot learning process and detail possible choices of deviation endomorphisms. Before outlining our main findings and their proofs in Sections 1.5, on the discrete DeGroot model, and 1.6, on the continuous variant, we introduce definitions and further mathematical notation and concepts in Section 1.4. We also give a few introductory examples there. Finally, we conclude in Section 1.7.

1.2 Related Work

Early and frequently cited predecessors of DeGrootian opinion dynamics are French (1956) and Harary (1959), although the now famous 'averaging' model of opinion and consensus formation has only been popularized through the seminal work of DeGroot (1974). At about the same time, Lehrer and Wagner (Wagner, 1978; Lehrer and Wagner, 1981; Lehrer, 1983) have developed a model of rational consensus formation in society that, in both its implications and its mathematical structure, is very similar to the DeGroot model, although behaviorally substantiated in more detail. Friedkin and Johnsen (1990) and Friedkin and Johnsen (1999) develop models of social influence that generalize the DeGroot model. In more recent years, a renewed interest in the DeGroot model of opinion and consensus formation has emerged, leading to a number of extensions proposed. For example, DeMarzo, Vayanos, and Zwiebel (2003), besides sketching psychological justifications of DeGroot learning, discuss time-varying weights on own beliefs that capture, e.g., the idea of a 'hardening of positions': over time, individuals may be more inclined to rely on their own beliefs rather than on those of their peers. Noteworthy are moreover the models of Deffuant et al. (2000) and of Hegselmann and Krause (2002), both of which are very similar in spirit; the two models mainly differ from each other in that, in the former, two randomly determined agents, rather than all agents, update opinions in each time step. The postulate of both models is that agents take only those individuals with 'similar' opinions into account (that is, assign them positive weights), which may be considered a tenet of homophily. In Hegselmann and Krause (2002), this leads to very interesting patterns of opinion formation in which, most prominently, the paradigms of plurality, polarization and consensus are observed, depending on specific parametrizations (most importantly, the definition of similarity, i.e., whether individuals are tolerant or not toward other opinions, affects which opinion pattern emerges). There is much research that directly relates to the Hegselmann and Krause (2002) model, from various disciplines; see, e.g., Hegselmann and Krause (2005), Hegselmann and Krause

(2006), Douven and Riegler (2009a), Douven and Riegler (2009b), Douven and Riegler (2010), Groeber, Lorenz, and Schweitzer (2013), and many others. As we have mentioned, whether homophily leads to disagreement may substantially depend on the specification of homophily. For example, Pan (2010) discusses a homophily variant in which agents assign trust weights to other agents *in proportion* to agents' current opinion distance — rather than by assigning uniform trust weights for agents within a fixed distance to own beliefs and *zero* trust weights to agents outside that radius,[15] as done in the Hegselmann and Krause models and in Deffuant et al. (2000) — which typically entails a consensus, in the limit. Homophily and DeGroot learning is also investigated in Golub and Jackson (2012), where the relationship between the speed of DeGrootian learning and homophily is discussed; in this model, homophily is modeled by designing random networks where the link probability between different groups is non-uniform, and is, in fact, higher between individuals of the same group.[16] Here, only networks that lead to a consensus are analyzed. Further extensions of the classical DeGroot model include Golub and Jackson (2010), whose contribution is to analyze weight structures such that DeGroot learners whose initial beliefs are *stochastically centered around truth* converge to a consensus that is correct, and the works of Daron Acemoglu and colleagues. For example, Acemoglu, Ozdaglar, and ParandehGheibi (2010) distinguish between regular and forceful agents (the latter influence others disproportionately), such as, in an economic interpretation, monopolistic media, and Acemoglu, Como, et al. (2012) distinguish between regular and stubborn agents (the latter never update); in Yildiz et al. (2012), a discrete version of the DeGroot model with stubborn agents is analyzed in which regular agents randomly adopt one of their neighbors' binary opinions. Concerning the 'consensus problem', forceful agents do generally not entail long-term disagreement between agents, and stubborn agents, trivially, entail long-term disagreement only if they are exogenously 'hard-wired' to hold distinct initial opinions.[17]

Another interesting DeGroot variant is discussed in Buechel, Hellmann, and Klößner (2012) and Buechel, Hellmann, and Klößner (2013) where agents' *stated opinions* may differ from their *true* (or private) opinions and where it is assumed that agents generally wish to state an opinion that is close to that of their reference group even if their true opinions may be very different (which is the 'conformity' aspect of their model); a similar approach is given in Buechel, Hellmann, and Pichler (2012), where DeGroot learning is applied to an overlapping generations model in which parents transmit traits to their children. These papers are related to our own work in that, in both cases, agents may deviate from (other) agents' opinion signals. In our work, *receivers* may deviate from the signals sent by senders, and in Buechel, Hellmann, and Klößner (2013) *senders* may deviate from their own true opinions. Moreover, since the latter model also allows *counter-conformity* (and not only conformity), it, too, incorporates an 'opposition modus', as in our model. It does, however, not induce long-term disagreement for strongly connected and closed groups of agents, instead leading them to a consensus or to a divergence of opinions rather than to a stable polarization.[18] A further modeling that comes close to our own approach, and which constitutes a specialization of our setup,[19] is the work of Cao et al. (2011), who study 'rebels' in a DeGroot learning setting. In their case, rebels are agents who hold views that invert *the average opinion of their neighbors*, which is equivalent, from our perspective, to opposing everyone but one's self. In this model, compared with our approach, since rebels have no ingroup other than themselves, long-term polarization does not ensue. Cao et al. (2011) show that their framework generally, except for very special cases, entails a 'doctrine of the mean' in which agents tend toward holding 'mean opinions' (in our terminology, agents hold neutral opinions).[20,21]

[15] This means that there is no communication whatsoever between agents whose opinions are 'too distant'.

[16] A crucial difference between this model and the other homophily variants is that homophily is endogenous in the latter, while it is exogenous in the Golub and Jackson (2012) model.

[17] The concept of 'stubbornness' does also not provide insight into inter-group antagonisms, as we consider.

[18] The 'problem' is that the model admits no ingroup/outgroup structure as in our framework. Agents want to conform/counter-conform to a *single* reference group, without having possible adversary relations to *different* groups.

[19] In terms of modeling, not in terms of results.

[20] This result is due to the fact that their mode of oppposition is always 'soft opposition', as we define below.

[21] There are still other papers, from various different disciplines, that incorporate ideas of adversary relatioships in the opinion formation process. For example, Zhang et al. (2013) interpret 'rebels' in a fashion context. Fent, Groeber, and Schweitzer (2007) study a simulational model incorporating an ingroup/outgroup mechanism. Finally, Fan et al. (2012) discuss opinion dynamics on signed networks in a simulational context, where the signs represent friendly and antagonistic relationships. They quote Mao Zedong on this issue as saying: "We should oppose what enemies support, and support what enemies oppose". See also the work of Altafini (2013) and Shi et al. (2013).

Social learning is also discussed in various other strands of literature, beyond the DeGroot opinion dynamics model, such as in herding models (cf. Banerjee, 1992; Gale and Kariv, 2003; Banerjee and Fudenberg, 2004), where agents usually converge to holding the same belief as to an optimal action. This conclusion generally applies to the observational learning setting (cf. Rosenberg, Solan, and Vieille, 2006; Acemoglu, Dahleh, et al., 2011), where agents are observing choices and/or payoffs of other agents over time and are updating accordingly. See also the references and the discussion in Golub and Jackson (2010). General overviews over social learning, whether Bayesian or non-Bayesian, whether based on communication or observation, are, in the economics context, for example, given in Lobel (2000) and Acemoglu and Ozdaglar (2011).

1.3 Model

1.3.1 The basic setup

For the continous DeGroot model as we discuss, let S be a convex subset of the real numbers, that is, $\sum_j \alpha_j x_j \in S$ for all finite numbers of elements $x_j \in S$ and all weights $\alpha_j \in [0,1]$ such that $\sum_j \alpha_j = 1$. Below, we will usually think of S as the whole of $\mathbb{R}$ or of some (closed and bounded) interval $[\alpha, \beta]$ for $\alpha \leq \beta$. For the discrete 'majority voting' DeGroot model, we let S be any finite set, without further restrictions.

A set $[n] = \{1, 2, \ldots, n\}$ of n agents forms opinions about an agenda X where all opinions on X lie in S. Initially, each agent $i = 1, \ldots, n$ has an exogenously specified initial opinion $b_i(0)$ on X. Then, agents interact — that is, update their opinions — according to a weighted social 'multigraph'. One type of interaction patterns is represented via an $n \times n$ interaction (or 'importance') matrix $\mathbf{W}$, where $W_{ij} > 0$ indicates that i pays attention to j and where the size of W_{ij} indicates the *intensity* of relationship between i and j. We allow matrix $\mathbf{W}$ to be asymmetric, that is, W_{ij} need not necessarily be equal to W_{ji}. Of crucial importance is also a second type of links between agents, namely, link types that indicate whether agents *follow* or *deviate from* each other; the latter represents opposition behavior. Following is encoded by the identity function $F : S \to S$, with $F(x) = x$ for all $x \in S$. Opposition is encoded by a deviation function $D : S \to S$, where we leave the form of D open other than that it not be the identity function. Note that if S is finite with cardinality $|S| = m$, there are $m! - 1 = m \cdot (m-1) \cdots 1 - 1$ possible choices for D. Now, for all agents i and j, i either follows or deviates from j; we summarize these patterns in an $n \times n$ matrix $\mathbf{F}$ with $F_{ij} \in \{F, D\}$. Again, $\mathbf{F}$ need not be symmetric. Also beware the difference between $\mathbf{F}$ and $\mathbf{W}$; the matrix $\mathbf{W}$ is an $n \times n$ matrix of real numbers, $\mathbf{W} \in \mathbb{R}^{n \times n}$, while $\mathbf{F}$ is an $n \times n$ matrix of functions from S to S, that is, $\mathbf{F} \in \{\phi \,|\, \phi : S \to S\}^{n \times n}$. We also call the entries in $\mathbf{F}$ *endomorphisms* because both the domain and the range of the functions are identical. As discussed in Section 1.1, the case $F_{ij} = D$ may result from a variety of circumstances, such as that i is a rebel, that i disagrees with j in the form of, e.g., countercultural opposition, that j belongs to i's outgroup, or simply, that i distrusts j's opinion signal for reasons some of which we have suggested in the named section, but whose source we leave, ultimately, open. We assume moreover that F_{ij}, like W_{ij}, is exogenously given and remains static over time, that is, agents do not change their attitude toward other agents. We also presuppose, as indicated, that agents truthfully report their opinions at each time period t and that all opinion signals are observable by all agents.

To describe opinion dynamics, in the continuous case, agents repeatedly take weighted arithmetic averages of their neighbors' (possibly inverted) opinion signals. Denoting by $b_i(t) \in S$ the opinion at time $t = 0, 1, 2, \ldots$ of agent i on issue X, opinions thus evolve according to

$$b_i(t+1) = \sum_{j=1}^{n} W_{ij} \cdot F_{ij}(b_j(t)), \tag{1.3.1}$$

for all $i = 1, \ldots, n$ and all discrete time periods $t = 0, 1, 2, 3, \ldots$. Rewriting the updating process (1.3.1) in 'matrix notation', we write

$$\mathbf{b}(t+1) = (\mathbf{W} \circ \mathbf{F})(\mathbf{b}(t)), \tag{1.3.2}$$

where we let, *qua definitione*, the 'operator' $\mathbf{W} \circ \mathbf{F}$ act on a vector $\mathbf{b} \in S^n$ in the manner prescribed in (1.3.1), i.e., $\big((\mathbf{W} \circ \mathbf{F})(\mathbf{b})\big)_i \stackrel{\text{def}}{=} \sum_{j=1}^n W_{ij} \cdot F_{ij}(b_j)$. Equation (1.3.2) may again be rewritten as,

$$\mathbf{b}(t) = (\mathbf{W} \circ \mathbf{F})^t(\mathbf{b}(0)), \tag{1.3.3}$$

by which we denote the t-fold application of operator $\mathbf{W} \circ \mathbf{F}$ on $\mathbf{b}(0)$, that is, $f^t(\mathbf{b}) = f(\cdots f(f(\mathbf{b})))$, where $f = \mathbf{W} \circ \mathbf{F}$.

Remark 1.3.1. In case $\mathbf{F}$ is the $n \times n$ matrix of identity functions, updating process (1.3.3) collapses to the standard DeGroot learning model where $(\mathbf{W} \circ \mathbf{F})^t$ is simply the t-th matrix power of matrix $\mathbf{W}$.

Remark 1.3.2. For short, we will usually write $(\mathbf{W} \circ \mathbf{F})\mathbf{b}$ instead of $(\mathbf{W} \circ \mathbf{F})(\mathbf{b})$.

In the discrete case, we consider the following updating process,

$$b_i(t+1) = \arg\max_{s \in S} \sum_{j=1}^n W_{ij} \mathbb{1}\big(F_{ij}(b_j(t)), s\big), \tag{1.3.4}$$

where $\mathbb{1}(r,t) = 1$ if $r = t$ and zero otherwise. In other words, at time $t+1$, agent i adopts the weighted majority opinion among his neighbors' (possibly inverted) opinions at time t. Note that, in Equation (1.3.4), there may be no unique maximum in which case further specification is necessary (see below). For both the discrete and the continuous case, we use the compact notations (1.3.2) and (1.3.3).

As concerns intensity weights W_{ij}, we require weights to be non-negative, $W_{ij} \geq 0$, with $W_{ij} = 0$ indicating that agent i ignores agent j or, simply, that j is not in i's social network (note that in this case, it does not matter whether $F_{ij} = F$ or $F_{ij} = D$). Usually, we also assume that $\mathbf{W}$ is *row-stochastic*, that is, $0 \leq W_{ij} \leq 1$, for all $i, j \in [n]$, and for all $i \in [n]$, $\sum_{j=1}^n W_{ij} = 1$, but, in some contexts, we drop this requirement and, thus, specify weight restrictions as we analyze the models.

We finally note that opinion evolution under process (1.3.2) may be visualized by operations in a *multigraph* as in Figure 1.3 below (Section 1.4), where there are two possible types of links between agent nodes.

1.3.2 Justifications of the DeGroot learning process

Myopic best-response updating

As has been pointed out by Golub and Jackson (2012), the standard DeGroot learning model may have an interpretation as a myopic best-response updating in a pure coordination game (for a more general setup, see Groeber, Lorenz, and Schweitzer, 2013). In our framework, the updating process may be interpreted as resulting from a mix of a coordination game and an anti-coordination game. For example, in the continuous case, if agents $i = 1, \ldots, n$ have utilities on beliefs $\mathbf{b} = (b_1, \ldots, b_n) \in S^n$ as

$$\begin{aligned} u_i(\mathbf{b}) &= -\sum_{j=1}^n W_{ij}\big(b_i - F_{ij}(b_j)\big)^2 \\ &= -\sum_{j:i \text{ follows } j} W_{ij}(b_i - b_j)^2 - \sum_{j:i \text{ opposes } j} W_{ij}(b_i - D(b_j))^2, \end{aligned} \tag{1.3.5}$$

then best-response dynamics — for each agent i, maximizing utility (1.3.5) with respect to b_i — precisely prescribes the updating process (1.3.1) as long as importance weights W_{ij} are such that $\mathbf{W}$ is row-stochastic.[22] One interpretation of the utility functions (1.3.5) is that agent i has disutility from making different opinion choices than neighbors he follows and has disutility from not deviating from, in the manner described by deviation function D, the opinion choices of neighbors he opposes. We note that when $F_{ij} = F$ for all $i, j \in [n]$, then each consensus $(c, \ldots, c)^\intercal \in S^n$ is a Nash equilibrium of the normal form game $([n], S^n, u(\cdot))$, for $u(\cdot) = (u_1(\cdot), \ldots, u_n(\cdot))$, because, in this situation, all agents' utility functions are at a maximum. When $F_{ij} = D$ for some agents $i, j \in [n]$, but deviation function

[22] Moreover, this presupposes that $W_{ii} = 0$.

D has a fixed-point, $D(x_0) = x_0$ for some $x_0 \in S$, then consensus $(x_0, \ldots, x_0)^\intercal$ is a Nash equilibrium of (1.3.5) for the same reason. Below, in Sections 1.5 and 1.6, we show that such equilibria are the only consensus Nash equilibria in this situation and we provide necessary and sufficient conditions when, in the analytically tractable situation where $D(x)$ is affine-linear, opinion updating process (1.3.3) leads agents precisely to such a consensus Nash equilibrium. We note that since, by our discussion, the operator $\mathbf{W} \circ \mathbf{F}$ in opinion updating process (1.3.2) retrieves best responses of agents, under utility functions $u_i(\cdot)$ as in (1.3.5), to an opinion profile $\mathbf{b}(t)$, the fixed points of $\mathbf{W} \circ \mathbf{F}$ — that is, the point $\mathbf{b}$ such that $(\mathbf{W} \circ \mathbf{F})(\mathbf{b}) = \mathbf{b}$ — are, by definition of a Nash equilibrium, the Nash equilibria of the normal form games $([n], S^n, u(\cdot))$, since, for each such a fixed-point, all players in $[n]$ play best responses to the other players' actions (opinions). In the subsequent sections, we pursue the task of finding $(\mathbf{W} \circ \mathbf{F})$'s fixed points in more detail.

In the discrete case, we may think of agents having utility functions

$$\begin{aligned} u_i(\mathbf{b}) &= -\sum_{j=1}^{n} W_{ij}(1 - \mathbb{1}(F_{ij}(b_j), b_i)) \\ &= -\sum_{j:i \text{ follows } j} W_{ij}(1 - \mathbb{1}(b_j, b_i)) - \sum_{j:i \text{ opposes } j} W_{ij}(1 - \mathbb{1}(D(b_j), b_i)) \\ &= -\sum_{j:i \text{ follows } j, b_i \neq b_j} W_{ij} - \sum_{j:i \text{ opposes } j, b_i \neq D(b_j)} W_{ij}, \end{aligned} \tag{1.3.6}$$

where, again, we let $\mathbb{1}(r, t) = 1$ if $r = t$ and zero otherwise. Namely, in case of utility functions of the form (1.3.6), a best reponse of agent i with respect to the opinion vector $\mathbf{b} = (b_1, \ldots, b_n)^\intercal$ is to choose the weighted majority opinion of his neighbors' (possibly inverted) opinion signals.

Boundedly rational Bayesian learning

In another interpretation — which, however, requires that there exist truths μ for topics X, which we do not assume in our modeling — for the continuous model, as outlined by DeMarzo, Vayanos, and Zwiebel (2003) for the situation when F_{ij} is the identity function for all agents $i, j \in [n]$, the updating process (1.3.1) may be rationalized as follows.[23] Agents initially receive noisy signals $b_i(0) = \mu + \epsilon_i$ about issue X, where ϵ_i is a noise term with expectation zero and where μ is the true value of X. Then, agents $i = 1, \ldots, n$ hear the opinions of the agents with whom they are connected, assigning subjective precisions (inverse of variance) π_{ij} to agents j; if i is not connected with j, then agent i assigns precision $\pi_{ij} = 0$. In the case where the signals are normally distributed, Bayesian updating from independent signals at $t = 1$ implies the updating rule (1.3.1) with $W_{ij} = \frac{\pi_{ij}}{\sum_{k=1}^{n} \pi_{ik}}$, since this weight structure yields the minimum variance convex combination of n independent normally distributed random variables, each with mean μ. As agents may not be connected with all other agents, e.g., due to exogenous constraints or costs, they will generally wish to continue to communicate and update based on their neighbors' evolving beliefs, since this allows them to incorporate indirect information.[24] The behavioral aspect of this model concerns updating after time period $t = 1$. A Bayesian agent would adjust the updating procedure to account for the possible duplication of information and for the "cross-contamination" of his neighbors signals. In contrast, continuing to use the updating rule (1.3.1), which treats all information as 'new', can be seen as a boundedly rational heuristic that addresses the complexity involved in fully Bayesian updating and that is in accordance with the psychological condition DeMarzo, Vayanos, and Zwiebel (2003) refer to as 'persuasion bias', the failure to adjust properly for information repetition.

Allowing F_{ij} to take a 'deviation form' may then require an additional behavioral assumption, namely, that the 'true signals' to be considered by agents $i = 1, \ldots, n$ in updating are not $b_j(t)$ but, instead, $F_{ij}(b_j(t))$. In other words, this would mean to assume, from the perspective of agent i, that agent j receives initial signal $b_j(0)$ such that $F_{ij}(b_j(0))$ has the form $\mu + \epsilon_j$ (which might be plausible, e.g., when agent j is a liar or when his signal has been 'corrupted' or 'distorted', by whatever mechanism, as

[23] We follow here the argumentation structure given in Golub and Jackson (2012).

[24] Such indirect information may also be captured even when i is in fact connected with *all* other agents j, due to the different precisions that agents may assign other agents.

perceived by agent i); subsequently, sticking to the same updating rule and weighting structure would correspond, again, to the discussed bounded rationality heuristic.

Aggregation theory

A third motivation for the continuous DeGrootian updating model (1.3.1), initially again for $F_{ij} = F$ for all $i, j \in [n]$, revolves around theoretical results from economic aggregation theory. We briefly sketch the essence of the argument here. Aggregation theory is concerned with the problem of finding a function G that maps 'opinions' of n experts on m topics (so far, we have considered a single topic X) to a 'joint' set of opinions on the m topics. Importantly, the opinions on the m topics must obey a 'funding restriction' such as probabilistic coherence: for example, the m topics might be m states of the world and the opinions might be probability assignments to the m states, one set of assignments for each expert, such that the sum of the probabilities, per expert, over all states, is one (the funding restriction). The purpose of the aggregator G is then, in this case, to assign a probability distribution over m states to each valid $n \times m$ matrix of probability distributions that captures the opinions of the n experts on the m states. Classic theorems from aggregation theory (see, for example, McConway, 1981; Lehrer and Wagner, 1981; Rubinstein and Fishburn, 1986; Dietrich and List, 2008; Herzberg, 2011) then state that if G satisfies two apparently very intuitive and mild criteria, independence and unanimity, G is a *convex combination* of the opinions of the n experts. Unanimity means that if all experts agree on one topic, G must preserve this consensus. Independence (of irrelevant alternatives) means that G_j, the j-th component of G, depends only on the opinions of the n experts on the j-th topic.

Thus, by the classical theorems mentioned, an apparently 'rational' way to aggregate the opinions of the n agents would be by means of weighted averages, as in the updating process (1.3.1). In fact, this (or similar) argumentation has been extensively made use of by Lehrer and Wagner (Wagner, 1978; Lehrer and Wagner, 1981; Lehrer, 1983) in the 1970s and 1980s as a justification for DeGroot-like opinion formation processes.[25] As in the justification based on boundedly rational Bayesian learning, allowing a deviation function may then be just an additional behavioral assumption about which signals (e.g., b_j or $D(b_j)$) are to be aggregated in a rational way.[26]

1.3.3 Deviation functions

As indicated, the case when i follows j is naturally modeled by letting F_{ij} be the identity, $F_{ij}(x) = x$ for all $x \in S$, that is, i precisely follows the signal sent by j. Contrarily, the choice of deviation function D that models opposition has been left unspecified so far. To define a few workable candidates, we first consider the discrete case when S is the finite set $S = \{A_1, \ldots, A_K\}$, for $K \geq 1$. If S is an arbitrary (finite) set, then any choice of $D : S \to S$ (which is not the identity function) seems equally plausible — as mentioned, there are $|S|! - 1$ possibilities to specify D — so we consider the situation when the elements in S have some *meaning*, at least in a relative sense, as when S is *linearly ordered* by some ordering relationship $<$ on S such that, without loss of generality, $A_1 < A_2 < \cdots < A_K$.

Example 1.3.1. Of course, when S is a finite subset of the real numbers (or integers), the usual $<$ relation on the reals (or integers) constitutes a natural linear order. Further interesting examples might arise in the case when S, e.g., consists of (discrete) probabilistic propositions about likelihoods of events such as when $S = \{$"impossible","unlikely","possible","likely","certain"$\}$, which may be thought of as probabilistically ordered, i.e., "impossible" $<$ "unlikely", etc. Other such examples might include:

- $S = \{$"disagree","agree"$\}$, which may be thought of as being ordered by a 'consent' relationship,
- $S = \{$"false","true"$\}$, which may be thought of as being ordered by a 'trueness' relationship,

[25] However, the argumentation appears problematic from at least two perspectives. First, the named theorems hold only for $m \geq 3$. Secondly, why agents should stick, repeatedly, to the same weighted averaging updating rule is unexplained, particularly, since, as outlined before, this implies that they double count information.

[26] This argumentation, if valid, could then also be used to justify why agents should take a weighted *arithmetic* average of the beliefs of all other agents, rather than some different 'mean function' such as the harmonic, geometric, or quadratic mean. In fact, the mathematics literature has provided an infinitude of different mean functions (see, for example, Hardy, Littlewood, and Pólya, 1934), and, for instance, Krause and Hegselmann and Krause (Krause, 2009; Hegselmann and Krause, 2005) discuss DeGroot-like updating processes with a variety of different weighted averages.

- $S = \{$"hate","dislike","neutral","like","love"$\}$, which may be thought of as being ordered by a 'emotional attitude' relationship,
- $S = \{$"strong reject","reject","borderline","accept","strong accept"$\}$, which may be thought of as being ordered by a 'degree of acceptedness' (at journals, conferences, etc.) relationship, etc.

If S is so ordered (by $<$), we consider two natural, as we think, examples of deviation functions. The first one we call 'hard opposition': it is the deviation function that maps opinions to the 'extreme' opinions A_1, the smallest element of S, and A_K, the largest element of S.

Example 1.3.2 (Hard opposition)**.** *Hard opposition* models that an agent i maps another agent j's opinion to one of the two 'extreme' opinions A_1 and A_K, depending on the 'location' of j's opinion. Formally, we assume that there exists $\bar{A} \in S$ such that $D(x) = A_1$ if $\bar{A} < x$, for $x \in S$, and $D(x) = A_K$ if $x < \bar{A}$. When $x = \bar{A}$, we either assume that $D(x) = x$ (D has a fixed point) or, conventionally, $D(x) = A_K$ or $D(x) = A_1$. For our above specified choices of S, this might mean, for instance,

- for $S = \{$"disagree","agree"$\}$: $D($"agree"$) =$ "disagree", $D($"disagree"$) =$ "agree", with $\bar{A} =$ "agree",
- for $S = \{$"hate","dislike","neutral","like","love"$\}$: $D(x) =$ "love" whenever $x =$ "hate","dislike" and $D(x) =$ "hate" whenever $x =$ "like","love". For $\bar{A} =$ "neutral", we might let $D($"neutral"$) =$ "neutral",

and so on.

Our next 'natural' choice of deviation function, we call "soft opposition" (or 'tit for tat' opposition). It models the situation when "more moderate" opinions are mapped to "more moderate" inverted opinions; equivalently, more extreme opinions are mapped to more extreme opinions on the 'other end' of the opinion spectrum.

Example 1.3.3 (Soft opposition)**.** By *soft opposition*, we mean a deviation function where there is a 'moderate' center opinion $\bar{A}$ such that D maps opinions x more to the extremes (on the opposite site of the opinion spectrum, whereby $\bar{A}$ is the focal point) the more distant they are to $\bar{A}$. For instance, we might have $A_1 < A_2 < \cdots < A_k < \bar{A} < A_{k+1} < \cdots < A_K$ with $D(A_i) = A_{K-i+1}$ and $D(A_{K-i+1}) = A_i$, for $i = 1, \ldots, k$, and $D(\bar{A}) = \bar{A}$. If S has even cardinality, we may think of $\bar{A}$, slightly imprecise, as an 'imagined' additional element of S for which D is undefined.

For our above examples, this might mean,

- for $S = \{$"disagree","agree"$\}$: $D($"agree"$) =$ "disagree", $D($"disagree"$) =$ "agree", with $\bar{A}$ an 'imagined' focal point between "disagree" and "agree" (soft opposition and hard opposition may coincide if S has only two elements),
- for $S = \{$"hate","dislike","neutral","like","love"$\}$: $D($"hate"$) =$ "love", $D($"dislike"$) =$ "like", $D($"neutral"$) =$ "neutral", $D($"like"$) =$ "dislike", $D($"love"$) =$ "hate", with $\bar{A} =$ "neutral".

In Figure 1.1, we schematically sketch soft and hard opposition in the discrete case. In the continuous case, when S is a convex subset of the real line that is in addition closed and bounded — that is, S is an interval $[\alpha, \beta]$, with $\alpha, \beta \in \mathbb{R}$ — hard opposition naturally corresponds to solving the maximization problem, for all x different from $\frac{\alpha+\beta}{2}$,

$$D(x) = \text{argmax}_{b \in [\alpha,\beta]} |b - x|, \tag{1.3.7}$$

which has a unique solution for all such x, namely, α and β, respectively. Thus, in the continuous case, hard opposition can be thought of as arising from the principle to maximize disagreement, in an *absolute distance sense*, with an agent that is opposed. Slightly problematic, from an analytical perspective, would be here that D has, in the situation of hard opposition, a discontinuity at $\frac{\alpha+\beta}{2}$, no matter the definition of D for this point, which potentially makes it less attractive as a modeling choice.

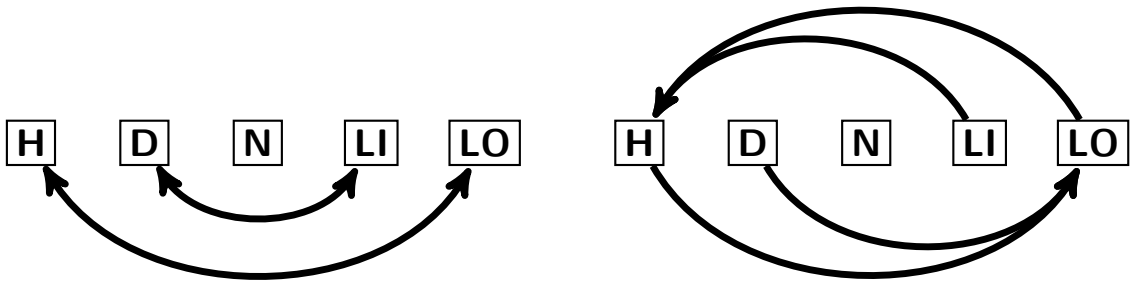


Figure 1.1: Schematic illustration of soft (left) and hard (right) opposition on the set $S = \{\text{“hate”}, \text{“dislike”}, \text{“neutral”}, \text{“like”}, \text{“love”}\}$. Elements $x \in S$ are abbreviated with initial letters.

Soft opposition would in the continuous case of $S = [\alpha, \beta]$ naturally correspond to the operation,

$$D(x) = \alpha + \beta - x. \tag{1.3.8}$$

which defines a continuous function and has a fixed point at $\frac{\alpha+\beta}{2}$. Moreover, in the case of $S = [0, 1]$, when S is the unit interval, soft opposition may be stochastically interpreted as probabilistic inversion, which can be thought of as disagreeing, with an agent j, on all truth conditions for issue X, so that it is apparently also an instance of the tenet to maximize disagreement. In the case of $S = [-\beta, \beta]$, soft opposition has the simple functional form $D(x) = -x$, which makes it an apparently very convenient and tractable candidate of a deviation function.

We graphically illustrate deviation choices (1.3.7), together with some variations, and (1.3.8) in Figure 1.2, for $S = [-1, 1]$.

1.4 Definitions, preliminaries and notation

We now define a couple of important concepts to be used in the remainder of this work. We start with definitions relating to deviation functions and to the operator $\mathbf{W} \circ \mathbf{F}$. Throughout, we let S be an arbitrary non-empty set, the *opinion spectrum* or *opinion space*.

Definition 1.4.1. Let Y be an arbitrary set and let Q be an arbitrary function $Q : Y \to Y$. By $\mathrm{Fix}(Q)$, we denote the set of fixed points of Q, that is, the set of all $x \in Y$ such that $Q(x) = x$.

Note that, in this work, we only consider $Y = S$ and $Y = S^n$. Our next definition simply restates what a deviation function is.

Definition 1.4.2. We call a function $D : S \to S$ *deviation function* (or *opposition function*) if $\mathrm{Fix}(D) \subsetneq S$, that is, if there exist elements $x \in S$ such that $D(x) \neq x$.

The points which D fixes, we call ‘neutral opinions’. In an economic interpretation, neutral opinions may be thought of as opinions that ‘allow no opposite’ or that are ‘undisputable’. For instance, if S were the set $\{\text{“Yes”}, \text{“Nay”}, \text{“Undecided”}\}$, then probably “Undecided” were a good candidate of a neutral opinion. If D admits no neutral opinions, we call D ‘radical’.

Definition 1.4.3. We call an opinion $x \in S$ for which $D(x) = x$ *neutral*.

Definition 1.4.4. If $\mathrm{Fix}(D) = \emptyset$, we call D *radical*.

If two opinions are ‘opposites’ of each other, from the perspective of deviation function D, we call them ‘opposing viewpoints’.

Definition 1.4.5. We call $a, b \in S$ *opposing viewpoints* if $D(a) = b$ and $D(b) = a$.

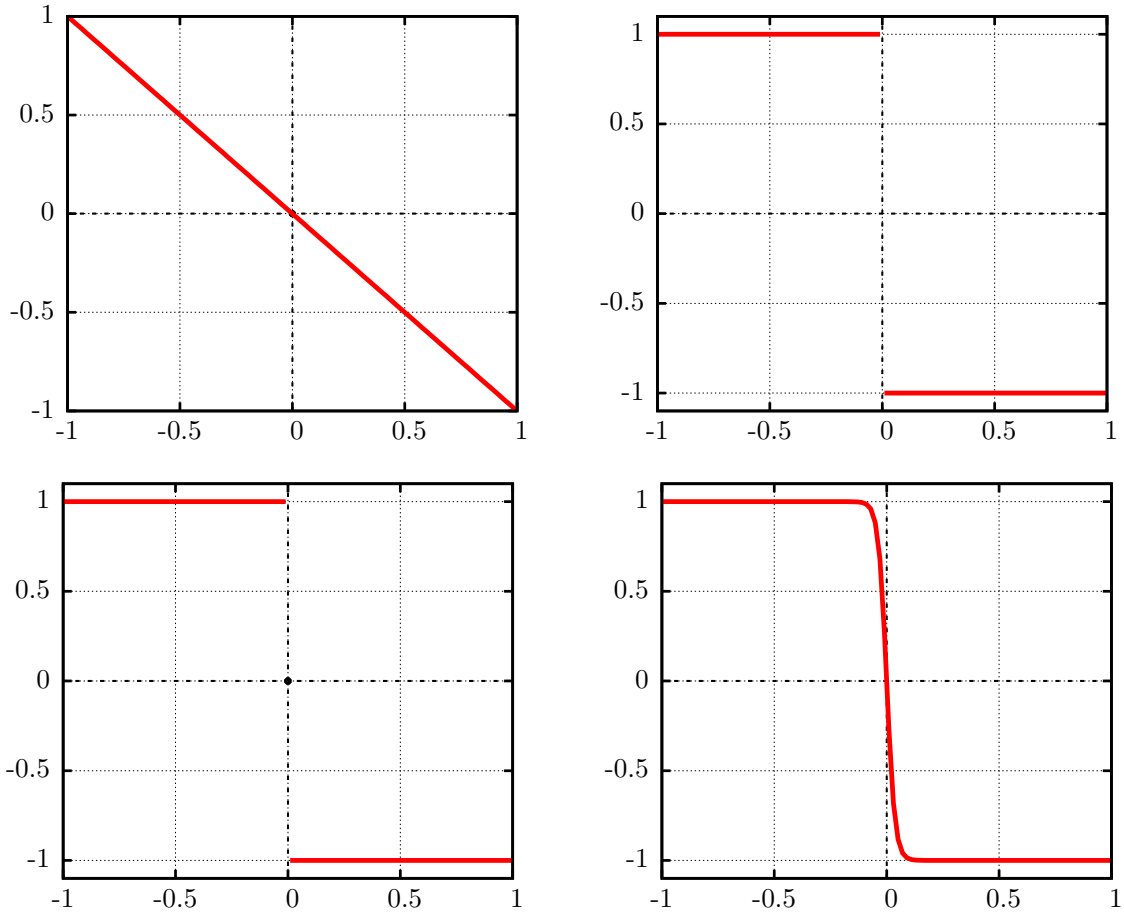


Figure 1.2: Deviation functions on $S = [\alpha, \beta] = [-1, 1]$. Top left: Soft opposition $D(x) = \alpha + \beta - x = -x$ on S. Top right: Hard oppositon on S. Bottom left: Hard opposition with fixed-point $D(0) = 0$. Bottom right: A continuous extension of bottom left.

Next, we define mathematical convergence of opinion updating process (1.3.3). Our definition applies both to the discrete and the continuous setup. Semantically, convergence means that each agent tends toward a 'limiting opinion' — rather than, e.g., changing his mind indefinitely — under repeated opinion updates as described by the updating processes in Section 1.3.

Definition 1.4.6. We say that $\mathbf{W} \circ \mathbf{F}$ is *convergent for opinion vector* $\mathbf{b}(0) \in S^n$ if $\lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$ exists. Moreover, we say that $\mathbf{W} \circ \mathbf{F}$ *induces a consensus for opinion vector* $\mathbf{b}(0)$ if $\mathbf{W} \circ \mathbf{F}$ is convergent for $\mathbf{b}(0)$ and $\lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$ is a *consensus*, that is, a vector $\mathbf{c} \in S^n$ with all entries identical.

Definition 1.4.7. We say that $\mathbf{W} \circ \mathbf{F}$ is *convergent* if $\mathbf{W} \circ \mathbf{F}$ is convergent for *all* initial opinion vectors $\mathbf{b}(0) \in S^n$. Moreover, we say that $\mathbf{W} \circ \mathbf{F}$ *induces a consensus* if $\mathbf{W} \circ \mathbf{F}$ induces a consensus for *all* initial opinion vectors $\mathbf{b}(0) \in S^n$.

Remark 1.4.1. In the discrete case, when S is a finite set and operation $(\mathbf{W} \circ \mathbf{F})\mathbf{b}$ refers to a majority update operation, convergence of $\mathbf{W} \circ \mathbf{F}$ — if indeed it obtains — obtains after a finite amount of time. This is because the sequence $\big(\mathbf{b}(t)\big)_{t\in\mathbb{N}}$, with $\mathbf{b}(t) = (\mathbf{W} \circ \mathbf{F})^t \mathbf{b} \in S^n$, cannot consist of an infinite number of *different* opinion vectors in this case and must, in fact, repeat after time $|S|^n$ at the latest. In other words, if S is finite, $\mathbf{b}(t_0) = \mathbf{b}(s_0)$, for some distinct time points t_0 and s_0. Let $\bar{s}$ be the smallest time point such that $\mathbf{b}(\bar{s}) = \mathbf{b}(\bar{t})$, for some $\bar{t} > \bar{s}$. Then, obviously, if and only if $\bar{t} = \bar{s} + 1$, $\mathbf{W} \circ \mathbf{F}$ is convergent (for $\mathbf{b}(0)$); otherwise, $(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$ 'cycles'.

Remark 1.4.2. We also note that, if $\mathbf{W} \circ \mathbf{F}$ is convergent, then $\mathbf{b}(\infty) := \lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t\mathbf{b}(0)$ is a fixed-point of $\mathbf{W} \circ \mathbf{F}$ as long as $\mathbf{W} \circ \mathbf{F}$ is a continuous operator:

$$(\mathbf{W} \circ \mathbf{F})\mathbf{b}(\infty) = (\mathbf{W} \circ \mathbf{F}) \lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t\mathbf{b}(0) = \lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^{t+1}\mathbf{b}(0) = \mathbf{b}(\infty).$$

Continuity, in turn, depends on the matrix $\mathbf{F}$ and, in particular, on $D(x)$ (since, certainly, $F(x) = x$ is a continuous function). If S is finite and $\mathbf{W} \circ \mathbf{F}$ is convergent (for $\mathbf{b}(0)$), then $\mathbf{b}(\infty)$ is a fixed-point of $\mathbf{W} \circ \mathbf{F}$ no matter the specification of $\mathbf{F}$, by Remark 1.4.1. Fixed-points of $\mathbf{W} \circ \mathbf{F}$ are interesting, for instance, because they constitute Nash equilibria of the coordination games given in Section 1.3 as justifications of our DeGroot learning model. Hence, if $\mathbf{W} \circ \mathbf{F}$ is continuous or if S is discrete, then if $(\mathbf{W} \circ \mathbf{F})^t(\mathbf{b}(0))$ converges, it converges to a Nash equilibrium of the coordination games in question.

As a short-hand for subsequent sections, we introduce the following convenient notations, before proceeding to more conceptual definitions again.

Definition 1.4.8. Let $A \subseteq [n]$ be a subset of the set of agents and let $i \in [n]$ be a particular agent. We denote by $W_{i,A} := \sum_{j\in A} W_{ij}$ the total weight mass i assigns to group A.

Definition 1.4.9. For $i = 1, \ldots, n$, we denote by $\mathcal{O}_i$ the set of agents that agent i opposes. Formally, $\mathcal{O}_i = \{j \in [n] \,|\, F_{ij} = D\}$. We also call $\mathcal{O}_i$ "i's opposed set/group of agents" or "i's *outgroup*". By $\mathcal{F}_i$, we denote the set of agents that i follows, $\mathcal{F}_i = \{j \in [n] \,|\, F_{ij} = F\}$. We also call $\mathcal{F}_i$ "i's *ingroup*".

Clearly, it holds that $\mathcal{O}_i \cap \mathcal{F}_i = \emptyset$ and $\mathcal{O}_i \cup \mathcal{F}_i = [n]$ for all $i \in [n]$.
Next, we formally introduce *networks* because of their relationship to our 'matrix' operators $\mathbf{W} \circ \mathbf{F}$.

Definition 1.4.10 ((Weighted) Network)**.** A *network*, or *graph*, is a tuple $G = (V, E)$ where V is a finite set and $E \subseteq V \times V = \{(u, v) \,|\, u \in V, v \in V\}$. We call V the *vertices* or *nodes* of graph G and E the *edges* or *links* of G. Moreover, we call the network G *weighted* if there exist *weights* w_{uv} for each edge $(u, v) \in E$.[27]

We note that the edges of a network G represent a relationship between nodes (agents, in our case), namely, whether or not two nodes are connected; weights generalize this binary relationship. In a *multigraph*, instead of having only one link type between nodes, there may exist multiple link types.

Definition 1.4.11 ((Weighted) Multigraph)**.** A *multigraph* is a tuple $G = (V, \mathcal{E})$ where V is a finite set and $\mathcal{E} = (E, m)$ is a multiset of ordered pairs of nodes, that is, with each edge $(u, v) \in E$ is associated its cardinality $m\big((u, v)\big) \in \{1, 2, 3, \ldots\}$ (the number of link types between nodes u and v). We call the multigraph G *weighted* if with each of the $m\big((u, v)\big)$ edge types of edge (u, v) is associated a 'weight' w^k_{uv}, for $k = 1, \ldots, m\big((u, v)\big)$.

Now, the operator $\mathbf{W} \circ \mathbf{F}$ of opinion updating process (1.3.2) can be thought of as representing a weighted multigraph $G = (V, \mathcal{E})$, where $V = [n] = \{1, \ldots, n\}$ is the set of agent nodes; $\mathcal{E} = (E, m)$, where E denotes the social neighborhoods of agents (who is connected with whom), $m\big((u, v)\big) = 2$ for all $(u, v) \in E$ and $w^1_{uv} = W_{uv}$ and $w^2_{uv} = F_{uv}$. We let $(u, v) \in E$ if and only if $W_{uv} > 0$. For an illustration, see Figure 1.3, where $\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and $\mathbf{F} = \begin{pmatrix} F & F & F & F \\ F & F & F & F \\ F & F & F & D \\ F & F & F & F \end{pmatrix}$.

In the continuous DeGroot model, that is, when $\mathbf{F}$ is the $n \times n$ matrix of identity functions, as is well known, the concepts of graphs are useful when discussing the convergence of operators $\mathbf{W} \circ \mathbf{F}$. Namely, in this case, updating process (1.3.3) corresponds to a power updating process with a *nonnegative* matrix $\mathbf{W} \circ \mathbf{F} = \mathbf{W}$ (each entry is nonnegative). This situation has been extensively analyzed by the German mathematicians Oskar Perron (1880-1975) and Georg Frobenius (1849-1917) around the turn of the 19th century, and also later in the field of *Markov chain theory*. Although the main results are well-known

[27] Weights may typically be real numbers but, initially, we more generally allow them to be arbitrary mathematical objects.

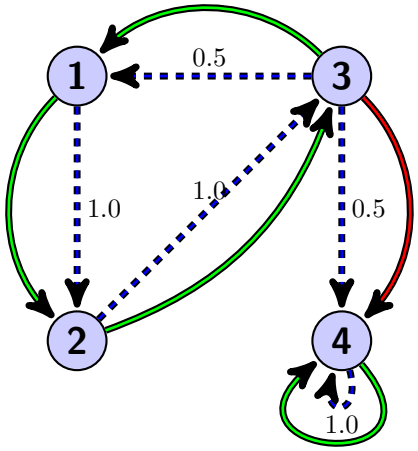


Figure 1.3: Multigraph as a representation of an operator $\mathbf{W} \circ \mathbf{F}$. Dashed blue links denote weights W_{uv}; green and red links denote following and deviating, respectively.

and have, e.g., been summarized both in the mathematics literature as well as in economics contexts (prominently, e.g., in Golub and Jackson, 2010), we very briefly indicate some of them here as well, both in order to introduce useful terminology and to sketch some basic insights about results on the standard DeGroot model. To restate the results, we first need to define a few properties of networks, which we cite from Golub and Jackson (2010).

Definition 1.4.12. A *walk* in a network $G = (V, E)$ is a sequence of nodes $i_1, i_2, \ldots, i_K$, not necessarily distinct, such that $(i_k, i_{k+1}) \in E$ for all $k \in \{1, \ldots, K-1\}$.

A *path* is a walk consisting of distinct nodes.

A *cycle* is a walk $i_1, \ldots, i_K$ such that $i_1 = i_K$. The *length* of cylce $i_1, \ldots, i_K$ is defined to be $K - 1$. A cycle is called *simple* if the only node appearing twice is $i_1 = i_K$.

Definition 1.4.13. The graph $G = (V, E)$ is said to be *strongly connected* if there is a path in G from any node to any other node.

Definition 1.4.14. The graph $G = (V, E)$ is said to be *aperiodic* if the greatest common divisor of the lengths of its simple cycles is 1. We call G *periodic* if it is not aperiodic.

Remark 1.4.3. We remark that we generally use the same terminology — 'strongly connected', 'aperiodic', etc. — whether we speak of multigraphs or ordinary graphs. In the case of multigraphs, when using the mentioned terminology, we refer to the ordinary graphs $G = (V, E)$ underlying the multigraphs $G = (V, \mathcal{E} = (E, m))$. Moreover, since we treat operators $\mathbf{W} \circ \mathbf{F}$ and the corresponding multigraphs as 'isomorphic', or, simply, identical, we may also speak of $\mathbf{W} \circ \mathbf{F}$ as 'strongly connected', etc.

Now, we are ready to state one of the main theorems for the DeGroot updates (1.3.3) in the non-opposition case. We assume that $\mathbf{W}$ is row-stochastic.

Theorem 1.4.1. Consider the opinion updating process (1.3.3) with $F_{ij} = F$ for all $i, j \in [n]$, where F is the identity function. Let the multigraph corresponding to the operator $\mathbf{W} \circ \mathbf{F} = \mathbf{W}$ — an ordinary graph — be strongly connected and aperiodic. Then $\mathbf{W} \circ \mathbf{F}$ is convergent and induces a consensus.

Our first example is a negative illustration of Theorem 1.4.1, i.e., it illustrates that if the assumptions of the theorem are not satisfied, then its consequences need not be satisfied as well. It is the classic example of a periodic network where agents' opinions oscillate.

Example 1.4.1. Let $n = 2$ and let $\mathbf{W}$ and $\mathbf{F}$ be the following matrices,

$$\mathbf{W} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & F \\ F & F \end{pmatrix},$$

where F is the identity function. The directed graph corresponding to $\mathbf{W} \circ \mathbf{F}$ is shown in Figure 1.4. Obviously, this graph is periodic since all cycles have length 2. Moreover, with notation as in Equations (1.3.2) and (1.3.3), we have $\mathbf{W} \circ \mathbf{F} = \mathbf{W}$ and

$$\mathbf{W}^t = \begin{cases} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \text{if } t \text{ is odd,} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \text{if } t \text{ is even.} \end{cases}$$

Hence, matrix $\mathbf{W}$ does not converge.

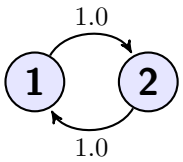


Figure 1.4: The graph corresponding to Example 1.4.1. Since $F_{ij} = F$ for all $i, j \in [n]$, we draw the graph as an ordinary graph, rather than as a multigraph.

More intricate necessary and sufficient conditions for convergence and consensus (than given in Theorem 1.4.1) in the non-opposition setup are, for instance, presented in Golub and Jackson (2010), and references therein. Hence, as far as strong results for the non-opposition case have already been established, in the current work, we generally analyze the situation when $\mathbf{W} \circ \mathbf{F}$ is a 'proper' multigraph, that is, where $F_{ij} = D$ for some agents i and j, so that some agents oppose some others.

Example 1.4.2. To see, however, first, that Theorem 1.4.1 may be false in the discrete weighted majority voter model, consider $S = \{A, B\}$ and $n = 3$. Let agents adopt the majority opinion among their neighbors and, in case of a tie, adopt opinion, say, B. Let $\mathbf{W}$ and $\mathbf{F}$ be the matrices

$$\mathbf{W} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & F & F \\ F & F & F \\ F & F & F \end{pmatrix}.$$

In other words, everyone follows everyone else; agent 1's neighborhood consists of agents 1 and 3, each of whom he weighs equally; and so on. Clearly, $\mathbf{W} \circ \mathbf{F} = \mathbf{W}$ and the graph corresponding to matrix $\mathbf{W}$ is strongly connected and aperiodic so that the assumptions of Theorem 1.4.1 are satisfied. If agents start with initial opinions, say, $(A, A, B)^\intercal$, then the sequence of opinion vectors generated by updating process (1.3.2) reads as:

$$\begin{pmatrix} A \\ A \\ B \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} B \\ B \\ A \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} B \\ B \\ B \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} B \\ B \\ B \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \cdots$$

which is in accordance with Theorem 1.4.1. If, in contrast, agents start with initial opinions, say, $(A, B, A)^\intercal$, then the sequence of opinion vectors generated by updating process (1.3.2) reads as:

$$\begin{pmatrix} A \\ B \\ A \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} A \\ B \\ A \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \cdots$$

which, consequently, does not lead to a consensus.

In the next sections, we discuss the discrete model in more depth. Now, we briefly sketch some more aspects, including examples, relating to the continuous DeGroot model where agents update by taking

weighted arithmetic averages of other agents' previous opinion signals. In this context, we first note that important fixed-point theorems from mathematics and economics allow us to make statements with regard to the behavior of opinion updating process (1.3.3) in the continuous case. These results may be applied in case operator $\mathbf{W} \circ \mathbf{F}$ satisfies certain conditions, as we outline.

Definition 1.4.15. Let $(Y, \|\cdot\|)$ be a metric space. A function $f : Y \to Y$ is called a *contraction mapping* on Y if there exists $\gamma \in [0, 1)$ such that

$$\|f(x) - f(y)\| \leq \gamma \|x - y\|,$$

for all $x, y \in Y$.

Theorem 1.4.2 (Banach fixed point theorem)**.** Let $(Y, \|\cdot\|)$ be a non-empty complete metric space and $f : Y \to Y$ a contraction mapping on Y. Then f has a unique fixed-point x^* in Y. Furthermore, x^* can be found as the limit of the sequence $\big(x(t)\big)_{t\in\mathbb{N}}$, defined via $x(t) = f^t(x_0)$, for any $x_0 \in Y$.

The beauty of Theorem 1.4.2 in our context is that it tells us that opinion update process $(\mathbf{W}\circ\mathbf{F})^t\mathbf{b}(0)$ converges, when $\mathbf{W} \circ \mathbf{F}$ is a contraction mapping, to the unique fixed point of $\mathbf{W} \circ \mathbf{F}$, that is, to the unique Nash equilibrium of the coordination games outlined in Section 1.3. Note, however, that limiting opinions are in this case independent of initial opinions, as the theorem tells.

Interestingly, in case $\mathbf{W} \circ \mathbf{F}$ is an affine-linear map, whether or not $\mathbf{W} \circ \mathbf{F}$ is a contraction mapping can be fully determined via the well-known notion of eigenvalues, which we introduce now.

Definition 1.4.16. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be an $n \times n$ matrix. An eigenvalue of $\mathbf{A}$ is any value $\lambda \in \mathbb{C}$ such that

$$\mathbf{Ax} = \lambda\mathbf{x}$$

for some non-zero vector $\mathbf{x} \in \mathbb{R}^n$. The set of distinct eigenvalues of matrix $\mathbf{A}$ is called its *spectrum* and denoted by $\sigma(\mathbf{A})$. By $\rho(\mathbf{A})$, we denote the *spectral radius* of $\mathbf{A}$, the largest absolute value of all the eigenvalues of $\mathbf{A}$,

$$\rho(\mathbf{A}) = \max\{\big|\lambda\big| \,|\, \lambda \in \sigma(\mathbf{A})\}.$$

Then, the following holds in case $\mathbf{W} \circ \mathbf{F}$ allows a representation as an affine-linear operator, that is, for all $\mathbf{x} \in S^n$,

$$(\mathbf{W} \circ \mathbf{F})\mathbf{x} = \mathbf{Ax} + \mathbf{d},$$

where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{d}$ is an n-vector.

Theorem 1.4.3. If $\mathbf{W} \circ \mathbf{F}$ is affine-linear of the form $(\mathbf{W} \circ \mathbf{F})\mathbf{x} = \mathbf{Ax} + \mathbf{d}$, then $\mathbf{W} \circ \mathbf{F}$ is a contraction mapping if and only if $\rho(\mathbf{A}) < 1$.

Proof. See, e.g., http://web.mit.edu/dimitrib/www/Appendixes_Abstract_DP.pdf. □

Remark 1.4.4. If $\mathbf{W} \circ \mathbf{F}$ is affine-linear of the form $(\mathbf{W} \circ \mathbf{F})\mathbf{x} = \mathbf{Ax} + \mathbf{d}$, then we call $(\mathbf{A}, \mathbf{d})$ the (*affine-linear*) *representation* of $\mathbf{W} \circ \mathbf{F}$.

To apply Theorem 1.4.3 to our setup, consider the following example.

Example 1.4.3. Let $n = 2$, $S = [0, 1]$ and let

$$\mathbf{W} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & D \\ F & F \end{pmatrix},$$

where D is probabilistic inversion, $D(x) = 1 - x$ for all $x \in S$, and where $a + b = c + d = 1$. We note that, in this situation, $\mathbf{W} \circ \mathbf{F}$ can be written in the following form

$$(\mathbf{W} \circ \mathbf{F})\mathbf{x} = \underbrace{\begin{pmatrix} a & -b \\ c & d \end{pmatrix}}_{=:\mathbf{A}} \mathbf{x} + \begin{pmatrix} b \\ 0 \end{pmatrix},$$

as the reader can easily verify. For instance, for $a = d = \frac{2}{3}$, $b = c = \frac{1}{3}$, the two eigenvalues of matrix $\mathbf{A}$ are $\frac{2}{3} \pm \frac{1}{3}i$, both of which have absolute value $\sqrt{\frac{5}{9}} < 1$. Thus, $\mathbf{W} \circ \mathbf{F}$ is a contraction mapping and opinion updating process (1.3.3) accordingly converges to the unique fixed-point of $\mathbf{W} \circ \mathbf{F}$, by the Banach fixed-point theorem, Theorem 1.4.2. Clearly, $\mathbf{b} = (\frac{1}{2}, \frac{1}{2})^\intercal$ is a fixed-point of $\mathbf{W} \circ \mathbf{F}$ and, by our reasoning, it is thus the unique fixed-point to which opinions converge.

Example 1.4.4. Taking the same example as Example 1.4.3, except for S and the deviation function, which we now specify as $S = [-1, 1]$ and $D(x) = -x$, we find that $\mathbf{b} = (0, 0)^\intercal$ is the unique fixed-point of $\mathbf{W} \circ \mathbf{F}$, in this situation, to which opinions converge. Opinion dynamics $\mathbf{b}(t) = (\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$, for $t = 0, \ldots, 50$, are sketched in Figure 1.5 for two different initial opinions $\mathbf{b}(0)$. We generally find that agent 1's opinions oppose agent 2's opinions in that they tend toward another direction of the opinion space, but that opposition becomes weaker as agent 2's opinions become more 'neutral' (which is the essence of what 'soft opposition' means).

Example 1.4.5. Now, we consider again the same example as Example 1.4.3, except for S and D, which we specify as $S = [-1, 1]$ and D is hard opposition on S; conventionally, we let $D(0) = 1$. Opinion dynamics are sketched in Figure 1.5. Rather than convergence (to consensus) as in Examples 1.4.3 and 1.4.4, we now find fluctuating and periodic opinion dynamics. We also find a phase shift in the opinion trajectories of both agents: whenever agent 2's opinions, which 'mimic' agent 1's opinion values, increase (in the opinion space S), agent 1's opinions decrease, and vice versa.

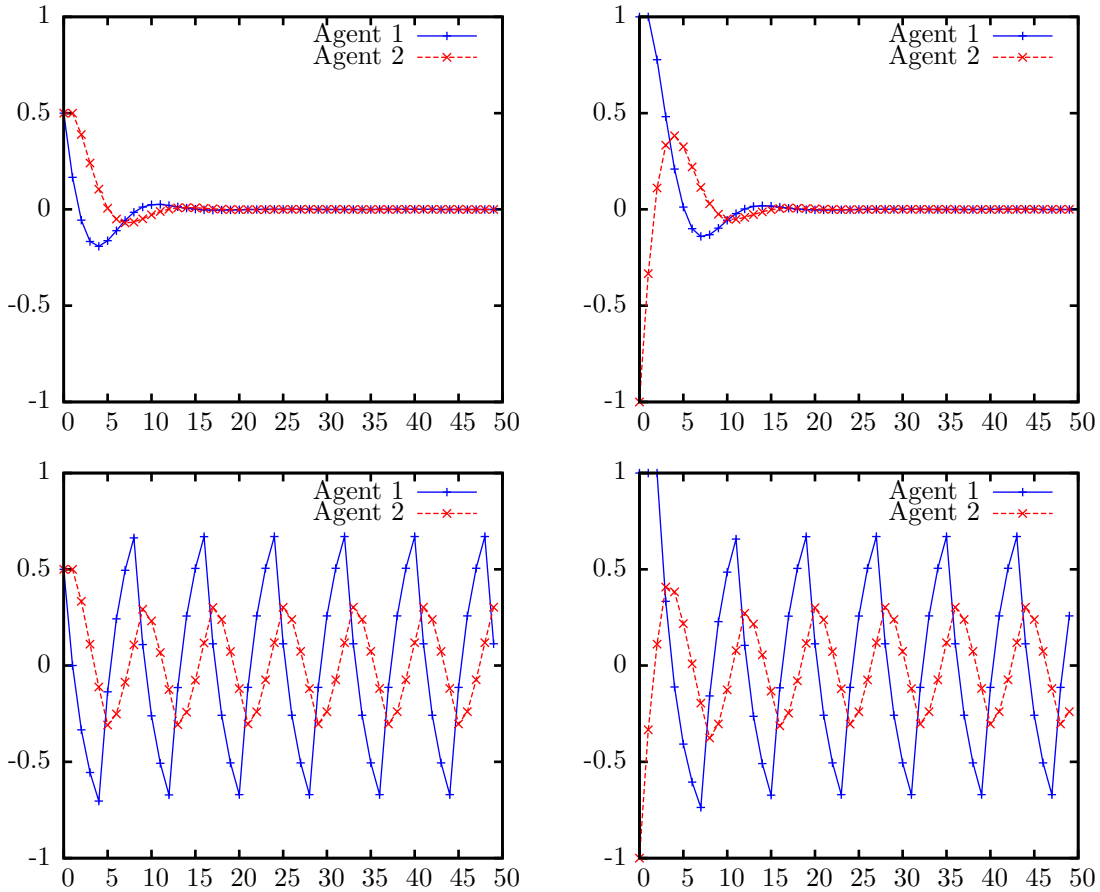


Figure 1.5: Opinion dynamics $\mathbf{b}(t)$ for Examples 1.4.4 and 1.4.5 with $a = d = \frac{2}{3}$, $b = c = \frac{1}{3}$ for $\mathbf{W}$ as in Example 1.4.3. Top: Example 1.4.4 with $\mathbf{b}(0) = (1/2, 1/2)^\intercal$ and $\mathbf{b}(0) = (1, -1)^\intercal$, respectively. Bottom: Example 1.4.5 with $\mathbf{b}(0) = (1/2, 1/2)^\intercal$ and $\mathbf{b}(0) = (1, -1)^\intercal$, respectively.

Example 1.4.6. As our final example, let $\mathbf{W}$ be arbitrary row-stochastic and let $\mathbf{F}$ be such that $F_{ij} = D$ for all $i, j \in [n]$ ('*everyone opposes everyone else*'). Let $D(x) = -x$ be soft opposition on $S = [-\beta, \beta]$. Then $(\mathbf{A}, \mathbf{d}) = (-\mathbf{W}, \mathbf{0})$ such that

$$\mathbf{A}^t = \begin{cases} \mathbf{W} & \text{if } t \text{ is even,} \\ -\mathbf{W} & \text{if } t \text{ is odd.} \end{cases}$$

Consequently, $(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0) = \mathbf{A}^t \mathbf{b}(0)$ oscillates as long as $\mathbf{W}^t \mathbf{b}(0)$ converges to a non-zero limit point, which typically holds, e.g., when $\mathbf{b}(0) \neq \mathbf{0}$ and $\mathbf{W}$ is strongly connected and aperiodic.

The same result holds true when D is hard opposition and agents, e.g., start with a consensus other than a fixed-point of D, no matter the structure of row-stochastic $\mathbf{W}$.

As Theorem 1.4.3 states, if $\mathbf{W} \circ \mathbf{F}$ is affine-linear with representation $(\mathbf{A}, \mathbf{d})$, the spectral radius of matrix $\mathbf{A}$ is of crucial importance for determining whether opinions converge or not, in our setup. When $\mathbf{d} = \mathbf{0}$ ($\mathbf{W} \circ \mathbf{F}$ is linear), a more general result than Theorem 1.4.3 on convergence of the operator $\mathbf{W} \circ \mathbf{F}$, which also includes the situation when $\rho(\mathbf{A}) = 1$, is the following.

Theorem 1.4.4 (Meyer, 2000, p.630)**.** For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\lim_{t \to \infty} \mathbf{A}^t$ exists if and only if

$$\rho(\mathbf{A}) < 1, \quad \text{or else,}$$
$$\rho(\mathbf{A}) = 1 \text{ and } \lambda = 1 \text{ is the only eigenvalue on the unit circle, and } \lambda = 1 \text{ is semisimple,}$$

where an eigenvalue is called *semisimple* if its algebraic multiplicity equals its geometric multiplicity. The *algebraic multiplicity* of an eigenvalue λ is the number of times it is repeated as a root of the characteristic polynomial $\chi(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}_n)$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. The *geometric multiplicity* is the number of linearly independent eigenvectors associated with λ.

The below two results, the latter of which is known as *Schur's inequality*, bound the spectral radius of a matrix $\mathbf{A}$ in terms of matrix p-norms, as we define now, and in terms of its entries.

Definition 1.4.17. The *p-norm*, for $p \in \mathbb{R} \cup \{\infty\}$, $p \geq 1$, of a matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_p = \max_{\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p},$$

where $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for a vector $\mathbf{x} \in \mathbb{R}^n$. As special cases, $\|\mathbf{A}\|_1$ is the maximum absolute column sum of $\mathbf{A}$ and $\|\mathbf{A}\|_\infty$ is the maximum absolute row sum of $\mathbf{A}$.

Theorem 1.4.5. It holds that

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|_p$$

for any $p \geq 1$. Furthermore, it holds that

$$\sum_{i=1}^n |\lambda_i|^2 \leq \sum_{i,j}^n |A_{ij}|^2,$$

where $\lambda_1, \ldots, \lambda_n$ are the (not necessarily distinct) eigenvalues of matrix $\mathbf{A}$.

1.5 The discrete majority voting DeGroot model

In a sense, the discrete majority voting DeGroot process is much harder to analyze than its continuous counterpart since the opinion update operator poses more problems here: in the continuous case, if $\mathbf{F}$ is linear, then $\mathbf{W} \circ \mathbf{F}$ is a linear operator and, in any case, $\mathbf{W} \circ \mathbf{F}$ represents a continuous operator as long as the functions in $\mathbf{F}$ are continuous. Thus, all in all, we content ourselves in the following with deriving results on fixed-points of the operator $\mathbf{W} \circ \mathbf{F}$; as mentioned, these fixed-points constitute Nash equilibria

of the coordination games outlined as justifications of the DeGroot learning process. Throughout, we assume that $\mathbf{W}$ is row-stochastic and that the opinion space $S = \{A_1, A_2, \ldots\}$ contains at least two elements. Moreover, we need the following assumption in order for operator $\mathbf{W} \circ \mathbf{F}$ to be well-defined in the discrete case, namely, the existence of *tie-breaking* elements that discriminate between any choices of opinions.

Assumption 1.5.1 (Tie-breaking element)**.** Let $M \subseteq S$ be an arbitrary non-empty subset of the opinion space. We assume that there exists $b \in M$ such that, in case of a (weighted) tie between the elements of M, agents adopt opinion b as an opinion update rather than any of the other elements in M.

Example 1.5.1. If S is ordered by a ordering relation $<$, a natural notion of a tie-breaking element would be the largest (or smallest) element of any $M \subseteq S$.

Influential groups, decisive groups and persistent disagreement

We start this discussion with three very simple examples, Examples 1.5.2, 1.5.3, and 1.5.4, before considering results of a general nature in Proposition 1.5.1 and thereafter. In Example 1.4.2, we have already seen that — in the situation when $\mathbf{F}$ consists of identity functions exclusively — strong connectedness and aperiodicity of the networks $\mathbf{W}$ are not sufficient conditions for $\mathbf{W}$ to induce a consensus, unlike in the continuous case. We now demonstrate by way of illustration that if $\mathbf{W}$ is periodic, then, like in the continuous case, $\mathbf{W}$ may not converge.

Example 1.5.2. Let $\mathbf{W}$ and $\mathbf{F}$ be the matrices,

$$\mathbf{W} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & F & F \\ F & F & F \\ F & F & F \end{pmatrix}.$$

Network $\mathbf{W}$ is periodic, as can easily be verified, since all simple cycles have length 2. Then:

$$\begin{pmatrix} A \\ B \\ B \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} B \\ A \\ A \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} A \\ B \\ B \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \cdots,$$

where $A, B \in S$. In other words, whenever agent 1, on the one hand, and agents 2 and 3, on the other hand, disagree initially, disagreement will perpetuate forever, under the social network $\mathbf{W} \circ \mathbf{F}$. As in the continuous case, this is due to the cyclical information structure in network $\mathbf{W} \circ \mathbf{F}$ whereby agent 1 derives her information from agents 2 and 3, who, in turn, listen to agent 1.

Now, we consider the opposition case when $F_{ij} = D$ for some agents $i, j \in [n]$ and some deviation function D. Interestingly, we notice that opposition may play a similar role as periodicity in the above example and, thus, may prevent convergence. We discuss this example below, too, when we talk about *anti-opposition bipartite* networks.

Example 1.5.3. Let $\mathbf{W}$ be any strictly positive matrix — that is, each entry is positive — and let $\mathbf{F}$ be the matrix,

$$\mathbf{F} = \begin{pmatrix} D & F & F \\ F & D & D \\ F & D & D \end{pmatrix},$$

where D is not the identity function. Note that matrix $\mathbf{F}$ has a very similar structure as matrix $\mathbf{W}$ in the previous example. In fact, if we replace zero entries in $\mathbf{W}$ from Example 1.5.2 by 'D' and positive entries by 'F', $\mathbf{W}$ is transformed into $\mathbf{F}$. Now, let $A, B \in S$ be opposing viewpoints, that is, $D(A) = B$ and $D(B) = A$. Then, as the reader may verify,

$$\begin{pmatrix} A \\ B \\ B \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} B \\ A \\ A \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} A \\ B \\ B \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \cdots,$$

precisely as in Example 1.5.2. This shows that, under opposition, $\mathbf{W} \circ \mathbf{F}$ may not even converge, even when $\mathbf{W}$ is strongly connected and aperiodic, as long as $\mathbf{F}$ satisfies a certain 'aperiodicity' condition (as well as D).

Next, we consider the example where opposition is 'marginal' in that only a few agents deviate from the opinion signals of a few others.

Example 1.5.4. Let $n = 3$ and let $\mathbf{F}$ be the matrix,

$$\mathbf{F} = \begin{pmatrix} F & D & F \\ F & F & F \\ F & F & F \end{pmatrix}, \tag{1.5.1}$$

where D is an arbitrary deviation function, and let $\mathbf{W}$ be uniform, for example,

$$\mathbf{W} = \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

so that everyone weighs everyone else equally. Then, any consensus opinion profile $(A, A, A)^\intercal$, with $A \in S$, is a fixed-point of $\mathbf{W} \circ \mathbf{F}$, as in the non-opposition case. Hence, in the discrete model, 'a bit of opposition' may not necessarily be an obstacle to consensus formation, something that is *not* true (in the same manner) for the continuous setup, as we discuss below. Still, even in this model, opposition does matter for the given example; e.g., for $D(A) = B$ and $D(B) = A$, we have $(A, A, B)^\intercal \mapsto_{\mathbf{W}\circ\mathbf{F}} (B, A, A)^\intercal \mapsto_{\mathbf{W}\circ\mathbf{F}} (B, A, A)^\intercal$, while in the non-opposition analogue of $\mathbf{W} \circ \mathbf{F}$ in which $F_{12} = F$, we have $(A, A, B)^\intercal \mapsto (A, A, A)^\intercal \mapsto (A, A, A)^\intercal$.

The last example raises two questions. Firstly, may opposition have no impact at all in that results are always the same as in the non-opposition scenario, for specific networks $\mathbf{W} \circ \mathbf{F}$? Secondly, in the case of opposition, what are requirements on the weight structure $\mathbf{W}$ that prevent consensus formation?

The latter question has a simple solution. It requires, for example, the following result which states that the set of consensus fixed of points of $\mathbf{W} \circ \mathbf{F}$ coincides with the set of neutral opinions of D provided that some agent assigns 'too large weight mass' to agents he opposes. We first define the concept of consensus vectors.

Definition 1.5.1. Let $\mathcal{C}$ be the set of consensus opinion vectors in S^n, i.e., $\mathcal{C} = \{(a_1, \ldots, a_n)^\intercal \in S^n \mid a_1 = \ldots = a_n\}$.

Proposition 1.5.1. Let $\mathbf{W} \circ \mathbf{F}$ be an arbitrary operator. Then, for any $c \in S$,

$$c \in \mathrm{Fix}(D) \implies (c, \ldots, c)^\intercal \in \mathrm{Fix}(\mathbf{W} \circ \mathbf{F}).$$

Moreover, if $W_{i,\mathcal{O}_i} > \frac{1}{2}$ for some $i \in [n]$, then, for all $c \in S$,

$$c \notin \mathrm{Fix}(D) \implies (c, \ldots, c)^\intercal \notin \mathrm{Fix}(\mathbf{W} \circ \mathbf{F}).$$

In other words, if $W_{i,\mathcal{O}_i} > \frac{1}{2}$ for some $i \in [n]$, then

$$\mathrm{Fix}(D) = P_1[\mathrm{Fix}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}],$$

where P_1 projects consensus vectors $(c, \ldots, c)^\intercal \in \mathcal{C}$ to $c \in S$.

Proof. Let $\mathbf{c} = (c, \ldots, c)^\intercal$.

If $c = D(c)$, then, clearly, by definition of $\mathbf{W} \circ \mathbf{F}$, $(\mathbf{W} \circ \mathbf{F})\mathbf{c} = \mathbf{c}$.

Conversely, let $i \in [n]$ with $W_{i,\mathcal{O}_i} > \frac{1}{2}$ and let $D(c) \neq c$. Then, for agent i, the weight of opinion $D(c)$ is larger than $1/2$. Thus, her updated opinion will be $D(c)$ rather than c, after applying operator $\mathbf{W} \circ \mathbf{F}$. Thus, $(\mathbf{W} \circ \mathbf{F})\mathbf{c} \neq \mathbf{c}$. □

Remark 1.5.1. For an agent i, we call a group of agents $\mathcal{N}$ that satisfies the requirement $W_{i,\mathcal{N}} > \frac{1}{2}$ as in Proposition 1.5.1 *decisive for agent i* as it may decide i's opinion provided that agents in $\mathcal{N}$ agree.

As a simple Corollary to Proposition 1.5.1, we find that the possible consensus limiting opinions of $\mathbf{W} \circ \mathbf{F}$, denoted by,

$$\text{Lim}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C} = \{\mathbf{b} \in S^n \,|\, \mathbf{b} = \lim_{t\to\infty} (\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0), \text{ for some } \mathbf{b}(0) \in S^n\} \cap \{(a_1, \ldots, a_n)^\intercal \in S^n \,|\, a_1 = \ldots = a_n\},$$

are given by the set of fixed points of D as long as at least one agent has 'too much distrust'. In other words, under opposition, agents can only converge to consensus vectors in which the consensus value is a neutral opinion if some agent's outgroup is decisive for him. If, in addition, D is radical, opinion dynamics (1.3.3), in the discrete weighted majority setup, cannot induce a consensus. Hence, under these conditions, disagreement will be persistent. Formally:

Corollary 1.5.1. Let $\mathbf{W} \circ \mathbf{F}$ be such that for some agent $i \in [n]$ the group of agents he opposes is decisive for him. Then,

$$P_1[\text{Lim}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}] = \text{Fix}(D).$$

In particular, if D is radical, $\text{Fix}(D) = \emptyset$ and (1.3.3) never converges to a consensus.

Proof. Limits of $\mathbf{W} \circ \mathbf{F}$ are, in the discrete case, fixed-points of $\mathbf{W} \circ \mathbf{F}$ by Remark 1.4.2, that is, $\text{Lim}(\mathbf{W} \circ \mathbf{F}) = \text{Fix}(\mathbf{W} \circ \mathbf{F})$. Accordingly, if $\mathbf{W}_{i,\mathcal{O}_i} > \frac{1}{2}$ for some $i \in [n]$, then, by Proposition 1.5.1, $\text{Fix}(D) = P_1[\text{Fix}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}] = P_1[\text{Lim}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}]$.

□

Example 1.5.5. The conditions $W_{i,\mathcal{O}_i} > \frac{1}{2}$ and fixed-point freeness of D might appear overly strong. Assuming a probabilistic analysis, for the moment, we find that fixed-point freeness is more likely the smaller the size of the opinion space, $m = |S|$. For $m = 2$, we have 1 fixed-point free deviation function D on $\{A, B\}$, among $m! - 1 = 1$ candidates (that is, all possible specifications of D are fixed-point free). For $m = 3$, there are 2 fixed-point free functions, among $m! - 1 = 5$ candidates, which is 40%. As m becomes large, this fraction approximates 1/3, as is well-known.[28] Now, assuming D is radical, we want to estimate the probability that $W_{i,\mathcal{O}_i} > \frac{1}{2}$ for some $i \in [n]$. For simplicity, let us assume that $W_{ij} = \frac{1}{n}$ for all $i, j \in [n]$, and that each agent $i \in [n]$ randomly opposes other agents $j \in [n]$ with probability $p \in [0, 1]$, that is, $P[F_{ij} = D] = p$; we assume independence across both i and j. Then, the probability that $W_{i,\mathcal{O}_i} \leq \frac{1}{2}$ equals the probability that $X \leq \frac{n}{2}$ where X is binomially distributed with parameters n (n trials) and p (success probability, that i opposes j, is p). Let $P(n; p)$ denote this probability, which equals $\sum_{k\leq \frac{n}{2}} \binom{n}{k} p^k (1-p)^{n-k}$. Then, the probability that all agents have $W_{i,\mathcal{O}_i} \leq \frac{1}{2}$ is just $P(n; p)^n$. Consequently, the probability that there is an agent with $W_{i,\mathcal{O}_i} > \frac{1}{2}$ is $1 - P(n; p)^n$. In Figure, 1.6 we plot this likelihood for $p = 0.30$, $p = 0.35$, $p = 0.40$, $p = 0.45$ and $p = 0.50$. Interestingly, there appears to be a bifurcation value — $p_0 = 0.50$ — such that if $p \geq p_0$, the probability that at least one agent has $W_{i,\mathcal{O}_i} > \frac{1}{2}$ goes to 1 as $n \to \infty$, while if $p < p_0$ the same probability converges to zero as $n \to \infty$. Hence, if $p \geq p_0$, for example, the probability that at least one agent's outgroup is decisive for him converges to 1 as $n \to \infty$. Thus, under fixed-point freeness of D, disagreement among such agents will obtain almost surely as $n \to \infty$.

A simple other condition that prevents the operator $\mathbf{W} \circ \mathbf{F}$ from inducing a consensus is, for example, the following. This condition is weaker than the previous because it says that disagreement obtains for *some* initial opinion vectors, while fixed-point freeness of D and decisiveness of outgroups, as discussed above, imply that disagreement obtains for *all* initial opinion vectors.

Proposition 1.5.2. If all agents oppose a certain agent, j', and otherwise $F_{ij} = F$ for all $i, j \in [n]$ with $j \neq j'$, then $\mathbf{W} \circ \mathbf{F}$ does not induce a consensus (that is, there exists $\mathbf{b}(0) \in S^n$ such that $\lim_{t\to\infty} (\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$ is not a consensus).

[28] See, e.g., http://www.math.umn.edu/~garrett/crypto/Overheads/06_perms_otp.pdf

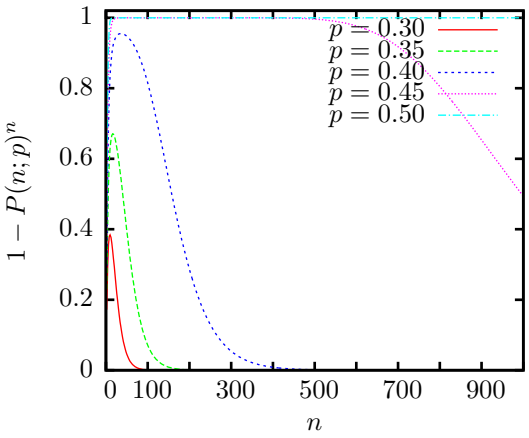


Figure 1.6: Probability $1 - P(n;p)^n$ that at least one agent $i \in [n]$ has $W_{i,\mathcal{O}_i} > \frac{1}{2}$ as a function of n and for five values of p. Description in text, Example 1.5.5.

Proof. Since D is not the identity, there exists c such that $c \neq D(c)$. Then a fixed-point of $\mathbf{W} \circ \mathbf{F}$ is given by $\mathbf{b}(0) = (D(c), \ldots, D(c), \underbrace{c}_{j'}, D(c), \ldots, D(c))^\intercal$. □

Remark 1.5.2. The last proposition also holds under the weaker assumption $W_{i,\mathcal{F}_i} + W_{ij'} > \frac{1}{2}$ for all $i = 1, \ldots, n$.

Now, to answer the first question — whether opposition may have no effect at all, in our current setup — let $\mathbf{F}_{D\to F}$ denote the matrix with all D's replaced by F's, i.e., the network links $\mathbf{F}$ with opposition 'inverted' to following. Then, it is easy to see that, in fact, $\mathbf{W} \circ \mathbf{F}$ and $\mathbf{W} \circ \mathbf{F}_{D\to F}$ may entail the exactly same limiting opinion results under opinion updating process (1.3.3), in the discrete case.

Example 1.5.6. Let $n = 4$, for example, and consider the operator $\mathbf{W} \circ \mathbf{F}$,

$$\mathbf{W} = \begin{pmatrix} \frac{3}{4} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \\ \star & \star & \star & \star \\ \star & \star & \star & \star \\ \star & \star & \star & \star \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & D & D & D \\ F & F & F & F \\ F & F & F & F \\ F & F & F & F \end{pmatrix},$$

where the weight structure of agents 2 to 4 may be arbitrarily specified; important is agent 1, who opposes agents 2 to 4, while the remaining agents follow all agents $j = 1, 2, 3, 4$. Then, no matter the opinion profile $\mathbf{b} \in S^n$, $(\mathbf{W} \circ \mathbf{F})\mathbf{b} = (\mathbf{W} \circ \mathbf{F}_{D\to F})\mathbf{b}$, which is obvious, since agent 1 assigns so much weight mass to himself (and follows himself) that he always adopts his own current opinion signal, no matter the opinion signals of agents 2 to 4. Consequently, $\lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t\mathbf{b}(0) = \lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F}_{D\to F})^t\mathbf{b}(0)$ for all $\mathbf{b}(0) \in S^n$.

Example 1.5.7. In the last example, agent $i = 1$ had assigned herself more than 50% weight mass (and followed herself) such that it is clear that her own current opinion always determines her next period opinion. A slightly more subtle example is the following, where none of the agents that i follows has more than 50% weight mass.

$$\mathbf{W} = \begin{pmatrix} \frac{40}{100} & \frac{30}{100} & \frac{21}{100} & \frac{9}{100} \\ \star & \star & \star & \star \\ \star & \star & \star & \star \\ \star & \star & \star & \star \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & F & F & D \\ F & F & F & F \\ F & F & F & F \\ F & F & F & F \end{pmatrix}.$$

Here, whenever two of the three agents $1, 2, 3$ agree, agent 1 adopts their opinion in the next period, irrespective of agent 4's opinion. If all three agents disagree, agent 1 adopts her own current opinion in the

next period, irrespective of agent 4's opinion. Hence, $\lim_{t\to\infty}(\mathbf{W}\circ\mathbf{F})^t\mathbf{b}(0) = \lim_{t\to\infty}(\mathbf{W}\circ\mathbf{F}_{D\to F})^t\mathbf{b}(0)$ for all $\mathbf{b}(0)\in S^n$.

There are, however, conditions on $\mathbf{W}$ that ensure that opposition 'matters'. One such condition is the following, whose idea is similar in spirit to that of Proposition 1.5.1, namely, it refers to 'too large weight mass' assigned to opposed agents. Before we state the proposition, we define the concept of an *influential group*.

Definition 1.5.2. Let an agent $i\in[n]$ be fixed. We call a group $\mathcal{N}\subseteq[n]$ *influential for agent* i if there exist two sets of agents $\mathcal{N}_1\subseteq[n]$ and $\mathcal{N}_2\subseteq[n]$ such that $(\mathcal{N},\mathcal{N}_1,\mathcal{N}_2)$ is a partition of $[n]$ (pairwise disjoint and whose union is $[n]$) and

$$W_{i,\mathcal{N}_1}+W_{i,\mathcal{N}}>\frac{1}{2}\quad\text{and}\quad W_{i,\mathcal{N}_2}+W_{i,\mathcal{N}}>\frac{1}{2}. \tag{1.5.2}$$

Remark 1.5.3. An influential group $\mathcal{N}$ for agent i is precisely what its name suggests: it may influence agent i's opinion. For instance, if agents in $\mathcal{N}_1$ hold opinion A and agents in $\mathcal{N}_2$ hold opinion B, then agents in $\mathcal{N}$ may 'turn the scales'.

Remark 1.5.4. If $\mathcal{N}$ is decisive for agent i, then it is influential for i.

Proposition 1.5.3. For some agent $i\in[n]$, let the group $\mathcal{O}_i$ of agents he opposes be influential for i, then it holds that

$$(\mathbf{W}\circ\mathbf{F})\mathbf{b}\neq(\mathbf{W}\circ\mathbf{F}_{D\to F})\mathbf{b}$$

for at least one opinion vector $\mathbf{b}\in S^n$.

Proof. Since $\mathcal{O}_i$, the agents i opposes, is an influential group, there is a partition $(\mathcal{O}_i,\mathcal{N}_1,\mathcal{N}_2)$ such that (1.5.2) holds; of course, i follows agents in $\mathcal{N}_1$ and $\mathcal{N}_2$. Let $S=\{A,B,\ldots\}$ consist of at least two elements and let, without loss of generality, $D(B)=A$. Let $\mathbf{b}$ be an opinion vector such that agents in $\mathcal{N}_1$ hold opinion A, agents in $\mathcal{N}_2$ hold opinion B and agents in $\mathcal{O}_i$ hold opinion B. Then

$$\big((\mathbf{W}\circ\mathbf{F})\mathbf{b}\big)_i=A,$$

since $W_{i,\mathcal{N}_1}+W_{i,\mathcal{O}_i}>\frac{1}{2}$ and $D(B)=A$, and

$$\big((\mathbf{W}\circ\mathbf{F}_{D\to F})\mathbf{b}\big)_i=B,$$

since $W_{i,\mathcal{N}_2}+W_{i,\mathcal{O}_i}>\frac{1}{2}$. □

Example 1.5.8. Many examples of $\mathbf{W}$ and $\mathbf{F}$ that satisfy the assumptions of Proposition 1.5.3 come to mind. One might be, for instance,

$$\mathbf{W}=\begin{pmatrix}\frac{1}{4}&\frac{1}{3}&\frac{1}{4}&\frac{1}{6}\\ \star&\star&\star&\star\\ \star&\star&\star&\star\\ \star&\star&\star&\star\end{pmatrix},\quad \mathbf{F}=\begin{pmatrix}F&D&F&F\\ \star&\star&\star&\star\\ \star&\star&\star&\star\\ \star&\star&\star&\star\end{pmatrix}.$$

Then, one partition as required by Proposition 1.5.3 is $\mathcal{N}_1=\{1\}$, $\mathcal{N}_2=\{3,4\}$, $\mathcal{O}_i=\{2\}$ with $W_{i,\mathcal{N}_1}=\frac{1}{4}$, $W_{i,\mathcal{N}_2}=\frac{5}{12}$, and $W_{i,\mathcal{O}_i}=\frac{1}{3}$. Hence,

$$W_{i,\mathcal{N}_1}+W_{i,\mathcal{O}_i}=\frac{1}{4}+\frac{1}{3}=\frac{7}{12}>\frac{1}{2},\quad\text{and}\quad W_{i,\mathcal{N}_2}+W_{i,\mathcal{O}_i}=\frac{5}{12}+\frac{1}{3}=\frac{3}{4}>\frac{1}{2}$$

so that, indeed, $\mathcal{O}_i$ is influential for $i=1$.

If we impose slightly more structure on D, we may give slightly more general conditions under which opposition 'matters'.

Proposition 1.5.4. Let there exist a non-fixed point B of D such that B is the tie-breaking element of $\{B, D(B)\}$. Moreover, for some agent $i \in [n]$, let the set $\mathcal{O}_i$ of agents he opposes be influential′ in the following manner. Rather than requirement (1.5.2), we assume the weaker form

$$W_{i,\mathcal{N}_1} + W_{i,\mathcal{O}_i} > \frac{1}{2} \quad \text{and} \quad W_{i,\mathcal{N}_2} + W_{i,\mathcal{O}_i} \geq \frac{1}{2}. \tag{1.5.3}$$

Then, it holds that

$$(\mathbf{W} \circ \mathbf{F})\mathbf{b} \neq (\mathbf{W} \circ \mathbf{F}_{D \to F})\mathbf{b}$$

for at least one opinion vector $\mathbf{b} \in S^n$.

Proof. As in the proof of Proposition 1.5.1, let $\mathbf{b}$ such that agents in $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{O}_i$ hold opinions $A = D(B)$, B, and B, respectively. □

We note that both hard opposition and soft opposition as specified in Section 1.3, and which assume an order $<$ on S, satisfy the condition on D specified in the last proposition under the 'natural notion' of tie-breaking element, cf. Example 1.5.1, as largest (or smallest) element of $M \subseteq S$. For instance, choose $B = \max S$, whence, since $D(B) = \min S$, B is the tie-breaking element of $\{B, D(B)\}$. Moreover, we note that weight requirement (1.5.3) in Proposition 1.5.4 is always satisfied in the case of uniform weights $\mathbf{W}$, which reproduces the ordinary ('unweighted') majority updating setup, when some agent i opposes at least one agent j. In other words, in the ordinary ('unweighted') majority updating setup, opposition always has an effect as long as D is, e.g., soft or hard opposition. This is what our next example shows more formally.

Example 1.5.9. Let $n \in \mathbb{N}$ and let $\mathbf{W} \circ \mathbf{F}$ be such that there exists an agent i with $W_{ij} = \frac{1}{n}$ for all agents $j = 1, \ldots, n$ and let $|\mathcal{O}_i| \geq 1$. We show that $\mathcal{O}_i$ is influential′ for i in the sense of requirement (1.5.3). To see this, let $\mathcal{F}_i$ be the set of agents that agent i follows. If $\mathcal{F}_i$ has even cardinality, let $\mathcal{N}_1$ and $\mathcal{N}_2$ be an arbitrary partition of $\mathcal{F}_i$ with $|\mathcal{N}_1| = |\mathcal{N}_2|$. Then, clearly, $W_{i,\mathcal{N}_k} + W_{i,\mathcal{O}_i} > \frac{1}{2}$ for $k = 1, 2$. If $\mathcal{F}_i$ has odd size, let $\mathcal{N}_1$ and $\mathcal{N}_2$ be an arbitrary partition of $\mathcal{F}_i$ such that $|\mathcal{N}_1| = |\mathcal{N}_2| + 1$. Then, clearly, $W_{i,\mathcal{N}_1} + W_{i,\mathcal{O}_i} > \frac{1}{2}$ and $W_{i,\mathcal{N}_2} + W_{i,\mathcal{O}_i} \geq \frac{1}{2}$. Hence, $\mathcal{O}_i$ is influential′ for i.

To be more precise on the example, let, e.g.,

$$\mathbf{W} = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & F & D & D & F \\ \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \end{pmatrix}.$$

Hence, for $i = 1$, choose $\mathcal{N}_1 = \{1, 2\}$, for example, and $\mathcal{N}_2 = \{5\}$, and $\mathcal{O}_i$, the set of agent i opposes, is $\mathcal{O}_i = \{3, 4\}$. Clearly, $W_{i,\mathcal{N}_1} + W_{i,\mathcal{O}_i} = \frac{2}{5} + \frac{2}{5} > \frac{1}{2}$ and $W_{i,\mathcal{N}_2} + W_{i,\mathcal{O}_i} = \frac{1}{5} + \frac{2}{5} > \frac{1}{2}$ so that $\mathcal{O}_i$ is indeed influential′ for i. Moreover, to have, in addition, the assumptions on D in Proposition 1.5.4 be satisfied, let, e.g., $S = \{A, B, C\}$ with $D(C) = A$, $D(B) = B$, and $D(A) = C$, where $A < B < C$ (note that D is soft opposition on S) and larger opinions are tie-breakers; then, C, a non-fixed point of D, is the tie-breaking element of $\{C, D(C) = A\}$. Hence, all assumptions of Proposition 1.5.4 are satisfied, and, accordingly, also its consequences. By the proof of the proposition, $\mathbf{b} = (A, A, C, C, C)^\intercal$ satisfies $(\mathbf{W} \circ \mathbf{F})\mathbf{b} \neq (\mathbf{W} \circ \mathbf{F}_{D \to F})\mathbf{b}$. In fact,

$$\big((\mathbf{W} \circ \mathbf{F})\mathbf{b}\big)_i = A,$$

while

$$\big((\mathbf{W} \circ \mathbf{F}_{D \to F})\mathbf{b}\big)_i = C.$$

Remark 1.5.5. We may summarize Propositions 1.5.1 to 1.5.4 as follows. First, we find that, in the discrete majority voting model, opposition may have no effect at all in that the same outcomes can obtain as in the setting without opposition (Examples 1.5.6 and 1.5.7). Intuitively and from the examples, we

feel that this must be related to the weight mass agents assign to opposed agents. Propositions 1.5.3 and 1.5.4 then show that opposition begins to matter once a single agent assigns 'large enough' weight mass $W_{i,\mathcal{O}_i}$ to opposed agents, i.e., his outgroup is influential for him. Weight mass requirements are not strong: they are satisfied for the ordinary ('unweighted') majority opinion update model, for example (see Example 1.5.9). Finally, Proposition 1.5.1 and Corollary 1.5.1 indicate that if $W_{i,\mathcal{O}_i}$ exceeds a critical value — $\frac{1}{2}$ in this setup; the group of agents i opposes becomes decisive for him — opposition becomes 'poisonous' in that it precludes consensus formation in the DeGroot learning model as long as, in addition, deviation function D is 'radical' in that it has no fixed points. In other words, if D is radical and if a single agent's outgroup is decisive for him, disagreement among agents (within a period or between periods) is the prediction of our discrete DeGrootian opinion dynamics model, no matter the initial opinions of agents. Fixed-point freeness may not be too surprising an occurrence, however, as the example of the binary model, with $S = \{A, B\}$, suggests. Here, the only legitimate specification of D is fixed-point free. Moreover, even if D is not radical, Proposition 1.5.1 shows that, in the case of opposition, agents can only attain neutral consensus opinions as long as a single agent has sufficient 'distrust'. If $\mathrm{Fix}(D)$ is small, as we would typically expect, most consensus opinions can, accordingly, never be attained.

Polarization

We now investigate *polarization* as an outcome of our opinion updating dynamics. Note that, in real societies, polarization on many agendas is frequently observed such as whether the Christian churches, or the law, should allow condoms or gay marriages. In fact, as we have discussed, polarizing viewpoints may occur, prominently, in the political arena and in the situation of 'countercultural' subsocieties, with respect to the viewpoints held by the 'mainstream' culture. We first define the concept formally.

Definition 1.5.3 ((Functional) Polarization)**.** We call an opinion vector $\mathbf{p} \in S^n$ a *polarization* if $\mathbf{p}$ consists of two distinct elements $a, b \in S$ exclusively.

We call an opinion vector $\mathbf{p} \in S^n$ a *functional polarization* if $\mathbf{p}$ is a polarization and a and b are opposing viewpoints.

The concept of functional polarization, which depends on the definition of deviation function D, captures the notion of 'opposing viewpoints' expressed in a polarization vector $\mathbf{p}$, while an 'ordinary' polarization vector may consist of disagreeing viewpoints solely, that stand in no relationship to each other. Next, we define network structures that are sufficient for inducing polarization opinion vectors.

Definition 1.5.4 (Opposition bipartite operator $\mathbf{W} \circ \mathbf{F}$)**.** We call the operator $\mathbf{W} \circ \mathbf{F}$ on n agents *opposition bipartite* if there exists a partition $(\mathcal{N}_1, \mathcal{N}_2)$ of the set of agents $[n]$ into two disjoint non-empty subsets — $[n] = \mathcal{N}_1 \cup \mathcal{N}_2$, with $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$, $\mathcal{N}_i \neq \emptyset$, for $i = 1, 2$ — such that agents in $\mathcal{N}_i$ follow each other, for $i = 1, 2$, while for all agents $i_0 \in \mathcal{N}_i$, $i_1 \in \mathcal{N}_{-i}$, for $i = 1, 2$, it holds that i_0 deviates from i_1. More precisely, we require

$$\forall\, i_0, i_1 \in \mathcal{N}_i \Big(W_{i_0 i_1} > 0 \implies F_{i_0 i_1} = F \Big), \quad \text{for } i = 1, 2,$$
$$\forall\, i_0 \in \mathcal{N}_i, i_1 \in \mathcal{N}_{-i} \Big(W_{i_0 i_1} > 0 \implies F_{i_0 i_1} = D \Big), \quad \text{for } i = 1, 2.$$

Remark 1.5.6. What we call 'opposition bipartite' operator — or at least a special case of our concept — has also been called 'balanced signed network' in the literature (cf. Beasley and D. Kleinberg, 2010).

Definition 1.5.5 (Anti-Opposition bipartite operator $\mathbf{W} \circ \mathbf{F}$)**.** We call the operator $\mathbf{W} \circ \mathbf{F}$ on n agents *anti-opposition bipartite* if there exists a partition $(\mathcal{N}_1, \mathcal{N}_2)$ of the set of agents $[n]$ into two disjoint non-empty subsets such that agents in $\mathcal{N}_i$ deviate from each other, for $i = 1, 2$, while for all agents $i_0 \in \mathcal{N}_i$, $i_1 \in \mathcal{N}_{-i}$, for $i = 1, 2$, it holds that i_0 follows i_1.

An example of an opposition bipartite operator is given in Example 1.5.10 below. An example of an anti-opposition bipartite operator is given in Examples 1.5.12 below and 1.5.3 above. A schematic illustration of both concepts is given in Figure 1.7.

We now show that opposition bipartite networks have polarization opinion vectors as fixed-points.

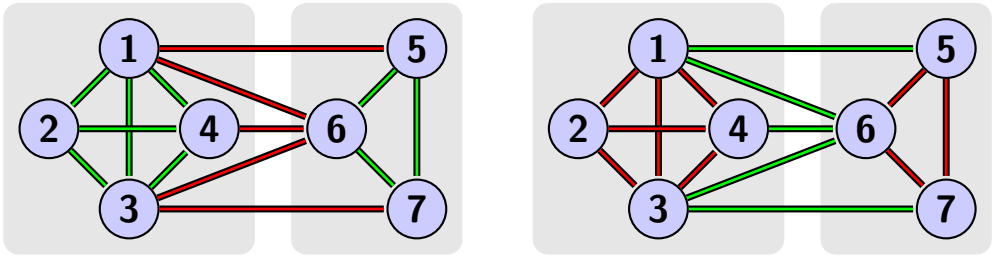


Figure 1.7: Schematic illustration of the concepts of opposition bipartite (left) and anti-opposition bipartite operators (right). We omit network links referring to weights $\mathbf{W}$ for clarity and we also draw links as undirected for the same reason. We omit links F_{ij} where $W_{ij} = 0$. Red links denote opposition, green links following.

Proposition 1.5.5. Let $\mathbf{W} \circ \mathbf{F}$ be opposition bipartite and let $a, b \in S$ be opposing viewpoints. Then, there exists a polarization opinion vector $\mathbf{p}$ consisting of opinions a and b such that $(\mathbf{W} \circ \mathbf{F})\mathbf{p} = \mathbf{p}$.

Proof. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the partition of the agent set $[n] = \{1, \dots, n\}$ such that agents in $\mathcal{N}_i$, $i = 1, 2$, follow each other, while agents across the two sets oppose each other. Let $a, b \in S$ be such that $D(a) = b$ and $D(b) = a$. Moreover, let $\mathbf{p}$ be such that each agent in $\mathcal{N}_1$ holds opinion a (or b) and each agent in $\mathcal{N}_2$ holds opinion b (or a). Then, for each agent $i_1 \in \mathcal{N}_1$, all neighbors' (possibly inverted) opinion signals are a (or b) and analogously for agents $i_2 \in \mathcal{N}_2$. □

Example 1.5.10. Let $\mathbf{W}$ be arbitrary. Consider

$$\mathbf{F} = \begin{pmatrix} F & F & D & D \\ F & F & D & D \\ D & D & F & F \\ D & D & F & F \end{pmatrix}.$$

Clearly, $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite; for example, take $\mathcal{N}_1 = \{1, 2\}$ and $\mathcal{N}_2 = \{3, 4\}$. Moreover, let $S = \{\text{“impossible”}, \text{“unlikely”}, \text{“possible”}, \text{“likely”}, \text{“certain”}\}$ as above with D as soft opposition. Then $\mathbf{p} = \big(\text{“unlikely”}, \text{“unlikely”}, \text{“likely”}, \text{“likely”}\big)^\intercal$ is a polarization fixed-point of $\mathbf{W} \circ \mathbf{F}$, amongst others.

Note that Proposition 1.5.5 would also be true under weaker conditions such as a 'perturbed opposition bipartite operator', as we define in the following.

Definition 1.5.6 (Perturbed opposition bipartite operator)**.** We call the operator $\mathbf{W} \circ \mathbf{F}$ on n agents *perturbed opposition bipartite* if there exists a partition $(\mathcal{N}_1, \mathcal{N}_2)$ of the set of agents $[n]$ into two disjoint non-empty subsets such that for each agent $i = 1, \dots, n$, there exists a group of agents $A_i \cup B_i \subseteq [n]$, with $A_i \subseteq \mathcal{N}_1$ and $B_i \subseteq \mathcal{N}_2$ and i follows agents in A_i and deviates from agents in B_i, such that the group $A_i \cup B_i$ is decisive for i, i.e., $W_{i, A_i \cup B_i} > \frac{1}{2}$.

Example 1.5.11. Consider

$$\mathbf{W} = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & F & D & F \\ D & F & D & D \\ D & D & D & F \\ D & F & F & F \end{pmatrix},$$

Taking $\mathcal{N}_1 = \{1, 2\}$ and $\mathcal{N}_2 = \{3, 4\}$, we see that $\mathbf{W} \circ \mathbf{F}$ is perturbed opposition bipartite. For example, for agent 1, we would, e.g., have $A_1 = \{1, 2\}$, $B_1 = \{3\}$ with $W_{1, A_1 \cup B_1} = \frac{3}{4} > \frac{1}{2}$; for agent 2, e.g., $A_2 = \{2\}$ and $B_2 = \{3, 4\}$ and $W_{2, A_2 \cup B_2} = \frac{3}{4} > \frac{1}{2}$, etc. Perturbed opposition bipartite networks also have polarization vectors as fixed-points, as seen in this example, e.g.:

$$\begin{pmatrix} \text{“unlikely”} \\ \text{“unlikely”} \\ \text{“likely”} \\ \text{“likely”} \end{pmatrix} \mapsto_{\mathbf{W} \circ \mathbf{F}} \begin{pmatrix} \text{“unlikely”} \\ \text{“unlikely”} \\ \text{“likely”} \\ \text{“likely”} \end{pmatrix}.$$

Anti-opposition bipartite networks induce oscillating, or fluctuating, opinion updates (cf. Kramer, 1971), very similar to ordinary periodic networks as discussed above.

Proposition 1.5.6. Let $\mathbf{W} \circ \mathbf{F}$ be anti-opposition bipartite and let $a, b \in S$ be opposing viewpoints. Then, there exist polarization opinion vectors $\mathbf{p}, \bar{\mathbf{p}} \in S^n$ consisting of opinions a and b in a complementary manner — $p_i = D(\bar{p}_i)$ and $\bar{p}_i = D(p_i)$ for all $i = 1, \ldots, n$ — such that $(\mathbf{W} \circ \mathbf{F})\mathbf{p} = \bar{\mathbf{p}}$ and $(\mathbf{W} \circ \mathbf{F})\bar{\mathbf{p}} = \mathbf{p}$.

Proof. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the partition of the agent set $[n] = \{1, \ldots, n\}$ such that agents in $\mathcal{N}_i$, $i = 1, 2$, deviate from each other, while agents across the two sets follow each other. Let $a, b \in S$ be such that $D(a) = b$ and $D(b) = a$. Moreover, let $\mathbf{p}$ be such that each agent in $\mathcal{N}_1$ holds opinion a (or b) and each agent in $\mathcal{N}_2$ holds opinion b (or a) and let $\bar{\mathbf{p}}$ have a complementary distribution of a's and b's. Then, for each agent $i_1 \in \mathcal{N}_1$, all neighbor's (possibly inverted) opinion signals are b (or a) and analogously for agents $i_2 \in \mathcal{N}_2$. □

Example 1.5.12. Let $\mathbf{W}$ be arbitrary. Consider

$$\mathbf{F} = \begin{pmatrix} D & D & F & F \\ D & D & F & F \\ F & F & D & D \\ F & F & D & D \end{pmatrix}.$$

Clearly, $\mathbf{W} \circ \mathbf{F}$ is anti-opposition bipartite; for example, take $\mathcal{N}_1 = \{1, 2\}$ and $\mathcal{N}_2 = \{3, 4\}$. For $\mathbf{p}$ as in Example 1.5.10, we have

$$\begin{pmatrix} \text{“unlikely”} \\ \text{“unlikely”} \\ \text{“likely”} \\ \text{“likely”} \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} \text{“likely”} \\ \text{“likely”} \\ \text{“unlikely”} \\ \text{“unlikely”} \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \begin{pmatrix} \text{“unlikely”} \\ \text{“unlikely”} \\ \text{“likely”} \\ \text{“likely”} \end{pmatrix} \mapsto_{\mathbf{W}\circ\mathbf{F}} \cdots$$

Of course, we could, in addition, define a concept of ‘perturbed anti-opposition bipartite operator’ and easily see that Proposition 1.5.6 also holds under this weaker concept, but we omit the details here because of analogy with the concept of ‘perturbed opposition bipartite operator’.

Since neither fluctating opinion updates nor polarization consitute a consensus, we have the following simple corollary to Propositions 1.5.5 and 1.5.6.

Corollary 1.5.2. Let $\mathbf{W} \circ \mathbf{F}$ be (perturbed) opposition bipartite or anti-opposition bipartite. Then there exist initial opinion vectors $\mathbf{b}(0) \in S^n$ such that $\mathbf{W} \circ \mathbf{F}$ does not induce a consensus for $\mathbf{b}(0)$. If $\mathbf{W} \circ \mathbf{F}$ is anti-opposition bipartite, then there exist initial opinion vectors $\mathbf{b}(0) \in S^n$ such that $\mathbf{W} \circ \mathbf{F}$ does not even converge for $\mathbf{b}(0)$.

We conclude with an example of how to induce more general polarization outcomes, between more than two groups of agents, and a simulation of the discrete weighted majority opinion updating model (1.3.3). In the latter example, rather than discussing (possible or impossible) fixed points of operators, we simulate *actual* dynamics.

Example 1.5.13. We briefly discuss how to induce, in a general manner, polarizing viewpoints between more than two groups of agents as fixed-points of the operator $\mathbf{W} \circ \mathbf{F}$. One way to achieve such more fragmented opinion and belief systems in society in our setup is to endow the different groups with *different* deviation functions $D_k : S \to S$, where k ranges over the groups (or agents). In essence, these different deviation functions would represent distinct interpretations of what the opposite of a certain opinion value $a \in S$ is. For example, one group might interpret opposition in a radical manner, allowing D_k to have no fixed-points while other groups may be more ‘tolerant’, leaving some opinion values unchanged, even in opposition modus.[29]

To make a concrete example, let $S = \{A, B, C, \ldots\}$ and consider three different groups with distinct deviation functions $D_1(x) = A$, $D_2(x) = B$, $D_3(x) = C$ for all $x \in \{A, B, C\}$ (or even all $x \in S$). For

[29] A generalization is to let the deviation endomorphisms depend, not only on the agents who oppose, but also on the opposed agents.

instance, group 1 might always deviate to an extreme left wing opinion, at least within the set $\{A, B, C\}$, provided that it deviates from certain agents; group 2 to a moderate position in the opinion space; and group 3 to an extreme right wing position. Let, e.g.,

$$\mathbf{W} = \frac{1}{6}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & F & F & D_1 & D_1 & D_1 \\ F & F & F & D_1 & D_1 & D_1 \\ F & F & F & D_1 & D_1 & D_1 \\ D_2 & D_2 & D_2 & F & D_2 & D_2 \\ D_3 & D_3 & D_3 & D_3 & F & F \\ D_3 & D_3 & D_3 & D_3 & F & F \end{pmatrix}.$$

We then have a partition $(\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3)$ of the agent set — $(\{1, 2, 3\}, \{4\}, \{5, 6\})$ in the example — such that agents within each subset $\mathcal{N}_i$ follow each other and agents across the subsets deviate from each other, applying their specific choices of deviation functions. It is easy to check that, e.g., $\mathbf{p} = (A, A, A, B, C, C)^\intercal$ is a fixed-point of $\mathbf{W} \circ \mathbf{F}$, constituting a 'generalized' polarization opinion vector.

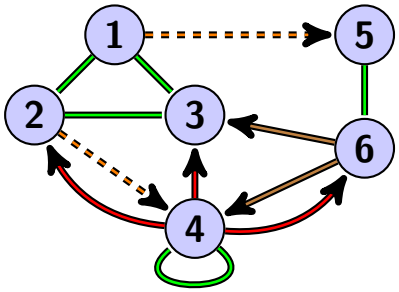


Figure 1.8: Graphical illustration of Example 1.5.13. Groups have individualized deviation functions, in different colors. We omit many links for clarity.

Example 1.5.14. To visualize the likelihood of a consensus in our current setup, we plot in Figure 1.9 the following quantities. We run a simulation where we choose weights W_{ij} from a uniform random distribution on $(0, 1)$ and then normalize in order for $\mathbf{W}$ to be row-stochastic. We draw F_{ij} according to the Bernoulli distribution $P[F_{ij} = D] = p$, with $p \in [0, 1]$. We let D be hard opposition on the set $S = \{-\alpha, -\alpha + 1, \ldots, 0, \ldots, \alpha - 1, \alpha\}$, for $\alpha \in \{1, 2, 3\}$; conventionally, we let $D(0) = 0$. For n agents ($n = 5$ in the figure), we then iterate over all possible distributions of initial opinion profiles $\mathbf{b}(0) \in S^n$ — there are $|S|^n$ different such profiles — and determine the fraction of profiles that result in a consensus among the $|S|^n$ total initial opinion profiles, that is, for which it holds that $\lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})\mathbf{b}(0)$ is a consensus. In the figure, we plot this fraction as a function of p; the displayed curves are averages over 20 simulations. We note that the probability of a consensus appears to be a decreasing function of p, opposition likelihood, as we expect. In the case of hard opposition, at least, consensus likelihood obviously decreases in α, the 'size' of S.

1.6 The continuous DeGroot model

1.6.1 The requirement $\sum_{j=1}^{n} W_{ij} = 1$

At first, we consider here the condition when the importance matrix $\mathbf{W}$ is row-stochastic, that is,

$$0 \leq W_{ij} \leq 1, \quad \text{and,} \quad \sum_{j=1}^{n} W_{ij} = 1 \tag{1.6.1}$$

for all $i = 1, \ldots, n$. As mentioned, this means that the weights that agents assign each other are normalized to unity, which is the usual assumption in DeGroot-like models.

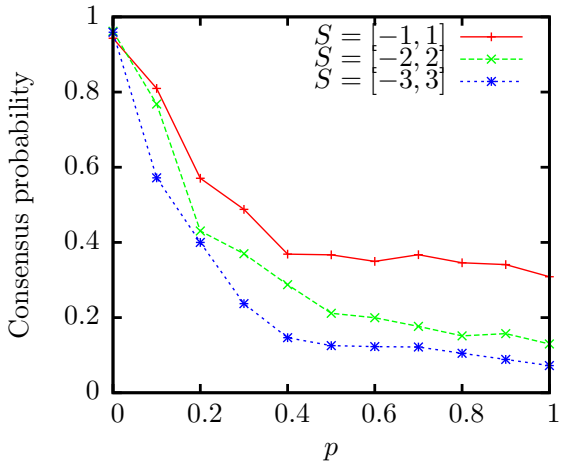


Figure 1.9: Consensus probability as a function of p. We denote the discrete interval $\{-\alpha, -\alpha+1, \ldots, \alpha-1, \alpha\}$ by $[-\alpha, \alpha]$, for short. Description in the text.

Proposition 1.6.1. Let $\mathbf{W} \circ \mathbf{F}$ be an arbitrary operator. Then, for any $c \in S$,

$$c \in \mathrm{Fix}(D) \implies (c, \ldots, c)^\intercal \in \mathrm{Fix}(\mathbf{W} \circ \mathbf{F}).$$

Moreover, if $W_{i,\mathcal{O}_i} > 0$ for some $i \in [n]$, then, for all $c \in S$,

$$c \notin \mathrm{Fix}(D) \implies (c, \ldots, c)^\intercal \notin \mathrm{Fix}(\mathbf{W} \circ \mathbf{F}).$$

In other words, if $W_{i,\mathcal{O}_i} > 0$ for some $i \in [n]$, then

$$\mathrm{Fix}(D) = P_1[\mathrm{Fix}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}].$$

Proof. Let $\mathbf{c} = (c, \ldots, c)^\intercal$.

If $c = D(c)$ for some $c \in S$, then clearly $(\mathbf{W} \circ \mathbf{F})\mathbf{c} = \mathbf{c}$ by the definition of $\mathbf{W} \circ \mathbf{F}$ since for each agent $i \in [n]$,

$$\big((\mathbf{W} \circ \mathbf{F})\mathbf{c}\big)_i = \sum_{j \in \mathcal{F}_i} W_{ij} c + \sum_{j \in \mathcal{O}_i} W_{ij} D(c) = c \sum_{j \in [n]} W_{ij} = c = (\mathbf{c})_i.$$

Conversely, let $c \neq D(c)$ for some $c \in S$. Let $i \in [n]$ be such that $F_{ij} = D$ and $W_{ij} > 0$ for some $j \in [n]$. If $\mathbf{c} = (c, \ldots, c)^\intercal$ were a fixed-point of $\mathbf{W} \circ \mathbf{F}$, then

$$c = \sum_{j \in \mathcal{O}_i} W_{ij} D(c) + \sum_{j \in \mathcal{F}_i} W_{ij} c = D(c) W_{i,\mathcal{O}_i} + c(1 - W_{i,\mathcal{O}_i}),$$

which implies that

$$0 = W_{i,\mathcal{O}_i}(D(c) - c),$$

which is impossible since $W_{i,\mathcal{O}_i} > 0$ by assumption. □

As a simple corollary to Proposition 1.6.1, we find that the possible consensus limiting opinions of $\mathbf{W} \circ \mathbf{F}$ are given by the set of fixed points of D when D is continuous. In other words, under opposition, agents can only converge to consensus vectors in which the consensus value is a neutral opinion. The corollary mimics the corresponding 'discrete case' corollary in the same way that Proposition 1.6.1 mimics Proposition 1.5.1, *mutatis mutandis*.

Corollary 1.6.1. Let D be continuous. Then, if $W_{i,\mathcal{O}_i} > 0$ for some $i \in [n]$,

$$P_1[\mathrm{Lim}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}] = \mathrm{Fix}(D).$$

In particular, if D is radical, $\mathrm{Fix}(D) = \emptyset$ and (1.3.3) never converges to a consensus.

Proof. If D is continuous, limits of $\mathbf{W} \circ \mathbf{F}$ are fixed-points of $\mathbf{W} \circ \mathbf{F}$ by Remark 1.4.2, that is, $\mathrm{Lim}(\mathbf{W} \circ \mathbf{F}) = \mathrm{Fix}(\mathbf{W} \circ \mathbf{F})$. Accordingly, if $W_{i,\mathcal{O}_i} > 0$ for some $i \in [n]$, then, by Proposition 1.6.1, $\mathrm{Fix}(D) = P_1[\mathrm{Fix}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}] = P_1[\mathrm{Lim}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}]$. □

Remark 1.6.1. As in the discrete case, the proposition may imply long-run disagreement, for all initial opinions, for instance, when D is fixed point free.

By Brouwer's fixed point theorem, however, stated in the appendix, D always has a fixed point as long as S is convex (our standard assumption) and compact and D is continuous. Hence, in this situation, updating process (1.3.3) always converges to at least one consensus opinion vector, given appropriate initial opinions, even under opposition. But also note, however, that one of our prime exemplars of a deviation function, hard opposition, is not a continuous function.

Remark 1.6.2. We mention the following generalization of Corollary 1.6.1 in case agents have individualized deviation functions $D_i : S \to S$ as in Example 1.5.14. In this situation, if each D_i is continuous and if $W_{i,\mathcal{O}_i} > 0$ for all i in a subset $A \subseteq [n]$, then

$$P_1[\mathrm{Lim}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}] = \bigcap_{i \in A} \mathrm{Fix}(D_i),$$

which implies that $P_1[\mathrm{Lim}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}] = \emptyset$ as soon as two $D_i, D_{i'}$, have disjoint sets of fixed-points. Of course, this generalization already applies to Proposition 1.6.1 such that $P_1[\mathrm{Fix}(\mathbf{W} \circ \mathbf{F}) \cap \mathcal{C}] = \bigcap_{i \in A} \mathrm{Fix}(D_i)$ whenever $W_{i,\mathcal{O}_i} > 0$ for all i in A.

We now want to study *actual* limiting behavior of opinion updating process (1.3.3), that is, we ask: what does (1.3.3) converge to (if it converges), rather than what can it possibly converge to (if at all)? An analytically (more or less) tractable situation arises when $S = [\alpha, \beta]$ and D is soft opposition, $D(x) = \alpha + \beta - x$ for all $x \in S$. In this case, D is affine-linear and, as may be clear and as we also show in the proof of Proposition 1.6.2 below, then also $\mathbf{W} \circ \mathbf{F}$ is affine-linear, allowing the representation, for all $\mathbf{x} \in S^n$, $(\mathbf{W} \circ \mathbf{F})\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{d}$, for some matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and some vector $\mathbf{d} \in \mathbb{R}^n$. As we have indicated in Section 1.4, and, in particular, in Theorem 1.4.3, in this situation, if matrix $\mathbf{A}$ has spectral radius smaller than 1, then $\mathbf{W} \circ \mathbf{F}$ is a contraction mapping and, thus, by the Banach fixed point theorem, $(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$ converges to the unique fixed point of $\mathbf{W} \circ \mathbf{F}$. Now, by Proposition 1.6.1, the fixed points of $\mathbf{W} \circ \mathbf{F}$ comprise the fixed points of D, that is, the neutral opinions. As we can easily verify, soft opposition D has precisely one fixed-point, namely, $c = \frac{\alpha+\beta}{2}$, from which we consequently conclude that, in the situation of the Banach fixed point theorem, $\mathbf{W} \circ \mathbf{F}$ induces the consenus $(c, \ldots, c)^\intercal$, for all initial opinion vectors $\mathbf{b}(0) \in S^n$. This is our next proposition.

Proposition 1.6.2. Let $S = [\alpha, \beta]$ and let D be soft opposition. Then, $\mathbf{W} \circ \mathbf{F}$ is an affine-linear operator of the form $\mathbf{A}\mathbf{x} + \mathbf{d}$. If $\rho(\mathbf{A}) < 1$, then $\mathbf{W} \circ \mathbf{F}$ induces the unique consensus $\frac{\alpha+\beta}{2}$, for all initial opinion profiles $\mathbf{b}(0) \in S^n$.

Proof. The proposition is clear, except maybe for the representation of $\mathbf{W} \circ \mathbf{F}$ as an affine-linear operator. For agent $i = 1, \ldots, n$, we have

$$\begin{aligned}
\big((\mathbf{W} \circ \mathbf{F})\mathbf{x}\big)_i &= \sum_{j \in \mathcal{F}_i} W_{ij} x_j + \sum_{j \in \mathcal{O}_i} W_{ij} D(x_j) = \sum_{j \in \mathcal{F}_i} W_{ij} x_j + \sum_{j \in \mathcal{O}_i} W_{ij}(\alpha + \beta - x_j) \\
&= \sum_{j \in \mathcal{F}_i} W_{ij} x_j + \sum_{j \in \mathcal{O}_i} (-W_{ij}) x_j + (\alpha + \beta) \sum_{j \in \mathcal{O}_i} W_{ij} \\
&= \sum_{j \in \mathcal{F}_i} W_{ij} x_j + \sum_{j \in \mathcal{O}_i} (-W_{ij}) x_j + (\alpha + \beta) W_{i,\mathcal{O}_i}.
\end{aligned}$$

Thus, we can set $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{d} \in \mathbb{R}^n$ with

$$A_{ij} = \begin{cases} W_{ij} & \text{if } F_{ij} = F, \\ -W_{ij} & \text{if } F_{ij} = D, \end{cases}, \qquad d_i = (\alpha+\beta)W_{i,\mathcal{O}_i}. \tag{1.6.2}$$

□

Example 1.6.1. Examples that satisfy the assumptions of Proposition 1.6.2 are, for instance, given in Examples 1.4.3 and 1.4.4 above.

Next, we show that in case D is soft opposition on $S = [\alpha, \beta]$, we come 'quite close' to having the condition $\rho(\mathbf{A}) < 1$ satisfied, in case $\mathbf{W}$ is row-stochastic.

Proposition 1.6.3. Let $S = [\alpha, \beta]$ and let D be soft opposition. Then, for the operator $\mathbf{W} \circ \mathbf{F}$ with the representation $(\mathbf{A}, \mathbf{d})$, we have

$$\rho(\mathbf{A}) \leq 1.$$

Proof. Consider matrix $\mathbf{A}$ as defined in (1.6.2). We have, for all $i = 1, \ldots, n$, $\sum_{j=1}^n |A_{ij}| = \sum_{j=1}^n W_{ij} = 1$, and therefore,

$$\|\mathbf{A}\|_\infty = 1,$$

where $\|\cdot\|_\infty$ is the row sum norm, defined in Definition 1.4.17. Moreover, by Theorem 1.4.5, it holds that

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|_p,$$

for any p-norm and any matrix $\mathbf{A}$. □

Remark 1.6.3. If $F_{ij} = F$ for all $i, j \in [n]$, then $\mathbf{A} = \mathbf{W}$ by (1.6.2) and $\rho(\mathbf{A}) = \rho(\mathbf{W}) = 1$ such that $\mathbf{W} \circ \mathbf{F}$ never is a contraction mapping in this case. To see that $\rho(\mathbf{W}) = 1$ is easy: any non-zero consensus is a fixed-point of a row-stochastic matrix such that there exists an eigenvalue $\lambda = 1$ of $\mathbf{W}$. In other words, *in the original DeGroot opinion dynamics model, without opposition,* $\mathbf{W} \circ \mathbf{F}$ *cannot be a contraction mapping.*

A crucial question is, of course, what the condition $\rho(\mathbf{A}) < 1$ in Proposition 1.6.2 actually *means* in terms of multigraph structure. Below, in Proposition 1.6.6, we consider this question for the situation when $\mathbf{A}$ is strictly positive in each entry, and, in Theorem 1.6.2, in the situation when $\mathbf{A}$ is symmetric and when $A_{ii} = 0$. In short, the condition $\rho(\mathbf{A}) < 1$, which is the 'Banach fixed point theorem condition', is equivalent, under the named assumptions, to the condition that the multigraph $\mathbf{W} \circ \mathbf{F}$ is 'unbalanced' in that, e.g., two agents A and B have mutual friends but their friendship networks are not identical such that, e.g., A opposes a friend of B. Apparently, this causes some inconsistency in the network — e.g., a violation of 'friendship transitivity' — and, ultimately, leads agents to neutrality, where everyone holds an uncontroversial opinion.[30]

Polarization

As in the discrete majority model, we now discuss polarization of opinions. Our first proposition is identical to the corresponding proposition in the discrete case.

Proposition 1.6.4. Let $\mathbf{W} \circ \mathbf{F}$ opposition bipartite and let $a, b \in S$ be opposing viewpoints. Then, there exists a polarization opinion vector $\mathbf{p}$ of opinions a and b such that $(\mathbf{W} \circ \mathbf{F})\mathbf{p} = \mathbf{p}$.

[30] E.g., polarization cannot be upheld because of such inconsistencies as indicated.

Proof. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the partition of the agent set $[n] = \{1, \dots, n\}$ such that agents in $\mathcal{N}_i$, $i = 1, 2$, follow each other, while agents across the two sets oppose each other. Let $a, b \in S$ be such that $D(a) = b$ and $D(b) = a$. Moreover, let $\mathbf{p}$ be such that each agent in $\mathcal{N}_1$ holds opinion a (or b) and each agent in $\mathcal{N}_2$ holds opinion b (or a). Then, for each agent $i_1 \in \mathcal{N}_1$:

$$\big((\mathbf{W} \circ \mathbf{F})\mathbf{p}\big)_{i_1} = \sum_{j \in \mathcal{N}_1} W_{i_1 j} a + \sum_{j \in \mathcal{N}_2} W_{i_1 j} D(b) = a(\sum_{j \in \mathcal{N}_1} W_{i_1 j} + \sum_{j \in \mathcal{N}_2} W_{i_1 j}) = a = p_{i_1} = (\mathbf{p})_{i_1},$$

and analogously for agents in $\mathcal{N}_2$. □

Our next proposition is a strengthening of the above in the case D is affine-linear. Namely, in this situation, we can give conditions such that $\mathbf{W} \circ \mathbf{F}$ converges to a polarization, no matter the initial opinions $\mathbf{b}(0)$, as long as $\mathbf{F}$ is opposition bipartite.

Proposition 1.6.5. Let D be soft opposition on $S = [\alpha, \beta]$ such that $\mathbf{W} \circ \mathbf{F}$ is affine-linear with representation $(\mathbf{A}, \mathbf{d})$. Then, if $\mathbf{F}$ is opposition bipartite, $\lambda = 1$ is an eigenvalue of $\mathbf{A}$. If $\lambda = 1$ is the only eigenvalue of $\mathbf{A}$ on the unit circle and if $\lambda = 1$ has algebraic multiplicity of 1, then $\lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0) = \mathbf{p}$ for some polarization opinion vector $\mathbf{p}$ (that depends on $\mathbf{b}(0)$) and all initial opinion vectors $\mathbf{b}(0) \in S^n$.

Proof. We prove the proposition in the case $\beta > 0$ and $\alpha = -\beta$.

To show that $\lambda = 1$ is an eigenvalue of $\mathbf{A}$ is simple. We need $\mathbf{A}\mathbf{x} = \mathbf{x}$ for some $\mathbf{x}$. Let a, b such that $a = -b$ with $a \neq 0$. Now, let $x_i = b$ if $i \in \mathcal{N}_1$ and $x_i = a$ if $i \in \mathcal{N}_2$, for all $i = 1, \dots, n$, where $(\mathcal{N}_1, \mathcal{N}_2)$ is the partition of the agent set $[n]$ that arises since $\mathbf{F}$ is opposition bipartite. Then, clearly, $\mathbf{A}\mathbf{x} = \mathbf{x}$. Now, since $\lambda = 1$ is the only eigenvalue of $\mathbf{A}$ on the unit circle and since $\lambda = 1$ is semisimple (since the algebraic multiplicity m_a of λ, which is 1, equals the geometric multiplicity m_g, since $m_a \geq m_g$ in general and $m_g \geq 1$ in our situation), $\lim_{t\to\infty} \mathbf{A}^t$ converges by Theorem 1.4.4. Moreover, it is well-known that $\mathbf{A}^t \mathbf{b}(0)$ converges to an eigenvector of $\mathbf{A}$ corresponding to $\lambda = 1$ in this situation, for any $\mathbf{b}(0)$ (see, e.g., Meyer, 2000, p.630). Since the eigenspace corresponding to $\lambda = 1$ has dimension 1 (geometric multiplicity of $\lambda = 1$ of 1) and since $\mathbf{x}$ as above is a polarization eigenvector, each eigenvector of $\mathbf{A}$ corresponding to $\lambda = 1$ is a polarization. □

Remark 1.6.4. Again, the proposition is abstract in that it gives conditions on the spectral radius of matrix $\mathbf{A}$ that ensure polarization but does not state what these conditions mean in terms of multigraph structure. In Theorem 1.6.2 below, we fill this gap and characterize, in graph theoretic terms, the condition, for instance, "$\lambda = 1$ is the only eigenvalue of $\mathbf{A}$ on the unit circle and has algebraic multiplicity of 1".

Remark 1.6.5. For the subsequent analysis, let $S = [-\beta, \beta]$ for convenience such that $\mathbf{d} = \mathbf{0}$.

How does $\mathbf{p}(\infty) := \mathbf{p} = \lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$ in Proposition 1.6.5 depend on the initial opinions $\mathbf{b}(0)$? One way to think of this limiting polarization is in terms of *social influence* vectors $\mathbf{s} \in \mathbb{R}^n$ such that $\|s\|_1 = \sum_{k=1}^n |s_k| = 1$ (cf. Jackson, 2009; Golub and Jackson, 2010). Denoting the two opposing viewpoints a and b (with $D(b) = -b = a$ and $D(a) = -a = b$) in polarization vector $\mathbf{p}(\infty)$ by $a(\infty)$ and $b(\infty)$, respectively, and assuming that a relationship

$$\begin{aligned} a(\infty) &= \mathbf{s}^\intercal \mathbf{b}(0) = \sum_{i=1}^n s_i b_i(0), \\ b(\infty) &= \bar{\mathbf{s}}^\intercal \mathbf{b}(0) = \sum_{i=1}^n D(s_i) b_i(0), \end{aligned}$$

exists, for all initial opinions vectors $\mathbf{b}(0)$ — that is, limiting polarization is a linear combination of agents' initial opinions where $|s_i|$ denotes the *social influence* (proper) of agent $i = 1, \dots, n$ and $\operatorname{sgn}(s_i) \in \{\pm 1\}$ denotes group membership of $i \in [n]$ — we then have

$$\mathbf{s}^\intercal \mathbf{b}(0) = a(\infty) = \mathbf{s}^\intercal (\mathbf{A}\mathbf{b}(0))$$

since $a(\infty)$ is the same whether we start from $\mathbf{b}(0)$ or $\mathbf{Ab}(0)$. But since this must hold for any $\mathbf{b}(0)$, we have

$$\mathbf{s}^\intercal = \mathbf{s}^\intercal \mathbf{A},$$

or, equivalently,

$$\mathbf{s} = \mathbf{A}^\intercal \mathbf{s},$$

such that $\mathbf{s}$ is simply an eigenvector of matrix $\mathbf{A}^\intercal$ (corresponding to the eigenvalue $\lambda = 1$). In other words, in order to compute $\mathbf{p}(\infty)$ given $\mathbf{b}(0)$, it might be possible to compute the eigenvector $\mathbf{s}$ of $\mathbf{A}^\intercal$ corresponding to $\lambda = 1$ and then apply $\mathbf{s}$ to $\mathbf{b}(0)$ in the form $\mathbf{s}^\intercal\mathbf{b}(0)$ to derive one limiting viewpoint and in the form $\bar{\mathbf{s}}^\intercal\mathbf{b}(0)$ to derive the other. Hence, since social influence is given by an eigenvector of $\mathbf{A}^\intercal$, social influence of agents is measured by *eigenvector centrality* (cf. Bonacich, 1972), in the current setting, in a very similar way as in the original DeGroot opinion dynamics model.

We give the following detailed example for Proposition 1.6.5 and the subsequent remark.

Example 1.6.2. Let $n = 2$ and let

$$\mathbf{W} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & D \\ D & F \end{pmatrix},$$

Let D be soft opposition on $S = [\alpha, \beta]$. We first note that $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite, e.g., with $\mathcal{N}_1 = \{1\}$, $\mathcal{N}_2 = \{2\}$. Moreover, the affine-linear representation of $\mathbf{W} \circ \mathbf{F}$ is given by

$$\mathbf{A} = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \mathbf{d} = (\alpha + \beta) \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \end{pmatrix};$$

note that we assume that $\alpha = -\beta$ such that $\mathbf{d} = \mathbf{0}$. The eigenvalues of matrix $\mathbf{A}$ are determined as the roots of the characteristic polynomial $\chi(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}_n)$ where $\mathbf{I}_n$ is the $n \times n$ identity matrix. We have,

$$\chi(\lambda) = (\frac{3}{4} - \lambda)(\frac{1}{2} - \lambda) - \frac{1}{8} = \frac{1}{4} - \frac{5}{4}\lambda + \lambda^2 = (\lambda - 1)(\lambda - \frac{1}{4}).$$

Hence, $\lambda = 1$ and $\lambda = \frac{1}{4}$ are the two eigenvalues of $\mathbf{A}$. Thus, $\lambda = 1$ is the only eigenvalue on the unit circle and the algebraic multiplicity of $\lambda = 1$ is 1 since the exponent of $(\lambda - 1)$ in $\chi(\lambda)$ is 1. Hence, $\lim_{t\to\infty}(\mathbf{W}\circ\mathbf{F})^t\mathbf{b}(0)$ is a polarization for any $\mathbf{b}(0) \in S^n$, by Proposition 1.6.5. To determine the influence vector $\mathbf{s}$, we need to compute the normalized eigenvector of $\mathbf{A}^\intercal$ corresponding to $\lambda = 1$. It is easy to see that $\mathbf{s} = (\frac{2}{3}, -\frac{1}{3})^\intercal$ is the searched for normalized unit vector since $\mathbf{A}^\intercal\mathbf{s} = \mathbf{s}$. Now, one sees how $\mathbf{s}$ captures social influence: agent 1 is apparently more influential, since he weighs himself higher, than agent 2 (3/4 self-weight vs. 1/2); accordingly, his influence weight in $\mathbf{s}$ is larger in absolute value, $\frac{2}{3} > \frac{1}{3}$. Then, limiting opinions are simply given by,

$$\begin{aligned} a(\infty) &= \frac{2}{3}b_1(0) - \frac{1}{3}b_2(0), \\ b(\infty) &= -\frac{2}{3}b_1(0) + \frac{1}{3}b_2(0). \end{aligned}$$

For instance, if agents start with the consenus $\mathbf{b}(0) = (\frac{1}{2}, \frac{1}{2})^\intercal$, they will end up at the polarization $\mathbf{p}(\infty) = (\frac{1}{6}, -\frac{1}{6})^\intercal$. In Figure 1.10, we illustrate opinion dynamics for this setup and for a random opposition bipartite multigraph.

Next, we show that 'opposition bipartiteness' is a very delicate condition in the continuous model that, if slightly violated, does not lead agents to a polarization but, rather, to a neutral consensus (or to divergence), in the situation when D is soft opposition. To this end, we define the notion of 'opposition (anti-)equivalent' agents.

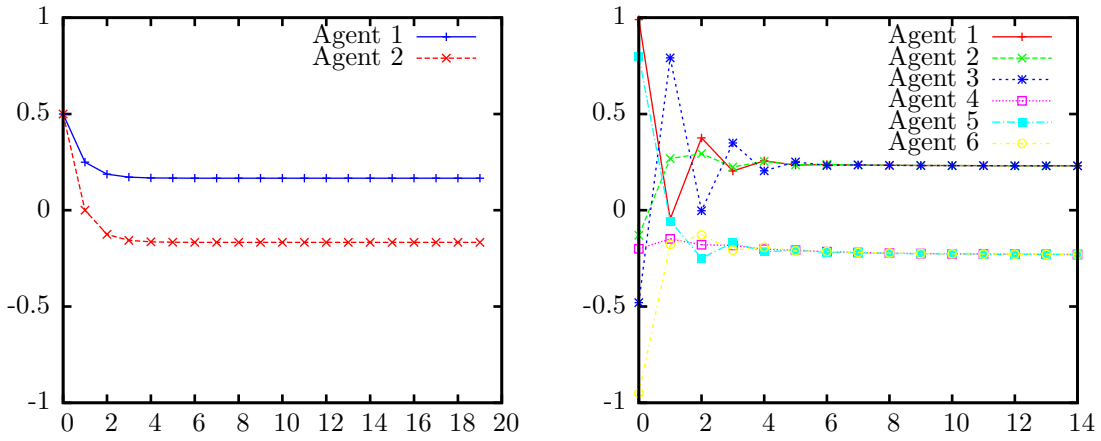


Figure 1.10: Left: $\mathbf{b}(t)$, for $t = 0, \ldots, 20$, for the process discussed in Example 1.6.2. Right: $\mathbf{b}(t)$, for $t = 0, \ldots, 15$, for a random 6×6 matrix $\mathbf{W}$ and an opposition bipartite network $\mathbf{F}$ (such that $\mathbf{W} \circ \mathbf{F}$ satisfies the conditions of Proposition 1.6.5) and random initial opinions $\mathbf{b}(0) \in [-1, 1]^6$.

Definition 1.6.1. We call two agents $i_0, i_1 \in [n]$ *opposition equivalent* if $F_{i_0 j} = F_{i_1 j}$ for all $j \in [n]$.

We call two agents $i_0, i_1 \in [n]$ *opposition anti-equivalent* if $F_{i_0 j} = \neg F_{i_1 j}$ for all $j \in [n]$, where we let $\neg D = F$ and $\neg F = D$.

Note that these two notions are closely related, e.g., to opposition bipartite networks. Namely, in the latter situation, there exist two groups of agents $\mathcal{N}_1$ and $\mathcal{N}_2$ such that for all $i_0, i_1 \in \mathcal{N}_1$, i_0 and i_1 are opposition equivalent (and follow each other) while for all $i_0 \in \mathcal{N}_1$ and $i_1 \in \mathcal{N}_2$, i_0 and i_1 are opposition anti-equivalent.

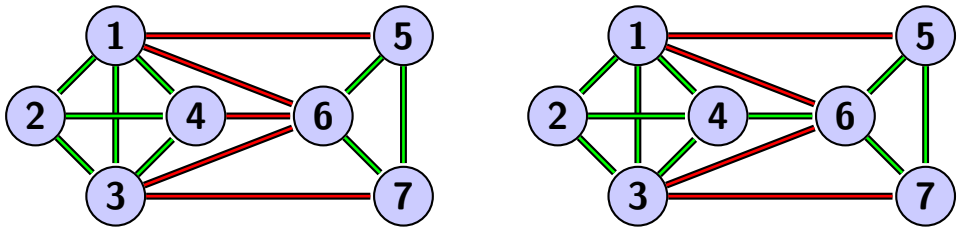


Figure 1.11: Balanced and unbalanced networks. The left network is opposition bipartite (balanced) while the right is not. In particular, agents 3 and 4 have mutual 'friends' (agents 1, 2) while agent 6 is in 3's outgroup and 4's ingroup.

Proposition 1.6.6. Let D be soft opposition on $S = [\alpha, \beta]$. Let $\mathbf{W} > 0$, entry-wise. Let $(\mathbf{A}, \mathbf{d})$ be the representation of $\mathbf{W} \circ \mathbf{F}$. Assume that $\mathbf{A}$ has no complex eigenvalues (on the unit circle). Then, if there exist agents i_0 and i_1 such that i_0 and i_1 are neither opposition equivalent nor opposition anti-equivalent, then $\mathbf{W} \circ \mathbf{F}$ induces the consensus $\frac{\alpha+\beta}{2}$, for all initial opinion vectors $\mathbf{b}(0)$.

Proof. By Proposition 1.6.3, $\rho(\mathbf{A}) \leq 1$. We want to exclude the case $\rho(\mathbf{A}) = 1$. This means we want to exclude that $\pm 1 \in \sigma(\mathbf{A})$ since $\mathbf{A}$ has no complex eigenvalues on the unit circle by assumption. For all agents $i = 1, \ldots, n$ and any vector $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_\infty = \max_{i \in [n]} |x_i|$, it holds that

$$|A_{i1}x_1 + \cdots + A_{in}x_n| \leq |A_{i1}|\,|x_1| + \cdots + |A_{in}|\,|x_n| < (|A_{i1}| + \cdots + |A_{in}|)\,\|\mathbf{x}\|_\infty = \|\mathbf{x}\|_\infty$$

unless $|x_1| = \cdots = |x_n|$ (since $W_{ij} = |A_{ij}|$ is strictly positive by assumption, for all $i, j \in [n]$), in which case equality may hold instead of $<$. Thus, if it does not hold that $|x_1| = \cdots = |x_n|$, then $\mathbf{Ax} = \mathbf{x}$ or $\mathbf{Ax} = -\mathbf{x}$ are impossible since both imply that $\|\mathbf{Ax}\|_\infty = \|\mathbf{x}\|_\infty$, contradicting $\|\mathbf{Ax}\|_\infty < \|\mathbf{x}\|_\infty$.

So, consider $\mathbf{x}$ with $|x_1| = \cdots = |x_n|$. Without loss of generality, we may assume that $\|\mathbf{x}\|_\infty = 1$ such that $\mathbf{x}$ is a vector with entries 1, either with positive or negative sign. In this case, for agent i_0,

$$\sum_{j=1}^{n} A_{i_0 j} \underbrace{x_j}_{\in\{\pm 1\}} = x_{i_0} \in \{\pm 1\}$$

implies that, by the structure of matrix $\mathbf{A}$ (row-stochasticity of $\mathbf{W}$ and $A_{ij} \neq 0$ for all $i, j \in [n]$), all summands $A_{i_0 j} x_j$ on the left-hand side of the last equation must have the same sign, either positive or negative. But then, for agent i_1, it cannot be that $\sum_{j=1}^{n} A_{i_1 j} \underbrace{x_j}_{\in\{\pm 1\}} \in \{\pm 1\}$, since some summands of the left-hand side of this equation must have opposite signs since i_0 and i_1 are neither opposition equivalent nor opposition anti-equivalent. Thus, under the assumptions of the proposition, $\mathbf{Ax} = \mathbf{x}$ or $\mathbf{Ax} = -\mathbf{x}$ cannot hold for any $\mathbf{x} \in \mathbb{R}^n$ and, thus, $\mathbf{A}$ has no eigenvalues ± 1, whence $\rho(\mathbf{A}) < 1$ and $\mathbf{W} \circ \mathbf{F}$ is a contraction mapping. □

Remark 1.6.6. Proposition 1.6.6 gives less abstract conditions for convergence to a neutral consensus than we have outlined before and which were based on the size of the spectral radius of matrix $\mathbf{A}$ in the affine-linear representation of $\mathbf{W} \circ \mathbf{F}$. Namely, in the situation of the proposition — e.g., with all weights W_{ij} strictly positive and no complex eigenvalues on the unit circle — a spectral radius of $\mathbf{A}$ strictly smaller than 1 is implied by the condition that two agents i_0 and i_1 are 'misaligned' in the sense that there are two distinct agents A and B such that i_0 and i_1 have the same relation to A but inverse relationships to B. For example, A might both be in i_0's and i_1's ingroup, while B is in i_0's ingroup and in i_1's outgroup; consider Figure 1.11 for an example. It is clear that such a configuration causes the corresponding multigraph to be 'unbalanced' because of 'contradicting' friendship/animosity relationships since, in the example made, i_0's and i_1's ingroups are overlapping but not identical. Accordingly, agents do not polarize but converge to a neutral consensus. Thus, neutrality may be perceived of as resulting from a lack of balance which would otherwise induce polarizations, in this context.

In the following beautiful theorem, Theorem 1.6.1, we generalize our above observation to the case when the multigraph underlying $\mathbf{W} \circ \mathbf{F}$ is strongly connected and aperiodic, rather than fully connected. The theorem, together with its generalization in Theorem 1.6.2, gives an *exhaustive classification of results* on convergence of $(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0)$ in case $\mathbf{W} \circ \mathbf{F}$ is strongly connected (and aperiodic); as restraining conditions, we merely assume that $W_{ii} = 0$ and that $A_{ij} = A_{ji}$, that is, *intensity and kind of relationship are symmetric*. The more general cases are left for ongoing research. Our theorem is based, to a significant degree, on the corresponding results given in Altafini (2013), who analyzes are very similar situation as we, but considers the (time-)continuous process $\dot{\mathbf{x}} = -\mathbf{L}\mathbf{x}$, rather than the (time-)discrete model $\mathbf{b}(t+1) = (\mathbf{W} \circ \mathbf{F})\mathbf{b}(t)$, as we investigate.

As to the results, the theorem shows that agents polarize *if and only if* the operator $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite; that agents diverge *if and only if* the operator $\mathbf{W} \circ \mathbf{F}$ is anti-opposition bipartite; and, finally, that agents reach a neutral consensus *if and only if* none of the former two conditions hold.

We first state the following simple lemma.

Lemma 1.6.1. Let $\mathbf{W} \circ \mathbf{F}$ be an arbitrary multigraph. Then, $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite if and only if $\mathbf{W} \circ \bar{\mathbf{F}}$ is anti-opposition bipartite, where $\bar{\mathbf{F}}$ is the matrix with entries $\bar{F}_{ij} = \neg F_{ij}$.

Proof. See Figure 1.7, in Section 1.5, for a graphical proof. □

If D is soft opposition on $S = [-\beta, \beta]$, let $(\mathbf{A}, \mathbf{0})$ be the representation of $\mathbf{W} \circ \mathbf{F}$. Then, the lemma specializes to the statement that $(\mathbf{A}, \mathbf{0})$ is opposition bipartite if and only if $(-\mathbf{A}, \mathbf{0})$ is anti-opposition bipartite.

Theorem 1.6.1. Let D be soft opposition on $S = [-\beta, \beta]$ for some $\beta > 0$. Let $\mathbf{W} \circ \mathbf{F}$ be an arbitrary operator such that $W_{ii} = 0$ for all $i \in [n]$. Let $(\mathbf{A}, \mathbf{0})$ be the affine-linear representation of $\mathbf{W} \circ \mathbf{F}$ and assume, moreover, that $\mathbf{A}$ is symmetric. Assume that $\mathbf{W} \circ \mathbf{F}$ is strongly connected (since $\mathbf{A}$ is symmetric, we might also simply say 'connected') and aperiodic. Then:

(i) $\mathbf{W} \circ \mathbf{F}$ induces a polarization if and only if $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite.

(ii) $\mathbf{W} \circ \mathbf{F}$ diverges if and only if $\mathbf{W} \circ \mathbf{F}$ is anti-opposition bipartite.

(iii) $\mathbf{W} \circ \mathbf{F}$ induces a neutral consensus if and only if $\mathbf{W} \circ \mathbf{F}$ is neither opposition bipartite nor anti-opposition bipartite.

Proof. The theorem follows from the following facts. (i) If $\mathbf{W} \circ \mathbf{F}$ induces a polarization, then, necessarily, $1 \in \sigma(\mathbf{A})$. But, (1) $1 \in \sigma(\mathbf{A}) \iff \mathbf{W} \circ \mathbf{F}$ is opposition bipartite. Conversely, let $\mathbf{W} \circ \mathbf{F}$ be opposition bipartite. Then, (2) $|\mathbf{A}|$ — the matrix with entries $|A_{ij}|$ — and $\mathbf{A}$ are *isospectral*, that is, they have the same eigenvalues and with the same associated multiplicities. Now, (3) a strongly connected and aperiodic row-stochastic matrix $|\mathbf{A}|$ has exactly one eigenvalue on the unit circle, $\lambda = 1$, with algebraic and geometric multiplicity of 1. Therefore, $\mathbf{A}$ has exactly one eigenvalue on the unit circle, $\lambda = 1$, with algebraic and geometric multiplicity of 1 and, consequently, converges by Theorem 1.4.4. Moreover, since each polarization vector $\mathbf{x}$ with $x_i = 1$ if $i \in \mathcal{N}_1$ and $x_i = -1$ if $i \in \mathcal{N}_2$ satisfies $\mathbf{A}\mathbf{x} = (\mathbf{W} \circ \mathbf{F})\mathbf{x} = \mathbf{x}$ when $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite with partition $(\mathcal{N}_1, \mathcal{N}_2)$, $\mathbf{W} \circ \mathbf{F}$ induces a polarization.

Part (ii) follows from the fact that $1 \in \sigma(\mathbf{A}) \iff \mathbf{W} \circ \mathbf{F}$ is opposition bipartite and the fact that $\mathbf{W} \circ \mathbf{F}$ with representation $\mathbf{A}$ is opposition bipartite if and only if $-\mathbf{A}$ is anti-opposition bipartite by Lemma 1.6.1. Thus, $-1 \in \sigma(\mathbf{A}) \iff \mathbf{W} \circ \mathbf{F}$ is anti-opposition bipartite, whence $\mathbf{A}$ diverges by Theorem 1.4.4.

Finally, part (iii) follows since if $\mathbf{W} \circ \mathbf{F}$ is neither opposition bipartite nor anti-opposition bipartite, then, by our above reasonings, $\pm 1 \notin \sigma(\mathbf{A})$, and since $\mathbf{A}$ is symmetric, $\mathbf{A}$ has no complex eigenvalues, whence $\rho(\mathbf{A}) < 1$ and, thus, $\mathbf{W} \circ \mathbf{F}$ is a contraction mapping. Consequently, $\mathbf{W} \circ \mathbf{F}$ induces the unique neutral consensus $(c, \dots, c)$ by Banach's fixed point theorem, Theorem 1.4.2, where $c = 0$ due to the choice of D.

Now, fact (3) is a classical theorem for row-stochastic matrices, which is, e.g., based on the famous Perron-Frobenius theorem; in our context, it is given by combining Theorems 1.4.1 and 1.4.4, for example. We prove facts (2) and (3) in the appendix, Lemmas 1.A.1, 1.A.2, and 1.A.3, respectively.

□

It is a well-known fact that graphs can be partitioned into strongly connected and closed groups of nodes and the (possibly empty) 'rest of the world' (cf., e.g., Jackson, 2009; Buechel, Hellmann, and Klößner, 2013). Hence, in the setup of Theorem 1.6.1, $\mathbf{W} \circ \mathbf{F}$ can be partitioned into precisely such a structuring. Then, if the underlying graphs corresponding to each strongly connected group in the partition satisfy aperiodicity, Theorem 1.6.1 may be applied to determine limits of $\mathbf{W} \circ \mathbf{F}$.

Example 1.6.3. Let $n = 12$ and let $\mathbf{W} \circ \mathbf{F}$ be such that $\mathbf{A}$ has the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_3 & 0 \\ & \mathbf{r}^\intercal & & \end{pmatrix},$$

where

$$\mathbf{C}_1 = \frac{1}{2}\begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix}, \quad \mathbf{C}_2 = \frac{1}{3}\begin{pmatrix} 0 & 1 & -1 & 1 \\ 1 & 0 & 1 & -1 \\ -1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \end{pmatrix}, \quad \mathbf{C}_3 = \frac{1}{3}\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & 1 & -1 & 0 \end{pmatrix},$$

and $\mathbf{r}$ is the vector with $r_3 = -0.6$, $r_9 = 0.4$ and $r_j = 0$ for all other $j \in [n]$. The multigraph corresponding to $\mathbf{W} \circ \mathbf{F}$ is shown in Figure 1.12. From this we see that, within all closed and strongly connected groups, the underlying graphs are aperiodic such that Theorem 1.6.1 may be applied to the strongly connected groups. Hence, we know that, no matter the initial opinions, agents $\{1, 2, 3\}$ will polarize since their underlying multigraph is opposition bipartite — we have, e.g., $\mathcal{N}_1 = \{1\}$ and $\mathcal{N}_2 = \{2, 3\}$. Groups $\{4, 5, 6, 7\}$ and $\{8, 9, 10, 11\}$ will either diverge or reach a neutral consensus. Since the group $\{4, 5, 6, 7\}$ is anti-opposition bipartite, it will, in fact, diverge and since the group $\{8, 9, 10, 11\}$ is, in fact, neither

opposition bipartite nor anti-opposition bipartite, it will reach a neutral consensus. The 'rest of the world', agent 12, will attain a limit opinion that is a linear combination of the opinions of agents $\{1, 2, 3\}$ and $\{8, 9, 10, 11\}$ — the latter attain a neutral consensus. We plot a sample evolution of the corresponding

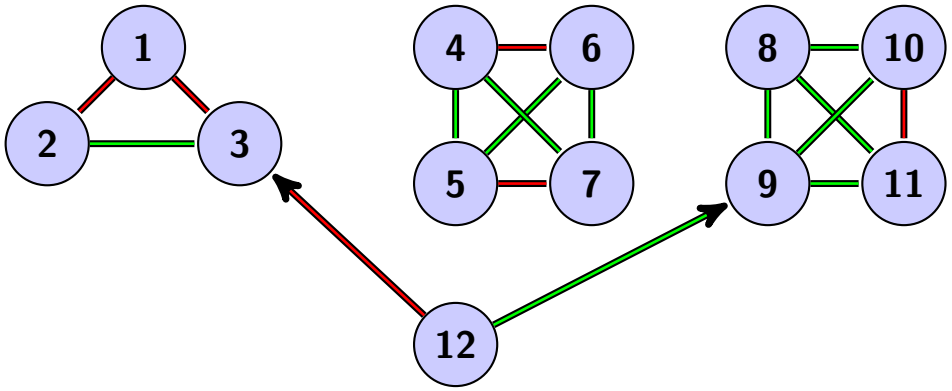


Figure 1.12: Description in text. As usual, we omit links corresponding to weights from the multigraphs, simply indicating links corresponding to opposition/following behavior. For convenience, we color following in green and deviating in red. All links are undirected unless indicated by a respective arrow.

opinion dynamics in Figure 1.13.

Now, we want to characterize limit behavior of a strongly connected $\mathbf{W} \circ \mathbf{F}$ in case $\mathbf{W} \circ \mathbf{F}$ is *periodic*, rather than aperiodic. For this, we need the insight that the concepts of 'opposition-bipartiteness' and 'anti-opposition bipartite' collapse if and only if $\mathbf{W} \circ \mathbf{F}$ is periodic. Figure 1.14 illustrates.

Lemma 1.6.2. A strongly connected multigraph $\mathbf{W} \circ \mathbf{F}$ is periodic if and only if the concepts of opposition-bipartiteness and anti-opposition bipartiteness coincide (that is, $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite if and only if $\mathbf{W} \circ \mathbf{F}$ is anti-opposition bipartite).

Proof. If $\mathbf{W} \circ \mathbf{F}$ is aperiodic, $\mathbf{W} \circ \mathbf{F}$ cannot be both opposition bipartite and anti-opposition bipartite because this would contradict Theorem 1.6.1, parts (i) and (ii), according to which the two concepts are distinct in this case (strongly connected and aperiodic opposition bipartite multigraphs have eigenvalues $\lambda = 1$ on the unit circle and no other, there, while anti-opposition bipartite multigraphs have eigenvalues $\lambda = -1$ on the unit circle and no other, there).

Conversely, assume that $\mathbf{W} \circ \mathbf{F}$ is periodic and assume that $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite with partition $(\mathcal{N}_1, \mathcal{N}_2)$. Then, the crucial aspect to note is that there can be no triangles in $\mathbf{W} \circ \mathbf{F}$, that is, nodes $i, j, k \in [n]$ such that $W_{ij} > 0$, $W_{jk} > 0$ and $W_{ik} > 0$ for otherwise — note that $\mathbf{W} \circ \mathbf{F}$ is symmetric — there would be a simple cycle of length 3 in $\mathbf{W} \circ \mathbf{F}$, whence the greatest common divisor of all simple cycles would be 1 (a symmetric connected graph trivially has simple cycles of length 2), contradicting that $\mathbf{W} \circ \mathbf{F}$ is periodic.

Hence, construct an anti-opposition bipartite partition of $\mathbf{W} \circ \mathbf{F}$ from $(\mathcal{N}_1, \mathcal{N}_2)$ as follows. Let $\tilde{\mathcal{N}}_1$ and $\tilde{\mathcal{N}}_2$ be empty sets. Take $a \in \mathcal{N}_1$, put it in $\tilde{\mathcal{N}}_1$, together with all its 'enemies' and put the 'friends' of a in $\tilde{\mathcal{N}}_2$. Consider any friend c of any friend b of a (other than a). Clearly, since there are no triangles in $\mathbf{W} \circ \mathbf{F}$, a and c are in no friendship relation. Hence, put c in $\tilde{\mathcal{N}}_1$ as well (c might have negative relationships with the other nodes in $\tilde{\mathcal{N}}_1$, which does not violate the conditions of anti-opposition bipartiteness). Continue until all nodes are covered with c taking the role of a at the beginning of the process and note that no condition of anti-bipartiteness is ever violated during the process.

By an analogue procedure, anti-opposition bipartiteness may be converted into opposition bipartiteness in the case of strongly connected periodic multigraphs. □

With Theorem 1.6.1 and the lemma, we obtain the following corollary, which takes care of the periodicity case of $\mathbf{W} \circ \mathbf{F}$.

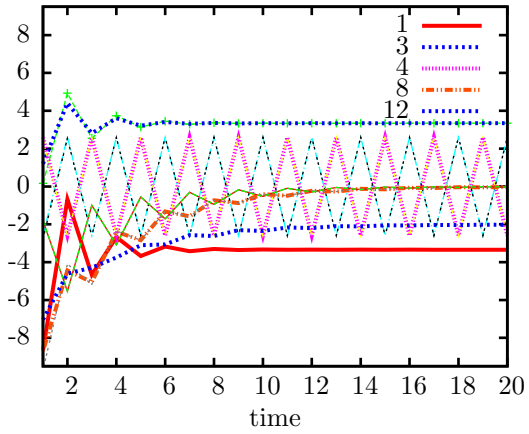


Figure 1.13: Sample opinion dynamics for Example 1.6.3. We see polarization, neutrality, and divergence, as well as an agent — agent 12, the 'rest of the world' — who holds an opinion that is a linear combination of the opinions of member 1 of group $\{1, 2, 3\}$ and of member 9 of group $\{8, 9, 10, 11\}$. Selected agents highlighted.

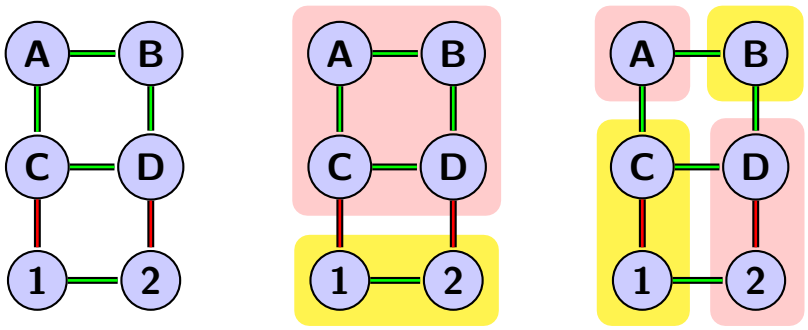


Figure 1.14: 'Re-arranging' an opposition bipartite partitioning of a strongly connected periodic multigraph to obtain an anti-opposition bipartite partitioning, Lemma 1.6.2 at work. Left: The periodic multigraph. Middle: The original opposition-bipartite partitioning $(\mathcal{N}_1, \mathcal{N}_2)$. Right: Choosing the sets $\tilde{\mathcal{N}}_1$ and $\tilde{\mathcal{N}}_2$ of the anti-opposition bipartite partitioning.

Corollary 1.6.2. Let D be soft opposition on $S = [-\beta, \beta]$ for some $\beta > 0$. Let $\mathbf{W} \circ \mathbf{F}$ be an arbitrary operator such that $W_{ii} = 0$ for all $i \in [n]$. Let $(\mathbf{A}, \mathbf{0})$ be the affine-linear representation of $\mathbf{W} \circ \mathbf{F}$ and assume, moreover, that $\mathbf{A}$ is symmetric. Assume that $\mathbf{W} \circ \mathbf{F}$ is strongly connected (or, since $\mathbf{A}$ is symmetric, simply 'connected') and periodic. Then:

(i) $\mathbf{W} \circ \mathbf{F}$ diverges if and only if $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite.

(ii) $\mathbf{W} \circ \mathbf{F}$ induces a neutral consensus if and only if $\mathbf{W} \circ \mathbf{F}$ is not opposition bipartite.

Proof. Putting all results together, we obtain the following equivalences for symmetric, strongly connected and periodic multigraphs $\mathbf{W} \circ \mathbf{F}$ with representation $\mathbf{A}$:

$$\mathbf{W} \circ \mathbf{F} \text{ is not OBIP} \iff \pm 1 \notin \sigma(\mathbf{A}) \iff \rho(\mathbf{A}) < 1 \iff \lim_{t \to \infty} (\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0) = \mathbf{0} \ \forall \mathbf{b}(0) \in S^n,$$

where we let OBIP abbreviate 'opposition bipartite'. The equivalences prove the corollary. The first equivalence follows from the fact that $\mathbf{W} \circ \mathbf{F}$ is OBIP if and only if $1 \in \sigma(\mathbf{A})$ and $\mathbf{W} \circ \mathbf{F}$ is anti-OBIP if and only if $-1 \in \sigma(\mathbf{A})$ for strongly connected multigraphs. Hence, by Lemma 1.6.2, $\pm 1 \in \sigma(\mathbf{A})$ if and only if $\mathbf{W} \circ \mathbf{F}$ is OBIP for strongly connected and periodic multigraphs. The second equivalence follows since $\mathbf{A}$ is symmetric, whence it has no other potential eigenvalues than ± 1 on the unit circle. □

Therefore, since periodicity and aperiodicity are mutually exclusive properties, we have fully characterized limit properties of strongly connected $\mathbf{W} \circ \mathbf{F}$ in the case of soft opposition D on $S = [-\beta, \beta]$ and where we assume that the linear representation $\mathbf{A}$ of $\mathbf{W} \circ \mathbf{F}$ satisfies symmetry and $A_{ii} = 0$. We summarize our findings in the following theorem.

Theorem 1.6.2. Let D be soft opposition on $S = [-\beta, \beta]$ for some $\beta > 0$. Let $\mathbf{W} \circ \mathbf{F}$ be an arbitrary operator such that $W_{ii} = 0$ for all $i \in [n]$. Let $(\mathbf{A}, \mathbf{0})$ be the affine-linear representation of $\mathbf{W} \circ \mathbf{F}$ and assume, moreover, that $\mathbf{A}$ is symmetric. Assume that $\mathbf{W} \circ \mathbf{F}$ is strongly connected. Then:

(i) $\mathbf{W} \circ \mathbf{F}$ diverges if and only if $\mathbf{W} \circ \mathbf{F}$ is anti-opposition bipartite.

(ii) $\mathbf{W} \circ \mathbf{F}$ induces a polarization if and only if $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite and aperiodic.

(iii) $\mathbf{W} \circ \mathbf{F}$ induces a neutral consensus if and only if it is neither opposition bipartite nor anti-opposition bipartite.

Remark 1.6.7. The fact that polarization requires 'exact' balance (opposition bipartiteness) and admits not a 'grain of unbalancedness', as stated in Theorem 1.6.1, may appear odd since one might expect, in reality, small perturbations to balance (e.g., small-scale intra-group antagonisms or individual friendships among enemies) to be the rule, rather than the exception, particularly in large enough systems.[31] We note that this result is, however, to a large part, due to the continuous opinion spectrum and the averaging updating process that we have considered in this section. If the reader thinks that reality is better perceived of as being discrete, with weighted majority voting a more plausible opinion updating mechanism, then we note that, as we have shown, the discrete model *is* in fact robust against small perturbations such that polarizing viewpoints can be Nash equilibria in this case even if the underlying multigraphs exhibit (marginal) unbalancedness. In addition, we note that our analysis thus far has also depended on the specification of weight sum requirements, as we illustrate in the following.

1.6.2 The requirement $W_{i,\mathcal{F}_i} = 1 + W_{i,\mathcal{O}_i}$

We have seen that, in the continuous model, agents cannot reach a non-neutral consensus, under opposition. This is unlike in the discrete case, where the same conclusion requires a certain 'weight mass condition', namely, that at least one agent's outgroup is decisive for him. One way to interpret this, consistent across both models, is to say that in the continuous model, under the row-stochasticity assumption, $W_{i,A} > 0$ already means that group $A \subseteq [n]$ is decisive for agent i, rather than $W_{i,A} > \frac{1}{2}$ as in the discrete model. In this interpretation, one way to 'address' the issue of reaching non-neutral consensus opinions is to either restrict the weight mass assigned to opposed agents (e.g., demanding that $W_{i,\mathcal{O}_i} \leq \frac{1}{2}$ in the discrete model) or to enlarge the weight mass assigned to trusted agents. In the continuous model, we would be forced to consider the latter option since requiring that $W_{i,\mathcal{O}_i} \leq 0$ would be tantamount to resorting to the standard DeGroot model, without opposition.

In the following, we sketch one possibility for agents to reach non-neutral consensus opinions in the continuous model, even under the presence of opposition. We do so for a very special but important instance of opposition function D, namely, soft opposition on $\mathbb{R}$, that is, $D(x) = -x$ (see below on why we need to extend S to $\mathbb{R}$, in this situation). In this setup, a weight mass requirement that allows non-neutral consensus formation can be read off from the proof of Proposition 1.6.1, which illustrates what 'goes wrong' under the row-stochasticity assumption. Namely, assuming that $D(c) \neq c$ (c is a non-neutral opinion), in order for $\mathbf{c} = (c, \ldots, c)^\intercal$ to be a fixed-point of $\mathbf{W} \circ \mathbf{F}$ it must hold that:

$$c = \sum_{j \in \mathcal{F}_i} W_{ij} c + \sum_{j \in \mathcal{O}_i} W_{ij} D(c) = \sum_{j \in \mathcal{F}_i} W_{ij} c - \sum_{j \in \mathcal{O}_i} W_{ij} c = c\Big(\sum_{j \in \mathcal{F}_i} W_{ij} - \sum_{j \in \mathcal{O}_i} W_{ij}\Big)$$

or, equivalently,

$$1 = \sum_{j \in \mathcal{F}_i} W_{ij} - \sum_{j \in \mathcal{O}_i} W_{ij} = W_{i,\mathcal{F}_i} - W_{i,\mathcal{O}_i}. \tag{1.6.3}$$

[31] Facchetti, Iacono, and Altafini (2011) empirically demonstrate, however, that currently available on-line social networks are indeed 'extremely balanced'.

Now, if $W_{i,\mathcal{O}_i} > 0$, this requirement can never be satisfied if we additionally require row-stochasticity of $\mathbf{W}$, which means that $1 = W_{i,\mathcal{F}_i} + W_{i,\mathcal{O}_i}$. If, instead of row-stochasticity, we demanded the following weight sum restriction,

$$W_{i,\mathcal{F}_i} = 1 + W_{i,\mathcal{O}_i}, \tag{1.6.4}$$

then (1.6.3) would trivially be satisfied and, consequently, even non-neutral opinions could be consensus outcomes of opinion updating process (1.3.3). Comparing this with the requirement of row-stochasticity, which reads, in other form,

$$W_{i,\mathcal{F}_i} = 1 - W_{i,\mathcal{O}_i}, \tag{1.6.5}$$

we find that

- under the row-stochasticity requirement (1.6.5), opposition 'takes away' weight mass from followed agents, while
- under requirement (1.6.4), opposition 'increases' the weight mass that must be assigned to followed agents. In other words, under requirement (1.6.4), the more an agent opposes her outgroup, the more is she required to follow, or trust, her ingroup in order for her to have her 'trust balance' cleared. In this sense, it is clear that (1.6.4) facilitates attaining a non-neutral consenus, compared with requirement (1.6.5). In addition,
- under requirement (1.6.4), all agents $i = 1, \ldots, n$ are 'generally trusting', that is, they assign more weight mass to their ingroup than to their outgroup. Opposition is less strong an incentive than is following one's ingroup.

From a mathematical perspective, requirement (1.6.4) is more problematic because even if S is a convex set, a weighted combination of elements of S where weights satisfy (1.6.4) need not be an element of S, since convex sets are guaranteed to be closed only under *convex combinations* of their elements, that is, where weights are taken from the unit simplex. Thus, to make this model mathematically well-defined, we need to think of S as the whole real line $\mathbb{R}$.

From an economic perspective, of our three justifications of DeGroot learning given in Section 1.3, weight requirement (1.6.4) fails two, namely, the justification based on boundedly rational Bayesian learning and the justification relating to aggregation theory because, in both instances, unit simplex weights are assumed. It does not fail the justification based on myopic best-response updating, since if we define agent i's utility on opinion vector $\mathbf{b}$ as

$$u_i(\mathbf{b}) = -\sum_{j\in\mathcal{F}_i} W_{ij}(b_i - Wb_j)^2 - \sum_{j\in\mathcal{O}_i} W_{ij}(b_i - WD(b_j))^2, \tag{1.6.6}$$

where $W = \sum_{j=1}^{n} W_{ij}$, then myopic best-response updating retrieves our learning rule (1.3.1) with weight sum restriction (1.6.4). One interpretation that we may give utility structure (1.6.6) is that agents have disutility from making opinion choices different from (positively) scaled opinion choices of agents they follow and that agents have disutility from not deviating from (positively) scaled opinion choices of agents they oppose.

In the sequel, we very briefly analyze the variant of the DeGroot model just introduced, thereby showing that this model allows agents, in a number of cases, to attain non-neutral consensus vectors as limits of the DeGroot opinion updating process.

Sufficient conditions for convergence to consensus

Proposition 1.6.7. Let D be soft opposition on $S = \mathbb{R}$. Then $\mathbf{W} \circ \mathbf{F}$ is (affine-)linear and let $(\mathbf{A}, \mathbf{0})$ be its representation. If $\mathbf{W}$ satisfies (1.6.4), then $1 \in \sigma(\mathbf{A})$. Moreover, if $\rho(\mathbf{A}) = 1$ and $\lambda = 1$ is the only eigenvalue of $\mathbf{A}$ on the unit circle and if $\lambda = 1$ has algebraic multiplicity of 1, then $\lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t\mathbf{b}(0)$ is a consensus for all $\mathbf{b}(0) \in S^n$.

Proof. As in (1.6.2), $\mathbf{A}$ is the matrix with $A_{ij} = W_{ij}$ if $F_{ij} = F$ and $A_{ij} = -W_{ij}$ if $F_{ij} = D$ for all $i, j \in [n]$. Moreover, we note that, under weight sum restriction (1.6.4),

$$\sum_{j=1}^{n} A_{ij} = \sum_{j \in \mathcal{F}_i} A_{ij} + \sum_{j \in \mathcal{O}_i} A_{ij} = \sum_{j \in \mathcal{F}_i} W_{ij} - \sum_{j \in \mathcal{O}_i} W_{ij} = W_{i,\mathcal{F}_i} - W_{i,\mathcal{O}_i} = 1$$

for all $i = 1, \dots, n$. In other words, the row sum of each row i of $\mathbf{A}$ is 1. But then, $\mathbf{Ax} = \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$ such that $x_1 = \ldots = x_n$. Thus, $1 \in \sigma(\mathbf{A})$. As in the proof of Proposition 1.6.5, algebraic multiplicity of $\lambda = 1$ of 1 and $\lambda = 1$ being the only eigenvalue on the unit circle, together with $\rho(\mathbf{A}) = 1$, imply that $\mathbf{W} \circ \mathbf{F}$ induces a consensus for any $\mathbf{b}(0) \in S^n$, by Theorem 1.4.4. □

Remark 1.6.8. In the proof, we have seen that weight sum restriction (1.6.4) implies that $\sum_{j=1}^{n} A_{ij} = 1$ for all $i = 1, \dots, n$. In contrast, under weight sum restriction (1.6.5), as discussed in the previous subsection, rows of $\mathbf{A}$ satisfy

$$\sum_{j=1}^{n} |A_{ij}| = 1,$$

for all $i = 1, \dots, n$, as can easily be verified.

Remark 1.6.9. Analogously as in Remark 1.6.5, if the assumptions of Proposition 1.6.7 hold, limiting consensus is given by

$$b(\infty) = \mathbf{s}^\intercal \mathbf{b}(0) = \sum_{i=1}^{n} s_i b_i(0),$$

where $\mathbf{s}$ is the eigenvector of $\mathbf{A}^\intercal$ corresponding to $\lambda = 1$. The vector $\mathbf{s}$ encodes social influence of the agents $i = 1, \dots, n$.

Example 1.6.4. Let $n = 3$ with

$$\mathbf{W} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F & D & F \\ F & F & F \\ F & F & F \end{pmatrix},$$

where D is soft opposition on $S = \mathbb{R}$. Obviously, for each agent, weight sum restriction (1.6.4) is satisfied; e.g., for agent $i = 1$, we have

$$W_{i,\mathcal{F}_i} = \frac{2}{3} + \frac{2}{3} = \frac{4}{3} = 1 + \frac{1}{3} = 1 + W_{i,\mathcal{O}_i}.$$

Then $\mathbf{A}$ has the structure

$$\mathbf{A} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

The eigenvalues of $\mathbf{A}$ are 1 and $\frac{1}{4} \pm \frac{1291}{4000} i$. Thus, since there are three distinct eigenvalues of a 3×3 system, each eigenvalue has algebraic multiplicity of 1, and, obviously, 1 is the only eigenvalue on the unit circle and $\rho(\mathbf{A}) = 1$. Hence, $\mathbf{W} \circ \mathbf{F}$ induces a consensus by Proposition 1.6.7. The limit consensus is obtained by computing $\mathbf{s}^\intercal \mathbf{b}(0)$ where $\mathbf{s} = (\frac{1}{2}, 0, \frac{1}{2})^\intercal$, i.e.,

$$b(\infty) = \sum_{i=1}^{n} s_i b_i(0) = \frac{1}{2} b_1(0) + \frac{1}{2} b_3(0).$$

For instance, if agents start with initial opinions $\mathbf{b}(0) = (1, 2, -2)^\intercal$, they will end up at the consensus vector $(-\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2})^\intercal$. We illustrate dynamics for this setup in Figure 1.15.

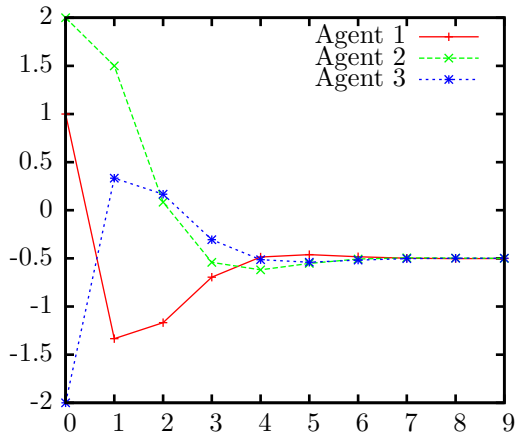


Figure 1.15: Opinions $\mathbf{b}(t)$, for $t = 0, \ldots, 10$, for the process discussed in Example 1.6.4.

1.7 Conclusions

Opinions are important in an economic context (and other contexts) since they shape the demand for products, set the political course, and guide, in general, socio-economic behavior. Models of opinion dynamics model how individuals form opinions or beliefs about an underlying state or a discussion topic. Typically, in the social networks literature, subjects may communicate with other individuals, their peers, in this context, enabling them to aggregate dispersed information. Bayesian models of opinion formation assume that agents form their opinions in a fully rational manner and have an accurate 'model of the world' at their disposal, both of which are questionable and unrealistic assumptions, if compared with actual social learning processes of human individuals (cf. Chandrasekhar, Larreguy, and Xandri, 2012; Corazzini et al., 2012, etc.). Non-Bayesian models, and most prominently the classical DeGroot model of opinion formation, while also not unproblematic (cf. Acemoglu and Ozdaglar, 2011), posit that agents employ simple 'rule-of-thumb' heuristics to integrate the opinions of others. Unfortunately, both the non-Bayesian and Bayesian paradigms typically lead individuals to a consensus, which apparently contradicts the facts as people disagree with others on many issues of (everyday) life. In the context of DeGroot learning models, some works have sought to address this issue, either by assuming a homophily principle whereby agents limit their communication to those who hold similar opinions as themselves or by introducing stubborn agents, modeling, e.g., opinion leaders, who never update their opinions. Both approaches are, again, debatable since the approach based on stubborn agents assumes truly autark individuals and models based on homophily can typically neither explain short-term opinion fluctuations (see the discussion in Acemoglu, Como, et al., 2012), nor functional disagreement whereby disagreeing opinions are, in fact, opposing viewpoints rather than arbitrary and unrelated. Finally, the homophily models that can account for disagreement rely on the condition that some subsets of society do not communicate with, or learn from, each other, at least from some time point onward, as in the model based on stubborn agents — a requirement that we find problematic since it is difficult to imagine subsets of society without *any* mutual influence.[32] In any case, models based on homophily and stubborn agents both ignore *negative* relationships between individuals as potential sources for conflict and disagreement.

In the current work, we have investigated opinion dynamics under opposition, as (such) a potentially alternative explanation for disagreement. In our setup, agents are driven by two forces: they want to adjust their opinions to match those of the agents they follow (their 'ingroup' or those they trust) and, in addition, they want to adjust their opinions to match the 'inverse' of those of the agents they oppose (their 'outgroup' or those they distrust). Best responses in this setting lead us to a DeGroot-like opinion updating process whereby agents form their next period opinions via weighted arithmetic averages of their neighbors' (possibly inverted) opinion signals. Our paradigm can account for a variety of phenomena such as consensus, neutrality, disagreement, and (functional) polarization, depending upon

[32] Particularly in today's 'globalized world'.

network (multigraph) structures and specifications of deviation functions, as we have demonstrated, both analytically and by means of simple simulations. Psychologically and socio-economically, we have interpreted opposition as arising either from rebels; countercultures; rejection of the norms and values of disliked others, as 'negative referents'; or, simply, distrust.

One issue that has been left undiscussed so far is the fact that, possibly unlike social norms and values, opinions oftentimes (though probably far less than always) admit a 'truth' against which they may be evaluated; accordingly, some research papers (e.g., Golub and Jackson, 2010) have asked for the conditions under which agents may converge to a consensus that is even correct. Under opposition, as we have specified, such a convergence to a correct consensus is severely compromised, as we have indicated. Namely, if agents converge to a consensus at all, then, as seen, such a consensus is typically a neutral consensus. In the continuous approach, if the opinion spectrum is a dense subset of the real line and the set of neutral opinions is, as we might plausibly assume, small (e.g., finite or even a singleton), then, from a probabilistic perspective, chances for agents of reaching a correct consensus are virtually zero. Alternatively, if agents disagree, or, more specifically, polarize, then, of course, at most one group of agents can be correct but, given a functional dependence of limiting opinions, we would expect none to be.

Finally, concerning future research directions within our context, both weight links and opposition links between agents, $\mathbf{W}$ and $\mathbf{F}$, have been assumed exogenous in the current work. Prospectively, it might be worthwhile to consider endogenous link formation processes. In particular, the origin and evolution of opposition behavior, and its relation to agents' opinions and external factors, such as, most importantly, to external truth, might be of interest, among other things.

Appendix 1.A Theorems and proofs

Theorem 1.A.1 (Brouwer's fixed point theorem)**.** Let $K \subseteq \mathbb{R}$ be convex and compact and let $f : K \to K$. Then, f has a fixed point.

Lemma 1.A.1. Let D be soft opposition on $S = [-\beta, \beta]$, for some $\beta > 0$. Let $\mathbf{W} \circ \mathbf{F}$ be an arbitrary operator with representation $\mathbf{A}$ such that $A_{ii} = 0$ and $A_{ij} = A_{ji}$. Then, $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite if and only if there exists a diagonal matrix $\mathbf{\Delta}$ such that $\mathbf{\Delta A \Delta} = |\mathbf{A}|$, where $|\mathbf{A}|$ denotes the matrix with entries $|A_{ij}|$.

Proof. Let $\mathbf{W} \circ \mathbf{F}$ be opposition bipartite with partition $(\mathcal{N}_1, \mathcal{N}_2)$. Choose $\Delta_{ii} = 1$ if $i \in \mathcal{N}_1$ and $\Delta_{ii} = -1$ if $i \in \mathcal{N}_2$. Then, as one can verify, $\mathbf{\Delta A \Delta} = |\mathbf{A}|$.

Conversely, let $\mathbf{\Delta A \Delta} = |\mathbf{A}|$ so that $|A_{ij}| = \Delta_{ii}\Delta_{jj}A_{ij}$. Hence, if $A_{ij} \neq 0$, $\Delta_{ii}, \Delta_{jj} \in \{\pm 1\}$. Choose $i \in \mathcal{N}_1$ if $\Delta_{ii} = 1$ and $i \in \mathcal{N}_2$ otherwise. □

Lemma 1.A.2. Let $|\mathbf{A}| = \mathbf{\Delta A \Delta}$ as in Lemma 1.A.1. Then $\mathbf{A}$ and $|\mathbf{A}|$ have the same eigenvalues with the same multiplicities.

Proof. Since $\mathbf{\Delta}^{-1} = \mathbf{\Delta}$, $\mathbf{\Delta A \Delta}^{-1}$ represents a similarity transformation. □

Lemma 1.A.3. Let D be soft opposition on $S = [-\beta, \beta]$, for some $\beta > 0$. Let $\mathbf{W} \circ \mathbf{F}$ an arbitrary operator with representation $\mathbf{A}$ such that $A_{ii} = 0$ and $A_{ij} = A_{ji}$. Then, $\mathbf{W} \circ \mathbf{F}$ is opposition bipartite if and only if $\lambda = 1$ is an eigenvalue of $\mathbf{A}$.

Proof. Altafini (2013), Lemma 1, shows that $0 \in \sigma(\mathbf{L})$ if and only if $\mathbf{A}$ is opposition bipartite where $\mathbf{L} = \mathbf{I}_n - \mathbf{A}$. Clearly, $1 \in \sigma(\mathbf{A}) \iff 0 \in \sigma(\mathbf{L})$. □

Bibliography

[1] Robert P. Abelson. "Mathematical models of the distribution of attitudes under controversy". In: *Contributions to mathematical psychology*. Ed. by N. Frederiksen and H. Gulliksen. New York: Rinehart Winston, 1964, pp. 142–160.

[2] Daron Acemoglu, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar. *Opinion Fluctuations and Disagreement in Social Networks*. LIDS report 2850. to appear in *Mathematics of Operations Research*. 2012. URL: `http://web.mit.edu/asuman/www/documents/disagreementsubmitted.pdf`.

[3] Daron Acemoglu, Munzer A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. "Bayesian Learning in Social Networks". In: *Review of Economic Studies* 78 (4 2011), pp. 1201–1236.

[4] Daron Acemoglu and Asuman Ozdaglar. "Opinion Dynamics and Learning in Social Networks". In: *Dynamic Games and Applications* 1 (1 2011), pp. 3–49.

[5] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. "Spread of (Mis)Information in Social Networks". In: *Games and Economic Behavior* 70 (2 2010), pp. 194–227.

[6] Claudio Altafini. "Consensus Problems on Networks With Antagonistic Interactions". In: *IEEE Transactions on Automatic Control* 58 (4 2013).

[7] Maxim Ananyev and Sergei Guriev. *Causal Effect of Income on Trust: Evidence from the 2009 Crisis in Russia*. Working Paper. 2013.

[8] Stef Aupers. "'Trust no one': Modernization, paranoia and conspiracy culture". In: *European Journal of Communication* 27 (2012), pp. 22–34.

[9] Delia Baldassari and Peter Bearman. "Dynamics of Political Polarization". In: *American Sociological Review* 72 (2007), pp. 784–811.

[10] Abhijit V. Banerjee. "A simple model of herd behavior". In: *Quarterly Journal of Economics* 107 (3 1992), pp. 797–817.

[11] Abhijit V. Banerjee and Drew Fudenberg. "Word-of-mouth learning". In: *Games and Economic Behavior* 46 (3 2004), pp. 1–22.

[12] David Beasley and Dan Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge, UK: Cambridge University Press, 2010.

[13] Philip Bonacich. "Factoring and weighting approaches to status scores and clique identification". In: *The Journal of Mathematical Sociology* 2 (1 1972), pp. 113–120.

[14] Marylinn B. Brewer. "In-Group Bias in the minimal intergroup situation: A cognitive-motivational analysis". In: *Psychological Bulletin* 86 (2 1979), pp. 307–324.

[15] Berno Buechel, Tim Hellmann, and Stefan Klößner. "Opinion Dynamics and Wisdom under Conformity". Working Paper. 2013.

[16] Berno Buechel, Tim Hellmann, and Stefan Klößner. "Opinion Dynamics under Conformity". Working Paper. 2012.

[17] Berno Buechel, Tim Hellmann, and Michael Pichler. "The Dynamics of Continuous Cultural Traits in Social Networks". Working Paper. 2012.

[18] Zhigang Cao, Mingmin Yang, Xinglong Qu, and Xiaoguang Yang. "Rebels Lead to the Doctrine of the Mean: Opinion Dynamic in a Heterogeneous DeGroot Model". In: *The 6th International Conference on Knowledge, Information and Creativity Support Systems*. Bejing, China, 2011, pp. 29–35.

[19] Emanuele Castano, Vincent Yzerbyt, David Bourguignon, and Eléonore Seron. "Who may Enter? The Impact of In-Group Identification on In-Group/Out-Group Categorization". In: *Journal of Experimental Social Psychology* 38 (2002), pp. 315–322.

[20] Arun G. Chandrasekhar, Horacio Larreguy, and Juan P. Xandri. *Testing models of Social Learning on Networks: Evidence from a lab experiment in the field*. Working Paper. 2012.

[21] Gary Charness, Luca Rigotti, and Aldo Rustichini. "Individual Behavior and group membership". In: *American Economic Review* 97 (4 2007), pp. 1340–1352.

[22] Geoffrey L. Cohen. "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs". In: *Journal of Personality and Social Psychology* 85 (5 2003), pp. 808–822.

[23] Luca Corazzini, Filippo Pavesi, Beatrice Petrovich, and Luca Stanca. "Influential listeners: An experiment on persuasion bias in social networks". In: *European Economic Review* 56 (6 2012), pp. 1276–1288.

[24] Fred Davis. *On Youth Subcultures: The Hippie Variant*. New York: General Learning Press, 1971.

[25] Guillaume Deffuant, David Neau, Frederic Amblard, and Gerard Weisbuch. "Mixing beliefs among interacting agents". In: *Advances in Complex Systems* 3 (2000), pp. 87–98.

[26] Morris H. DeGroot. "Reaching a Consensus". English. In: *Journal of the American Statistical Association* 69.345 (1974), pp. 118–121. ISSN: 01621459. URL: `http://www.jstor.org/stable/2285509`.

[27] Peter M. DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. "Persuasion Bias, Social Influence, And Unidimensional Opinions". In: *The Quarterly Journal of Economics* 118.3 (Aug. 2003), pp. 909–968. URL: `http://ideas.repec.org/a/tpr/qjecon/v118y2003i3p909-968.html`.

[28] Morton Deutsch. *The resolution of conflict: Constructive and destructive processes*. New Haven CT: Yale University Press, 1973.

[29] Franz Dietrich and Christian List. *Opinion pooling on general agendas*. Working Paper. 2008.

[30] Igor Douven and Alexander Riegler. "Extending the Hegselmann-Krause model I". In: *The Logic Journal of the IGPL* 18 (2 2010), pp. 323–335.

[31] Igor Douven and Alexander Riegler. "Extending the Hegselmann-Krause model II". In: *Proceedings of ECAP 2009* (2009).

[32] Igor Douven and Alexander Riegler. "Extending the Hegselmann-Krause model III: From single beliefs to complex belief states". In: *Episteme* 6 (2009), pp. 145–163.

[33] Giuseppe Facchetti, Giovanni Iacono, and Claudio Altafini. "Computing global structural balance in large-scale signed social networks". In: *Proceedings of the National Academy of Sciences* 108.52 (Dec. 27, 2011), pp. 20953–20958. ISSN: 1091-6490. DOI: `10.1073/pnas.1109521108`. URL: `http://dx.doi.org/10.1073/pnas.1109521108`.

[34] Pengyi Fan, Pei Li, Hui Wang, Zhihong Jiang, and Wei Li. "Opinion Interaction Network: Opinion dynamics in Social Networks with Heterogeneous relationships". In: *ISI-KDD '12 Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics Article*. Beijing, China, 2012.

[35] Sebastian Fehrler and Michael Kosfeld. "Can You Trust the Good Guys? Trust Within and Between Groups with Different Missions". Working Paper. 2013.

[36] Thomas Fent, Patrick Groeber, and Frank Schweitzer. "Coexistence of Social Norms based on In- and Out-group Interactions". In: *Advances of Complex Systems* 10 (2 2007), pp. 271–286.

[37] John R.P. French. "A formal theory of social power". In: *Psychological Review* 63 (3 1956), pp. 181–194.

[38] Noah E. Friedkin and Eugene C. Johnsen. "Social influence and opinions". In: *Journal of Mathematical Sociology* 15 (3-4 1990), pp. 193–205.

[39] Noah E. Friedkin and Eugene C. Johnsen. "Social influence networks and opinion change". In: *Advances in Group Processes* 16 (1999), pp. 1–29.

[40] Douglas Gale and Shachar Kariv. "Bayesian Learning in Social Networks". In: *Games and Economic Behavior* 45 (2 2003), pp. 329–346.

[41] Uri Gneezy. "Deception: The role of consequences". In: *American Economic Review* 95 (2005), pp. 384–394.

[42] Benjamin Golub and Matthew O. Jackson. "How Homophily Affects the Speed of Learning and Best-Response Dynamics". In: *The Quarterly Journal of Economics* 127 (3 2012), pp. 1287–1338.

[43] Benjamin Golub and Matthew O. Jackson. "Naïve Learning in Social Networks and the Wisdom of Crowds". In: *American Economic Journal: Microeconomics* 2 (1 2010), pp. 112–149.

[44] Patrick Groeber, Jan Lorenz, and Frank Schweitzer. "Dissonance minimization as a microfoundation of social influence in models of opinion formation". In: *Journal of Mathematical Sociology* (2013).

[45] Frank Harary. "A criterion for unanimity in French's theory of social power". In: *Studies in social power* (1959).

[46] Godfrey H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge: Cambridge University Press, 1934.

[47] Rainer Hegselmann and Ulrich Krause. "Opinion dynamics and bounded confidence: models, analysis and simulation". In: *J. Artificial Societies and Social Simulation* 5.3 (2002).

[48] Rainer Hegselmann and Ulrich Krause. "Opinion Dynamics Driven by Various Ways of Averaging". In: *Computational Economics* 25.4 (2005), pp. 381–405. URL: `http://EconPapers.repec.org/RePEc:kap:compec:v:25:y:2005:i:4:p:381-405`.

[49] Rainer Hegselmann and Ulrich Krause. "Truth and Cognitive Division of Labour: First Steps Towards a Computer Aided Social Epistemology". In: *Journal of Artificial Societies and Social Simulation* 9.3 (2006), p. 10. ISSN: 1460-7425. URL: `http://jasss.soc.surrey.ac.uk/9/3/10.html`.

[50] Frederik Herzberg. *An algebraic approach to general aggregation theory: Propositional-attitude aggregators as MV-homomorphisms*. Working Paper. 2011.

[51] Marc M. Howard. "Postcommunist civil society in comparative perspective". In: *Demokratizatsiya* (2002), pp. 285–305.

[52] Matthew O. Jackson. *Social and Economic Networks*. Princeton: Princeton University Press, 2009.

[53] Mark P. Jones. *Electoral Laws and the Survival of Presidential Democracies*. Notre Dame: University of Notre Dame Press, 1995.

[54] James A. Kitts. "Social influence and the emergence of norms amid ties of amity and enmity". In: *Simulation Modelling Practice and Theory* 14 (2006), pp. 407–422.

[55] Gerald H. Kramer. "Short-term fluctuations in U.S. voting behavior, 1896–1964". In: *American Political Science Review* 65 (1 1971), pp. 131–143.

[56] Ulrich Krause. "Compromise, consensus, and the iterations of means". In: *Elemente der Mathematik* 64 (1 2009), pp. 1–8.

[57] Paul Krugman. *Ricardo's difficult idea*. Paper for Manchester conference on free tade. Available on his official web page. Mar. 1996.

[58] Keith Lehrer. "Rationality as Weighted Averaging". In: *Synthese* 57 (1983), pp. 283–295.

[59] Keith Lehrer and Carl Wagner. "Rational consensus in science and society". In: *Reidel, Dordrecht* (1981).

[60] Jure Leskovec, Daniel P. Huttenlocher, and Jon M. Kleinberg. "Signed networks in social media." In: *CHI*. Ed. by Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden. ACM, 2010, pp. 1361–1370. ISBN: 978-1-60558-929-9. URL: http://dblp.uni-trier.de/db/conf/chi/chi2010.html#LeskovecHK10.

[61] Joel Lobel. "Economists' Models of Learning". In: *Journal of Economic Theory* 94 (2000), pp. 241–261.

[62] Kevin J. McConway. "Marginalization and linear opinion pools". In: *Journal of the American Statistical Association* 76.374 (1981), pp. 410–414.

[63] D. Harrison McKnight and Norman L. Chervany. "Trust and Distrust Definitions: One bite at a time". In: *Trust in Cyber-Societies: Integrating the Human and Artificial Perspectives*. Ed. by R. Falcone, M. Singh, and Y.H. Tan. Berlin: Springer, 2001, pp. 27–54.

[64] Glen D. Mellinger. "Interpersonal trust as a factor in communication". In: *The Journal of Abnormal and Social Psychology* 52 (3 1956), pp. 304–309.

[65] Carl D. Meyer. *Matrix analysis and applied linear algebra*. Philadelphia: SIAM, 2000.

[66] William Mishler and Richard Rose. "Trust, distrust and skepticism: Popular evaluations of civil and political institutions in post-communist societies". In: *The Journal of Politics* 59 (2 1997), pp. 418–451.

[67] Susan Mitchell. *The official guide to American Attitudes: Who thinks what about the issues that shape our lives*. Ithaca, New York: New Strategist Publications, 1996.

[68] Victor Nee, Sonja Opper, and Hakan Holm. *Trust and Economic Behavior in a Low-trust Society*. Working Paper. 2013.

[69] Theodore M. Newcomb. "An approach to a theory of communicative acts". In: *Psychological Review* (60 1953), pp. 393–404.

[70] Zhengzheng Pan. "Trust, influence, and convergence of behavior in social networks". In: *Mathematical Social Sciences* 60 (1 2010), pp. 69–78.

[71] Julian Rode. "Truth and trust in communication. An experimental study of behaviour under asymmetric information". In: *Games and Economic Behavior* 68 (1 2010), pp. 325–338.

[72] D. Rosenberg, E. Solan, and N. Vieille. "Informational Externalities and Convergence of Behavior". Preprint. 2006.

[73] Julian B. Rotter. "Generalized expectancies for interpersonal trust". In: *American Psychologist* 26 (5 1971).

[74] Ariel Rubinstein and Peter C. Fishburn. "Algebraic aggregation theory". In: *Journal of Economic Theory* 38 (1 1986).

[75] Yaacov Schul, Ruth Mayo, and Eugene Burnstein. "The value of distrust". In: *Journal of Experimental Social Psychology* 44 (2008), pp. 1293–1302.

[76] Guodong Shi, Alexandre Proutiere, Mikael Johansson, John S. Baras, and Karl H. Johannson. "The Evolution of Beliefs over Signed Social Networks". Available at http://arxiv.org/pdf/1307.0539.pdf. 2013.

[77] Matthias Sutter. "Deception through truth telling?! Experimental evidence from individuals and teams". In: *The Economic Journal* 119 (534 2009), pp. 47–60.

[78] Ryuhei Tsuji. "Interpersonal influence and attitude change toward conformity in small groups: a social psychological model". In: *Journal of Mathematical Sociology* 26 (2002), pp. 17–34.

[79] Carl Wagner. "Consensus through respect: A model of rational group decision-making". In: *Philosophical studies: An international Journal for Philosophy in the Analytic Tradition* 34 (4 1978), pp. 335–349.

[80] World Public Opinion. *Iraq: The Separate Realities of Republicans and Democrats*. Technical report. available at http://www.worldpublicopinion.org/pipa/articles/brunitedstatescanadara/186.php. Washington DC: World Public Opinion, 2006.

[81] Ercan Yildiz, Daron Acemoglu, Asuman Ozdaglar, Amin Saberi, and Anna Scaglione. *Discrete Opinion Dynamics with Stubborn Agents*. LIDS report 2870. to appear in *ACM Transactions on Economics and Computation*. 2012. URL: `http://web.mit.edu/asuman/www/documents/voter-submit.pdf`.

[82] John M. Yinger. "Contraculture and subculture". In: *American Sociological Review* 25 (1960), pp. 625–635.

[83] John M. Yinger. "Countercultures and social change". In: *American Sociological Review* 42 (6 1977), pp. 833–853.

[84] Bo-Yu Zhang, Zhi-Gang Cao, Cheng-Zhong Qin, and Xiao-Guang Yang. *Fashion and homophily*. Available at SSRN: http://ssrn.com/abstract=2250898 or http://dx.doi.org/10.2139/ssrn.2250898. 2013.

Chapter 2

(Failure of the) Wisdom of the crowds in an endogenous opinion dynamics model with multiply biased agents

Abstract

We study an *endogenous* opinion (or, belief) dynamics model where we endogenize the social network that models the link ('trust') weights between agents. Our network adjustment mechanism is simple: an agent increases her weight for another agent if that agent has been close to truth (whence, our adjustment criterion is 'past performance'). Moreover, we consider *multiply biased* agents that do not learn in a fully rational manner but are subject to persuasion bias — they learn in a DeGroot manner, via a simple 'rule of thumb' — and that have biased initial beliefs. In addition, we also study this setup under *conformity*, *opposition*, and *homophily* — which are recently suggested variants of DeGroot learning in social networks — thereby taking into account further biases agents are susceptible to. Our main focus is on *crowd wisdom*, that is, on the question whether the so biased agents can adequately aggregate dispersed information and, consequently, learn the true states of the topics they communicate about. In particular, we present several conditions under which wisdom fails.

2.1 Introduction

Crowds can be amazingly wise, even wiser than the most accurate individuals among them. An early formalization of this insight has been Concordet's Jury theorem from 1785 (Concordet, 1785), which states that a simple majority vote of the opinions of independent and fallible lay-people may provide near-perfect accuracy if the number of voters is sufficiently large.[1] Over a hundred years later, in 1906, Francis Galton found strong empirical support of Concordet's theoretical finding at an agricultural fair in Plymouth. At a weight-judging contest, participants were asked to privately estimate the weight of a chosen live ox after it had been slaughtered and dressed (meaning that the head and other parts were removed). The winner was the one whose estimate was closest to the true weight of the ox. When analyzing the results in a *Nature* article the following year (Galton, 1907), Galton found that the simple average of the entire crowd was even more accurate than the winner and that the median of the 787 valid guesses, 1197 pounds, was extremely close to the true weight, 1198 pounds (cf. Bahrami et al., 2012; Acemoglu and Ozdaglar, 2011). This finding was obtained even though most participants were no 'experts' in this contest, with little specialized knowledge in butchery; yet, their estimates could

[1] Which also requires that each individual in the group of voters is more likely correct than not.

obviously contribute to the crowds' overall success. Galton took this result as evidence that democratic political systems may work.

Yet, contradicting this optimistic viewpoint concerning the wisdom of crowds, it has also been observed that groups of individuals may be quite fallible, and possibly even more fallible than most or all of their members. One result of this kind is already hidden in Concordet's Jury theorem: namely, if each lay-person is just slightly 'too uniformed' (or slightly 'too much mistaken'), then the majority vote may be much less accurate than each individual's estimate. Drawing upon empirical observations, a comprehensive illustration of 'crowd madness' has been brought forward in Scottish journalist Charles Mackay's (Mackay, 1841) work *The extraordinary and popular delusions and the madness of crowds*, where the author chronicles 'humankind's collective follies', including financial bubbles, in the economics context, and other popular 'delusions' such as witch-hunts and fortune-telling (cf. Bahrami et al., 2012), thus challenging the claim that "two heads are better than one".

Today — while, according to scholars' opinions, the question of wisdom of crowds continues to be one of the most important issues facing social sciences in the twenty-first century —[2] more is known on group wisdom and collective failure. On the one hand, the mean of the opinions of several individuals may become increasingly accurate, for large groups, merely as a consequence of the law of large numbers. This holds under restrictive assumptions — in particular, that the beliefs of individuals are independent and probabilistically centered around truth such that, on an aggregate level, individual errors cancel out. On the other hand, much empirical literature, foremostly in psychology, has documented that, frequently, "groups outperform individuals [...], although groups typically fall short of the performance of their highest-ability members" (Kerr, MacCoun, and Kramer, 1996, p.691).[3] In fact, a very recent experiment by Lorenz et al. (2011) finds that 'social influence', in a broad meaning, in a group causes individuals' beliefs to become more similar over time, without improvements in accuracy, however. Hence, much depends on how groups aggregate or process individual opinions and also on these initial predispositions of agents. Kerr, MacCoun, and Kramer (1996)'s insight is that whether groups perform better than individuals may depend, among other things, on the following aspects: (1) the way that groups aggregate the opinions of individuals (that is, the group decision, or belief integrating, process), (2) the *bias* of individuals, and (3) the type of bias. Concerning issue (1), the way agents in groups process their peers' beliefs, we assume a specific structural form below, which has empirically proved plausible for learning in the domain we consider (social networks).

Issue (2), individuals' bias, will be another central notion in our work. The classical work of Tversky and Kahnemann (1974) documents several biases human judgment is susceptible to. In particular, *anchoring biases* describe the psychological condition of humans to pay undue attention to initial values — e.g., typically, individuals estimate the product $9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ to be higher than the product of factors in reverse order, which is attributed to subjects' performing an initial approximate computation based on the first few terms, which entails a biasing anchor (the same effects may happen if the anchor is exogenously specified, e.g., by providing the subjects with random numbers as anchors and then querying them for their own judgement). *Biases of availabilty* refer to the phenomenon of assessing (and, consequently, possibly, misjudging) the probability of an event by the 'ease with which instances or occurrences can be brought to mind', and, finally, *biases of representativeness* lead subjects to assess the probability that an object is of a particular class (e.g., that a person has a certain profession) by the degree to which the object is representative of the class, which may lead to judgement errors because such reasoning ignores, e.g., base-rate frequencies. In another typological classification of bias, Kerr, MacCoun, and Kramer (1996) distinguish between *judgmental sins of imprecision* (systematically deviating from prescribed and precise use of information, such as ignoring Bayes' theorem when forming beliefs or being affected by framing; see Kahnemann and Tversky, 1984), *judgmental sins of commission* (using irrelevant information to arrive at a decision, such as the attractiveness of an accused) and *judgmental sins of omission* (ignoring relevant information, such as base-rate information).

We now describe the setup investigated in the current work, relating to the issues discussed above subsequently. We consider a *(social) network* of individuals, or agents, that form opinions, or beliefs,

[2] See the recent survey at `http://bit.ly/hR3hcS`.

[3] Groups can also blatantly fail, as, e.g., in *groupthink* (Janis, 1972), *hidden profiles* (Wittenbaum and Stasser, 1996), etc. See the overview in Kerr and Tindale (2004).

about an underlying state or a discussion topic.[4] We assume that agents start with some initial beliefs, at time zero, and then, as time progresses, learn from each other through *communication*. Communication between any two individuals takes place if there is a link between them in the network. In the current work, we assume a specific form of learning paradigm, *DeGroot learning*, that posits that agents update their beliefs by taking weighted arithmetic averages of their peers' past beliefs, whereby the weights are given by the (social) ties between the agents in the network. Much has been said on the adequacy (or inadequacy) of DeGroot learning — a 'boundedly rational' learning paradigm that posits that agents are susceptible to *persuasion bias*, not properly adjusting for the repetition of information they hear — which, as experiments claim (e.g., Chandrasekhar, Larreguy, and Xandri, 2012; Corazzini et al., 2012), appears as a more plausible standard of human social learning than, e.g., fully rational Bayesian learning and we refer the reader to, e.g., DeMarzo, Vayanos, and Zwiebel (2003), Golub and Jackson (2010) or Acemoglu and Ozdaglar (2011) for extensive discussions. While the DeGroot model of opinion formation is quite old, dating back to Morris H. DeGroot's (DeGroot, 1974) seminal work, the framework has only more recently received increasing attention from the economics community.

In this context, one matter that has been put forth as a central guiding question in DeGroot learning models, and which connects to our initial discussion, is whether the 'naïve' DeGroot learners, who commit the 'sin of imprecision' of not (properly) applying Bayes' theorem, can, in fact, become 'wise' (Golub and Jackson, 2010). Here, a society (set of agents) is called *wise*, roughly, if it reaches a consensus — in the limit, as time (discussion periods) goes to infinity — that corresponds to truth. In Golub and Jackson (2010), the question relating to wisdom has been answered in the affirmative — (even) naïve (DeGroot) learners do become wise under rather mild conditions; namely, all that is required is that no naïve learner is *excessively influential*, whereby an agent is excessively influential if his social influence (how limiting beliefs depend on this agent's initial beliefs) does not converge to zero as society grows. In undirected networks (social ties are mutual) with uniform weights, an obstacle to wisdom would then, e.g., be that each agent newly entering society assigns, e.g., a constant fraction of his links to a particular agent, who would then be excessively influential. Hence, as long as links are somewhat 'democratically' balanced, naïve DeGroot learners would apparently become wise. While we hold the analysis of Golub and Jackson (2010) to be an important 'benchmark' for DeGroot learning, we think that it is overly optimistic in at least one of its critical two assumptions, namely, the (1) *unbiasedness* of agents' initial beliefs.[5]

In the current work, we drop, in particular, the largely implausible, as we find, assumption (1). Our central notion will be as illustrated in Figure 2.1, which we adapt from Einhorn, Hogarth, and Klempner (1977). In words, we assume that some agents' initial beliefs are *biased*, with an expected value that is different from truth μ, and that other agents' initial beliefs are unbiased, with an expected value that equals truth μ — we remark here that we abstract away from the precise type of bias some agents' initial beliefs are subject to, that is, we are agnostic about whether, e.g., agents commit sins of imprecision, commission, or omission in forming their initial beliefs, simply assuming *that* at least some agents' initial beliefs are biased. We might, if we wish, label the first kind of agents 'non-experts' and the second 'experts', although this might be slightly misleading, as even experts can be biased, of course; nonetheless, for convenience, we keep this terminology in the following. A situation as sketched may be quite challenging to assess, for individuals. Ignoring non-experts may be suboptimal, in some circumstances, because their verdicts may still not be totally irrelevant in that their opinions may have (relatively) large probability of being close to truth. Consider, in particular, the bottom part of Figure 2.1 where the 'expert' is unbiased but has high variance. In this case, for each 'closeness interval' around truth, the non-expert's initial belief has higher probability of falling within this interval than the expert's beliefs. Thus, if there is exactly one expert and one non-expert, it would be optimal, for an outside observer, to disregard the expert's opinion and, in the absence of further information, adopt the non-expert's opinion. However, if there are *many* experts with identical and independent distributions

[4] Opinions or beliefs are important, from an economics perspective, because they crucially shape economic behavior: consumers' opinions about a product determine the demand for that product and majority opinions set the political course, etc. See Buechel, Hellmann, and Klößner (2013).

[5] The other critical assumption is biasedness of the belief formation process (DeGroot learning — agents are prone to persuasion bias). Finally, of crucial relevance is also independence of agents' initial beliefs, which we do not challenge here, however.

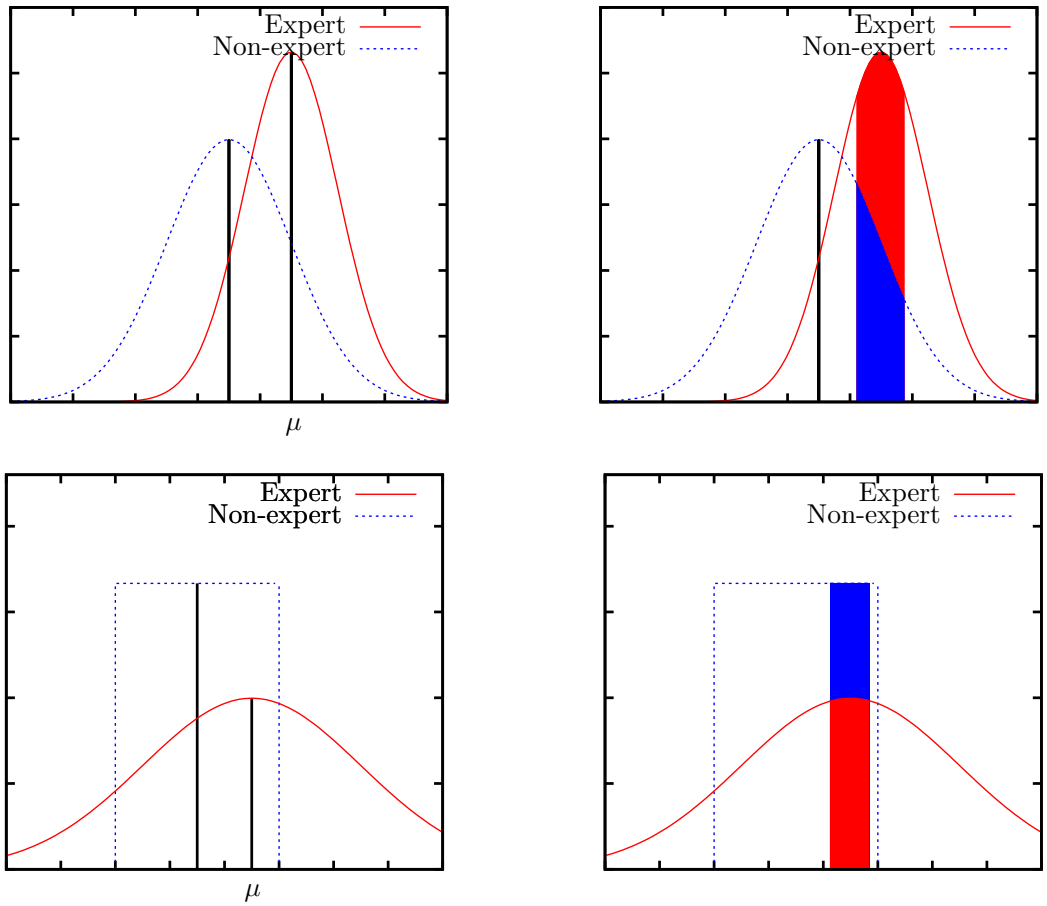


Figure 2.1: Schematic illustration of experts' and non-experts' distribution of initial beliefs. Right figures show probability masses of falling within an (arbitrary) small interval around truth, for both experts and non-experts.

and also *many* non-experts with identical and independent distributions, then an optimal aggregation of information would ignore the non-experts' and average the experts' opinions.

We study this setup in an *endogenous* DeGroot learning model, where we endogenize the (social) network. In particular, we assume that the 'trust' links between agents in the network are based on 'past performance', which has been outlined as a relevant reputation building criterion in the psychology literature (cf. Yaniv and Kleinberger, 2000; Yaniv, 2004). We think that the endogenous model is the 'right' setup for our investigation of wisdom in DeGroot learning under biased initial beliefs because if the network structure is assumed exogenous, then one relatively uninteresting solution to the wisdom problem would, e.g., be to ignore the biased agents — in contrast, in the endogenous model, the question arises what weighting scheme individuals *actually* learn for the biased (and unbiased) agents, under given assumptions concerning the agents' behavior. Since we learn the network structure endogenously, by looking at agents' past performance (how often have they been close to truth previously?), we necessarily study learning in a *repeated setting*, where agents are involved in repeated communications over *multitudes* of topics, whereby past beliefs and their external validation may inform today's beliefs and the network structure. Our network learning rule is quite simple: we increment the weight that one agent places upon another by some $\delta > 0$ if the latter agent has been in a predefined 'η-radius' around truth for the current topic. We show that this is a 'utility maximizing' rule[6] provided that agents *expect*, subjectively, that all other agents' beliefs are unbiased, which we call the *bona fides* (or, 'good faith') assumption.

[6]Or at least a good approximation to a solution of a maximization problem.

Intruigingly, the bona fides assumption concerning the unbiasedness of other agents' initial beliefs may be consistent with the 'egocentric bias' hypothesis which suggests that "interacting human agents operate under the assumption that their collaborators' decisions and opinions share the same level of reliability" as their own; the upholding of this bias, even despite potential collective failure, might then be due to the social obligation to treat others as equal to oneself, despite their conspicuous inadequacy, or due to the urge to contribute to the group (Bahrami et al., 2012).

To summarize our model, in our setup, agents hold beliefs and learn, via communication, about a multitude of topics $X_1, X_2, X_3, \ldots$, each with an associated 'truth' $\mu_1, \mu_2, \mu_3, \ldots$. Within each topic X_k, 'discussion rounds' are indexed by discrete time steps $t = 0, 1, 2, \ldots$ and in each time step $t \geq 1$, agents *update* their beliefs on X_k by integrating their peers' beliefs, starting with some exogenously specified initial beliefs on X_k. After a topic has been communicated about (for an infinite amount of time), truth μ_k is *revealed*, whereupon agents adjust the 'trust' weights they assign to other agents (they 'learn', or 'grow', the network topology) based on agents' past performance: if an agent has been close to truth for topic X_k, agents increase their trust for this agent by increasing the respective weight by δ. Our agents are multiply biased (or '*naïve*'):

(i) At least some agents' initial beliefs are *systematically biased* in that the expected values of their initial beliefs are different from truth μ_k, for all $k = 1, 2, 3, \ldots$. For initial beliefs, we abstract away from the particular kind of bias agents are subject to, simply assuming *that* some kind of bias plays a role.

(ii) Agents are subject to *persuasion bias* in updating their beliefs on X_k in that they apply the DeGroot learning paradigm rather than a fully rational Bayesian belief updating framework.

(iii) In adjusting weights for other agents, agents are *egocentrically biased*: they assume that their own judgments are relevant (more precisely, their initial beliefs are unbiased) and they assume that their peers' beliefs share the same level of reliability as their own (more precisely, that their peers' initial beliefs are also unbiased). This bias justifies the weight adjustment rule — adding δ — that we have sketched (see Section 2.3).

Besides this basic setup, we consider refinements of standard DeGroot learning recently suggested — DeGroot learning under opposition, conformity, and homophily — in each case incorporating our endogenized network structure and, in addition, the three kinds of biases discussed above. We show that, in these more refined versions of DeGroot learning, which are supposed to endow the DeGroot learning paradigm with a more 'realistic' structure, wisdom is even more difficult to arrive at, as we discuss below.

Our main contributions over existing work are as follows.

- We more thoroughly investigate the concept of *bias* in social (network) learning — or more specifically, DeGroot learning — than previous literature. In particular, as mentioned, we allow agents' initial beliefs to be biased and consider further biases, as discussed.

- We endogenize the network structure in DeGroot learning and we do so by referring to the notion of 'past performance'. Of course, in the vast literature on networks, ('endogenous') network formation processes are not novel; often, however, the network is adapted, in the literature more or less relevant to our setup, by adding or deleting (costly) links as in Jackson and Watts (2002) and Goyal (2004), etc., rather than by increasing link weight based on agents' past performance. In DeGroot learning, self-evolving networks are discussed, e.g., in the work on DeGroot learning and homophily (e.g., Pan, 2010 and the Hegselmann and Krause models), but weight adjustments based on truth, as we model, must be considered distinct from these mechanisms.

- As mentioned, we consider *multitudes* of topics, rather than a single topic, in DeGroot learning, and we crucially allow truth to be *revealed* at some stage. This differs from all the previous work, where agents have been in the unfortunate situation of eternally communicating about a given topic, without ever knowing its true state.

- We incoporate other DeGroot variants in our setup. In particular, we provide an alternative to the homophily model designed by Hegselmann and Krause (Hegselmann and Krause, 2002), see Section 2.9.

- We derive a microeconomic foundation for weight adjustments as we implement by defining an individual agent's optimization problem — in particular, we assume that agents have negative utility from not knowing truth — and by computing a closed-form solution to this problem.[7] We then show how our heuristic weight adjustment rule — adding δ — corresponds to the solution of the optimization problem.

Our main findings are as follows.

- For the standard model, we first show that agents reach a *consensus* for almost all topics X_k, under weak conditions, in our endogenized DeGroot learning paradigm (Proposition 2.6.1 and Remark 2.6.2). This confirms the commonly held belief (cf. Acemoglu and Ozdaglar, 2011) that the standard DeGroot model leads agents to consensus (so easily) but also shows that, in our endogenized model, conditons that prevent consensus are, in fact, not satisfied.

- Next, we illustrate that if all agents' initial beliefs are *unbiased*, then agents in fact reach a consensus that is even correct, for 'large' topics X_k and as agent group size n becomes large. This holds both when agents adjust weights based on limiting beliefs and on initial beliefs (Propositions 2.6.3 and 2.6.4, respectively); we define the notions of relevant weight adjustment time points below. When there are *biased* agents, then agents' limiting beliefs are generally a convex combination of the unbiased agents' initial beliefs and the biased agents' initial beliefs. We demonstrate the truthfulness of this claim under various parametrizations (Propositions 2.6.5, 2.6.7, 2.6.8). We also give sufficient conditions on when agents may converge to truth, for large topics, even under the presence of biased agents (Propositions 2.6.5 and 2.6.6), but these conditions are 'low probability events' (or require a sufficiently high valuation of truth) and they hold only under the particular parametrization that agents stop learning the network topology in case 'everything is fine', as we define below.

  That limiting consensus beliefs are convex combinations of biased and unbiased beliefs may imply that limiting beliefs are 'arbitrarily' far off from truth, provided that the number of biased agents is sufficiently large (Corollary 2.6.3), thus demonstrating that agents do not optimally aggregate information in our endogenized DeGroot learning model, at least under certain conditions.

- Next, for opinion dynamics 'under opposition', a recently suggested DeGroot learning variant where agents are motivated by 'ingroup'/'outgroup' relationships (Eger, 2013), we show that even if all agents' initial beliefs are unbiased and, more particularly, agents receive arbitrarily accurate initial signals about topics, some agents may be arbitarily far off from truth. In other words, we show that if agents have additional incentives besides truth, namely, to disassociate from unliked others — such agents must be thought of as additionally biased; namely, they must be thought of as, e.g., committing the sin of omission to ignore the unliked others' relevant information and the sin of commission to incorporate irrelevant information, namely, the 'opposite' of unliked others' beliefs —[8] then wisdom is even more difficult to attain. This, in particular, concerns several important fields of everyday life, such as the political arena.

- Then, for DeGroot learning 'under conformity' — that is, when agents want to conform to a reference opinion (again, which may be thought of as a sin of commission) — another recent variant of DeGroot learning (Buechel, Hellmann, and Klößner, 2012), we show that even if the unbiased agents have never been truthful in the past, they may become arbitrarily influential, something that is not possible in the standard model, and which, again, shows that additional biases may worsen the case for wisdom.

- Finally, in case homophily also plays a role — that is, when agents have the tendency to adjust the social network topology based on agents with similar beliefs — then, again, wisdom is more difficult to arrive at. We show this (only) by simulation since this process is (much) more difficult to analyze analytically as it deals with learning matrices that are changing over time (and not

[7] In spirit, our approach is similar to that of DeMarzo, Vayanos, and Zwiebel (2003).
[8] They may generally be thought of as biased toward ingroup members, cf. Brewer (1979).

only across topics). In our context, homophily can also be seen as a *search bias* in which subjects overrate beliefs that are close to their own (cf. Kunda, 1990).

The structure of this work is as follows. In Section 2.2, we present related work, beyond what we have already referred to. In Section 2.3, we give a formal outline of our model, and, in Section 2.4, a 'justification' of our network learning rule. In Section 2.5, we introduce relevant notation. Then, in Sections 2.6, 2.7, 2.8, and 2.9, we derive our results, as outlined above, on the standard model, and the DeGroot variants under opposition, conformity, and homophily, respectively. In Section 2.10, we conclude. We list several proofs in the appendices; there, we also report on a 'small-scale' experiment on the (un)biasedness (and the distribution) of individuals' (initial) beliefs concerning several 'common knowledge questions'.

Before actually listing related work in Section 2.2, we now briefly discuss this experiment and the lessons that we learn from it.

A small-scale experiment concerning the (un)biasedness of individuals' beliefs. As we have mentioned, some research papers have assumed that individuals' initial beliefs on topics are unbiased, that is, centered around truth. Certainly, this assumption may sometimes be plausible, e.g., depending on the topic, but, as we have indicated, we do not think that the condition holds across a large spectrum of circumstances. We conducted an experiment where we asked individuals on `Amazon Mechanical Turk`[9] 16 'common knowledge questions'. The questions ranged from, to our opinion, rather easy problems such as 'What do you think is the year the first world war started?' or 'What do you think is $17 - 4 \times 2$?' to rather difficult problems, such as 'What do you think is the number of people per square mile in China's capital Beijing?' or 'What do you think is the diameter of the sun in miles?'. We list all 16 questions in Appendix 2.B.

On all questions, more than $n = 100$ subjects answered (between $n = 110$ to $n = 119$). Analyzing the answers (see Figures 2.16 and 2.17), we find that, typically, neither the mean of the answers nor the median are very close to the true value. In fact, on only 8 out of 16 questions is the median (which tends to be more reliable since it is not so much affected by outliers) within a 10% interval around truth, and on only 6 out of 16 questions does this hold for the mean. Looking at 1% intervals, these numbers drop to 6 and 2, respectively (for the mean, these questions are about the start of the first world war and the average height of an adult male US American). Such low numbers were truly surprising if in fact the assumptions of unbiasedness and independence of (initial) beliefs were true, given the validity of the law of large numbers. A slightly more detailed analysis is given in Appendix 2.B.

2.2 Related Work

Early and frequently cited predecessors of DeGrootian opinion dynamics are French (1956) and Harary (1959), although the now famous 'averaging' model of opinion and consensus formation has only been popularized through the seminal work of DeGroot (1974). At about the same time, Lehrer and Wagner (Wagner, 1978; Lehrer and Wagner, 1981; Lehrer, 1983) have developed a model of rational consensus formation in society that, in both its implications and its mathematical structure, is very similar to the DeGroot model. In the sociology literature, Friedkin and Johnsen (1990) and Friedkin and Johnsen (1999) develop models of social influence that generalize the DeGroot model. In more recent years, a renewed economic interest in the DeGroot model of opinion dynamics has emerged, leading to a number of further extensions proposed. For example, DeMarzo, Vayanos, and Zwiebel (2003), besides sketching psychological justifications for DeGroot learning relating to persuasion bias as discussed above, discuss time-varying weights on own beliefs that capture, e.g., the idea of a 'hardening of positions': over time, individuals may be more inclined to rely on their own beliefs rather than on those of their peers. Further extensions of the classical DeGroot model include Golub and Jackson (2010), whose contribution is to analyze weight structures such that DeGroot learners whose initial beliefs are *stochastically centered around truth* converge to a consensus that is correct, and the works of Daron Acemoglu and colleagues. For example, Acemoglu, Ozdaglar, and ParandehGheibi (2010) distinguish between regular and forceful

[9] Available at `https://www.mturk.com/mturk/`.

agents, such as, in an economic interpretation, monopolistic media (forceful agents influence others disproportionately), and Acemoglu, Como, et al. (2012) distinguish between regular and stubborn agents (the latter never update), to account for the phenomenon of disagreement in societies; in Yildiz et al. (2012), a discrete version of the DeGroot model with stubborn agents is analyzed in which regular agents randomly adopt one of their neighbors' binary opinions. Another interesting DeGroot variant is discussed in Buechel, Hellmann, and Klößner (2012) where agents' *stated opinions* may differ from their *true* (or private) opinions and where it is assumed that agents generally wish to state an opinion that is close to that of their peer group even if their true opinions may be very different (which is the 'conformity' aspect of their model); we review this work in more depth in Section 2.8. A similar approach is given in Buechel, Hellmann, and Pichler (2012), where DeGroot learning is applied to an overlapping generations model in which parents transmit traits to their children. Receivers who deviate from the opinion signals sent by senders — *rebels* — are discussed in Cao et al. (2011); see also the modeling in Zhang et al. (2013) where such behavior is interpreted in a 'fashion' context. A model with more general 'ingroup/outgroup' relationships and opposition toward outgroup members is described in Eger (2013), which we discuss in more detail in Section 2.7. Multi-dimensional real opinion spaces have been considered in Lorenz (2006) and a survey of generalizations of DeGroot models developed within physicist communities (e.g., density-based approaches in place of agent-based systems) is provided by Lorenz (2007). Groeber, Lorenz, and Schweitzer (2013) provide 'dissonance minimization' as a general microfoundation of a variety of heterogenous DeGroot-like opinion dynamics models.

Concerning DeGroot models with *endogenous* weight formation, one pattern of endogenous weight formation that has been studied in the literature is weight formation based on a *homophily principle*, in which agents assign positive weights to those individuals whose current opinions are 'similar' with their own. In Hegselmann and Krause (2002) — an approach with many extensions such as Hegselmann and Krause (2005), Hegselmann and Krause (2006), Douven and Riegler (2009a), Douven and Riegler (2009b), and Douven and Riegler (2010) — this leads to very interesting patterns of opinion formation in which, most prominently, the paradigms of plurality, polarization and consensus are observed, depending on specific parametrizations; most importantly, the definition of similarity, i.e., whether individuals are tolerant or not toward other opinions, affects which opinion pattern emerges. The model of Deffuant et al. (2000) is identical in setup to the Hegselmann and Krause model, except that two randomly determined agents, rather than all agents, update beliefs in each time step. Pan (2010) discusses a homophily variant in which agents assign trust weights to other agents *in proportion* to agents' current opinion distance — rather than by thresholding, as done in the Hegselmann and Krause models and in Deffuant et al. (2000) — which typically entails a consensus, in the limit. Homophily and DeGroot learning is also investigated in Golub and Jackson (2012), where the relationship between the speed of DeGrootian learning and homophily is discussed; in this model, homophily is — exogenously, however — modeled by random networks where the link probability between different groups is non-uniform, and is, in fact, higher between individuals of the same group. Endogenous weight formation typically implies time-varying weight matrices as belief updating operators and mathematical results on corresponding processes are, for instance, given in Lorenz (2005).

Recent empirical and experimental evidence on the validity of the DeGroot heuristic for learning in social networks has been provided in, e.g., Chandrasekhar, Larreguy, and Xandri (2012) and Corazzini et al. (2012). Interesting in our context is also the experiment by Lorenz et al. (2011), where individuals are placed in a situation consistent with our setup: individuals observe their peers' past beliefs (on social/geopolitical issues) and may update their current opinions accordingly. In addition, truth on each of the discussed topics becomes revealed, by the experimenter, after a certain fixed amount of time.

Social learning is also discussed in various other strands of literature besides those discussed, such as in herding models (cf. Banerjee, 1992; Gale and Kariv, 2003; Banerjee and Fudenberg, 2004), where agents usually converge to holding the same belief as to an optimal action. This conclusion generally applies to the observational learning setting (cf. Rosenberg, Solan, and Vieille, 2006; Acemoglu, Dahleh, et al., 2011), where agents are observing choices and/or payoffs of other agents over time and are updating accordingly. See also the references and the discussion in Golub and Jackson (2010). General overviews over social learning, whether Bayesian or non-Bayesian, whether based on communication or observation, are, in the economics context, for example, given in Lobel (2000) and Acemoglu and Ozdaglar (2011). In Acemoglu and Ozdaglar (2011), an extensive discussion of the 'pros and cons' of fully rational learning

models versus boundedly rational (most importantly, DeGroot-like) heuristics is provided.

As discussed in the introduction, group opinion and belief formation and decision making also has a long history in psychology. A crucial difference between such models and models of social learning is that, in the psychology studies and models, it is usually assumed, and even explicity demanded, for the group of individuals to reach a consensus in the course of the discussion process. A general overview over group decision making is given in Kerr and Tindale (2004) and other relevant literature, besides that sketched in the introduction, is, for example, Mannes (2009) and Budescu et al. (2003).

2.3 Model

A finite set $[n] = \{1, 2, \dots, n\}$ of n agents discusses a sequence $X_1, X_2, X_3, \dots$ of topics. Each agent $i = 1, \dots, n$ holds *initial beliefs* $b_i^k(0) \in S$ on issue X_k, where $k = 1, 2, 3, \dots$ and where S is a convex set that we may innocuously assume to be the whole of $\mathbb{R}$. Moreover, each topic has a corresponding *truth* $\mu_k \in S$ which denotes the 'true evaluation' of topic X_k. Agents *update* their beliefs on X_k by taking a *weighted average* of all other agents' beliefs, starting from initial beliefs:

$$b_i^k(t+1) = \sum_{j=1}^{n} W_{ij}^{(k)} b_j^k(t), \tag{2.3.1}$$

where $t = 0, 1, 2, 3, \dots$ and where $W_{ij}^{(k)}$ denotes the *weight* ('trust') that agent i assigns agent j for topic X_k; in Section 2.9, we let $W_{ij}^{(k)}$ also depend on time t, i.e., $W_{ij}^{(k)} = W_{ij}^{(k)}(t)$. We let the limiting beliefs of agent i for issue X_k be denoted by $b_i^k(\infty)$. Moreover, we assume that weight matrix $\mathbf{W}^{(k)}$ — which we also interpret as a 'learning matrix', or, as a *(social) network* — is *row-stochastic* for every topic k, that is,

$$\forall i, j: \ 0 \leq W_{ij}^{(k)} \leq 1, \quad \text{and} \quad \forall i: \ \sum_{j=1}^{n} W_{ij}^{(k)} = 1,$$

which means that the weights that agents assign each other are normalized to unity; we furthermore assume that weights carry over from one topic to another, as we explicate below. Crucially, we consider an *endogenous* weight formation process where agents *adjust* the weights they attribute to other agents based on the foundational principle of truth.

- If agent j has known truth μ_k for issue X_k (or, was 'close enough'), then it seems natural for agent i to increase his trust in j. Formally, we let

$$W_{ij}^{(k+1)} = \begin{cases} W_{ij}^{(k)} + \delta \cdot T\big(\big|N(\mathbf{b}^k(\infty), \mu_k)\big|\big) & \text{if } \big\|b_j^k(\tau) - \mu_k\big\| < \eta, \\ W_{ij}^{(k)} & \text{otherwise,} \end{cases} \tag{2.3.2}$$

  for all $k \geq 1$; by $\|\cdot\|$, we denote the absolute distance and by $|A|$ the cardinality of set A. Here, $N(\mathbf{b}^k(\tau), \mu_k) \subseteq \{1, \dots, n\}$ is the set of all agents i whose belief $b_i^k(\tau)$ for X_k at time τ is within an η-radius of μ_k and $T : \{1, \dots, n\} \to [0, \infty]$ is a function for which we specify the following: $m_1 \leq m_2 \implies T(m_1) \geq T(m_2)$ (T is non-increasing in its argument; 'knowing truth pays a weakly larger trust increment the less people know it'; see our discussion below). The variable τ models the relevant adjustment time point; we consider $\tau = 0$ ('adjusting based on initial beliefs') and $\tau = \infty$ ('adjusting based on limiting beliefs'). We take the variables η, with $\eta \geq 0$, and δ, with $\delta > 0$, as exogenous variables. We also refer to the variable η as the agents' *tolerance* since it describes the interval within which agents are tolerant against deviations from truth.

  Adjusting weights in case $b_j^k(\tau)$ is close to truth rather than exactly truth may also be interpreted as a boundedly rational heuristic for agents who cannot assess truth with infinite precision. *Note that after adjusting weights as in* (2.3.2)*, we renormalize weight matrices in order for them to*

satisfy the row-stochasticity condition, that is, with a slight abuse of notation, we let

$$W_{ij}^{(k+1)} \leftarrow \frac{W_{ij}^{(k+1)}}{\sum_{j'} W_{ij'}^{(k+1)}},$$

after all n agents have adjusted weights as in (2.3.2).

Discussion

Our endogenous DeGroot model appears quite simple and natural — we let agents adjust the network $\mathbf{W}$ in a way that incorporates 'past performance': whenever an agent has been close enough to truth, agents increase their trust for this agent by δ — except, possibly, for the weight adjustment time points and the factor $T(\cdot)$. Concerning weight adjustment time points, the question is *what is the relevant time point that an agent's belief should be (or is) compared to truth* μ_k for some issue X_k. Note that, for any issue X_k, there are infinitely many possible such time points — $t = 0, 1, 2, 3, \ldots$ — so this question admits no straightforward answer. We consider two relevant time points, namely, the beliefs that agents hold *initially*, at the beginning of communication, and the beliefs that agents hold in the *limit*, as time goes to infinity; these beliefs are agents' limiting beliefs, after communication on topic X_k has terminated. Both time points have some intuitive appeal, as we think. Initial beliefs say something on an agent's 'innate ability', *before* learning from others, and limiting beliefs may possibly be a more realistic reference point for weight adjustments if agents are perceived of as having 'limited memory' (limiting beliefs are the 'most recent' observations).[10] Concerning the function $T(\cdot)$, our intuition is as follows. The larger the group of agents who know the correct answer (or, as we consider as equivalent throughout, are 'close enough' to truth) for a given topic — that is, the larger is the set of agents whose limiting beliefs are correct — the smaller should be the trust weight increment that agents assign each other. Intuitively, the number of agents who are correct for a topic may be indicative of the topic's 'difficulty' or 'hardness'.[11] If $T(x) = 0$, for some x, then this means that the network is not adjusted if at least x agents know the truth on any one topic.[12] Consider Figure 2.2 for three typical exemplars of $T(\cdot)$, as we have in mind.

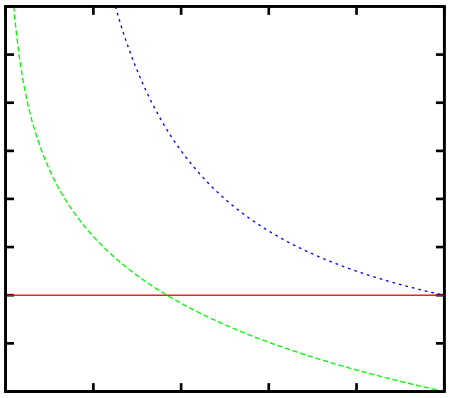


Figure 2.2: Three possible specifications of the function T in (2.3.2). Note that T is always non-increasing, as we have defined. For example, the red function is $T(\cdot) = 1$ and the dashed green function has $T(n) = 0$ and $T(x) > 0$ for $x < n$.

[10] Of course, it could be argued that an agent's 'average belief', somehow weighted over time, might also be a quantity that could be compared with truth μ_k.

[11] To make a crude example, 'everyone' may know what $3 + 3$ is — so that correct knowledge of this answer may not justify increased trust — but it took an Euclid to first discover the infinitude of the set of prime numbers.

[12] The condition $T(x) = 0$ may also be paraphrased as meaning that 'if at least x agents know truth on any one topic, then the network need not be changed' — that is, 'everything is fine' if at least x (e.g., $x = n$) agents know truth.

We also point out that, in our model, we logically differentiate between what is 'innate knowledge' (or, simply, '*ability*') and what is *socially learned* from others in that we think of initial beliefs as capturing ability and updating beliefs based on others' beliefs as the social learning process. Finally, we remark that we generally think of topics X_k as 'of the same kind' — that is, all of them are on sports or mathematics or natural science or politics or the stock market, etc. — in order to justify why network weights should carry over from one topic to another; see also our dicussion below.

Almost all throughout the work, we assume that agents are *homogenous* with respect to the tolerances η, the weight increments δ, and adjustment time points τ.

2.4 A justification of our weight adjustment procedure

The choice of a rational agent

We now derive a (micro-founded) justification of our weight adjustment rule (2.3.2). We first assume that agents $i = 1, \ldots, n$ have disutilities from not knowing truth for topic X_k, for $k = 1, 2, \ldots$. More precisely, we assume that agent i has utility function U_i from weight structure $\mathbf{W}^{(k)}$ for issue X_k as

$$U_i(\mathbf{W}^{(k)}) = U_i(\mathbf{W}^{(k)}; \mu_k, b_1^k(0), \ldots, b_n^k(0)) = -F\big(\, d(b_i^k(\infty), \mu_k)\,\big), \tag{2.4.1}$$

where $F : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is monotonically increasing and d is a *metric* — that is, in particular, $d(a, b) = 0$ if and only if $a = b$ — and where we assume that μ_k and initial beliefs $b_1^k(0), \ldots, b_n^k(0)$ are exogenous; also note how $b_i^k(\infty)$ depends on $\mathbf{W}^{(k)}$ (and $b_1^k(0), \ldots, b_n^k(0)$) via process (2.3.1). In other words, according to utility function (2.4.1), a larger distance between i's limiting belief $b_i^k(\infty)$ and truth μ_k does not lead to larger utility of agent i and when $b_i^k(\infty) = \mu_k$, then agent i attains maximum possible utility. For technical ease, we assume that d is the Euclidean distance and F has the simple quadratic form $F(z) = z^2$ such that

$$U_i(\mathbf{W}^{(k)}) = -\left\|b_i^k(\infty) - \mu_k\right\|^2 = -\big(b_i^k(\infty) - \mu_k\big)^2.$$

Now, we assume that $[\mathbf{W}^{(k)}]_i$, by which we denote the i-th row of $\mathbf{W}^{(k)}$, are the endogenous variables of agent i she wants to set in such a way as to maximize her utility U_i. Since agent i cannot affect the weight structure choices of agents i', with $i \neq i'$, we write U_i as a function of $[\mathbf{W}^{(k)}]_i$, rather than $\mathbf{W}^{(k)}$. Hence, we write

$$U_i\big([\mathbf{W}^{(k)}]_i\big) = -\big(b_i^k(\infty) - \mu_k\big)^2.$$

Assume next that agents $i = 1, \ldots, n$ have 'limited foresight' or 'finite horizon' in that they cannot foresee the dynamics of belief updating process (2.3.1) (which would also require knowledge of the other agents' weight choices) but that they take $b_i^k(1)$ as a reference variable, rather than $b_i^k(\infty)$.

Assumption 2.4.1. Agents $i = 1, \ldots, n$ have *limited foresight* or *finite horizon*. They choose weights $[\mathbf{W}^k(0)]_i$ to maximize

$$U_i\big([\mathbf{W}^{(k)}]_i\big) = -\big(b_i^k(1) - \mu_k\big)^2 = -\Big(\sum_{j=1}^{n} W_{ij}^{(k)} b_j^k(0) - \mu_k\Big)^2.$$

Our next assumption is that initial beliefs $b_1^k(0), \ldots, b_n^k(0)$ are *random variables*.

Assumption 2.4.2. Initial beliefs $b_1^k(0), \ldots, b_n^k(0)$ are *random variables*.

From Assumption 2.4.2, it follows that agents become *expected utility maximizers*: they choose weights $[\mathbf{W}^{(k)}]_i$ to maximize

$$\mathbb{E}_i\left[U_i\big([\mathbf{W}^{(k)}]_i\big)\right].$$

Our final assumption says that agents expect their own and other agents' initial beliefs to be correct, which we call the *bona fides* ("good faith") assumption.

Assumption 2.4.3. Agents $i = 1, \dots, n$ are *bona fide*, that is,

$$\mathbb{E}_i[b_j^k(0)] = \mu_k, \quad \text{for all } j = 1, \dots, n, \text{ and all } k = 1, 2, 3, \dots$$

Now, we derive agents' maximization problem under Assumptions 2.4.1 to 2.4.3. To this end, let X denote the random variable

$$X = \sum_{j=1}^{n} W_{ij}^{(k)} b_j^k(0). \tag{2.4.2}$$

With this notation, agents' utility maximization problems become, under our named assumptions, for each agent $i = 1, \dots, n$:

$$\begin{aligned} \max_{[\mathbf{W}^{(k)}]_i} \ & \mathbb{E}_i\left[U_i\big([\mathbf{W}^{(k)}]_i\big)\right] = \mathbb{E}_i\left[-\Big(\sum_{j=1}^{n} W_{ij}^{(k)} b_j^k(0) - \mu_k\Big)^2\right] = -\,\mathbb{E}_i\big[(X - \mathbb{E}_i[X])^2\big] \\ \text{s.t. } & W_{i1}^{(k)} + \ldots + W_{in}^{(k)} = 1, \end{aligned} \tag{2.4.3}$$

since $\mathbb{E}_i[X] = \sum_{j=1}^{n} W_{ij}^{(k)}\, \mathbb{E}_i[b_j^k(0)] = \mu_k \sum_{j=1}^{n} W_{ij}^{(k)} = \mu_k$ and where we assume row-stochasticity of $\mathbf{W}^{(k)}$. Now, $\mathbb{E}_i\big[(X - \mathbb{E}_i[X])^2\big] = \mathrm{Var}_i[X]$ and hence, agents' utility maximization problems may be rewritten as

$$\begin{aligned} \max_{[\mathbf{W}^{(k)}]_i} \ & -\mathrm{Var}_i[X] = \min_{[\mathbf{W}^{(k)}]_i} \mathrm{Var}_i[X] \\ \text{s.t. } & W_{i1}^{(k)} + \ldots + W_{in}^{(k)} = 1, \end{aligned} \tag{2.4.4}$$

that is, agents strive to set weights $W_{i1}^{(k)}, \dots, W_{in}^{(k)}$ such that $\mathrm{Var}_i[X]$ is minimized subject to the row-stochasticity condition on $\mathbf{W}^{(k)}$. To simplify the solution to problem (2.4.4), we additionally assume independence of $b_1^k(0), \dots, b_n^k(0)$.

Assumption 2.4.4. The variables $b_1^k(0), \dots, b_n^k(0)$ are *independent* random variables.

Finally, we assume that agents expect the variables $b_j^1(0), b_j^2(0), b_j^3(0), \dots$ to be independent with *identical* variances. If this were not the case, agents' reliability across topics would vary so that statistical regularities — inference from past performance to current performance — could not be exploited.

Assumption 2.4.5. Each agent $i \in [n]$ expects the random variables $b_j^1(0), b_j^2(0), b_j^3(0), \dots$ to be independent random variables with identical variances, that is,

$$\mathrm{Var}_i[b_j^1(0)] = \mathrm{Var}_i[b_j^2(0)] = \mathrm{Var}_i[b_j^3(0)] = \cdots$$

for all $j = 1, \dots, n$.

Under Assumptions 2.4.4 and 2.4.5, $\mathrm{Var}_i[X]$ may be written as

$$\mathrm{Var}_i[X] = \sum_{j=1}^{n} (W_{ij}^{(k)})^2\, \mathrm{Var}_i[b_j^k(0)] = \sum_{j=1}^{n} \alpha_{ij}^2 \sigma_{ij}^2,$$

where we let, for short, $\alpha_{ij} = W_{ij}^{(k)}$ (here, we may omit the dependence on k since $W_{ij}^{(k)}$ are optimization variables that do not, intrinsically, depend on topic X_k) and $\sigma_{ij}^2 = \mathrm{Var}_i[b_j^k(0)]$ (here, we may omit the dependence on k due to Assumption 2.4.5). Thus, to solve problem (2.4.4) under Assumptions 2.4.4 and 2.4.5, each agent $i = 1, \dots, n$ minimizes the 'Lagrange' function

$$\mathcal{L}_i(\alpha_{i1}, \dots, \alpha_{in}) = \sum_{j=1}^{n} \alpha_{ij}^2 \sigma_{ij}^2 - \lambda(\sum_{j=1}^{n} \alpha_{ij} - 1) \tag{2.4.5}$$

for some 'Lagrange multiplier' λ. Via the first-order conditions, this leads to

$$\alpha_{ij} = \frac{\lambda}{2\sigma_{ij}^2},$$

and from $\sum_{j=1}^{n} \alpha_{ij} = 1$, we find that

$$\sum_{j=1}^{n} \frac{\lambda}{2\sigma_{ij}^2} = 1 \quad \iff \quad \lambda = \frac{2}{\sum_{j=1}^{n} \sigma_{ij}^{-2}}.$$

Thus, under Assumptions 2.4.1 to 2.4.5, a rational agent chooses weights $W_{ij}^{(k)}$ that satisfy

$$W_{ij}^{(k)} = \alpha_{ij} = \frac{\frac{1}{\mathrm{Var}_i[b_j^k(0)]}}{\sum_{j'=1}^{n} \big(\mathrm{Var}_i[b_{j'}^k(0)]\big)^{-1}} \propto \frac{1}{\mathrm{Var}_i[b_j^k(0)]}, \tag{2.4.6}$$

which is quite an intuitive result: the larger the variance of agent j's estimate $b_j^k(0)$ — or, more precisely, what agent i thinks of this variance to be — the lower should the weight be that agent i assigns j, since j's initial belief tends to be 'away from truth' more frequently — or, more precisely, i expects j's initial belief to be so.

A comparison with the heuristic weight adjustment rule (2.3.2)

To compare the 'optimal' weight adjustment rule under Assumptions 2.4.1 to 2.4.5 with the heuristic rule (2.3.2), note first that weight adjustment rule (2.3.2) amounts to (weighted) 'counting' of how often a particular agent j has been in an η interval around truth μ_k, since, each time j has been within this interval, the weight of i for j is increased by the term $\delta \cdot T(\cdot)$. Hence, denoting the weights defined via rule (2.3.2) by $\tilde{W}_{ij}^{(k)}$ for the moment and the remainder of this section, we have

$$\tilde{W}_{ij}^{(k)} \propto R_j^k(\eta),$$

where $R_j^k(\eta)$ is the number of times agent j has been in an η-interval around truth within the first k discussion topics,

$$R_j^k(\eta) = \big|\{h \in \{1, \ldots, k\} \mid \big\|b_j^h(\tau) - \mu_k\big\| < \eta\}\big|.$$

Now, if Assumptions 2.4.2, 2.4.3, 2.4.4, and 2.4.5 hold and if $\tau = 0$, then clearly, $R_j^k(\eta)$ is inversely related to $\mathrm{Var}_i[b_j^k(0)]$, for all $j = 1, \ldots, n$, since if $R_j^k(\eta)$ is low, then i thinks that j has high variance (around j's presumed expected value of $\mathbb{E}_i[b_j^k(0)] = \mu_k$) and analogously if $R_j^k(\eta)$ is high. Hence, under these assumptions, weight adjustment rule (2.3.2) entails weights $\tilde{W}_{ij}^{(k)}$ which satisfy

$$\tilde{W}_{ij}^{(k)} \propto \frac{1}{\mathrm{Var}_i[b_j^k(0)]}.$$

Thus, to summarize, if

- Assumptions 2.4.1 to 2.4.5 hold and if,
- $\tau = 0$ (adjusting based on initial beliefs),

then, heuristic weight adjustment rule (2.3.2) corresponds, by analogy, to an adjustment rule that a rational agent would implement, under the named assumptions.

Discussion

Some of the assumptions we have made require further discussion. Assumption 2.4.1, which says that agents have limited foresight and want to minimize the distance between $b_i^k(1)$ and μ_k, rather than between $b_i^k(\infty)$ and μ_k, may not only be perceived as the choice of a boundedly rational agent. In contrast, if agent i knows, or at least assumes, that all agents are similarly rational as her (and share the same information structure, etc., that is, are perfectly homogeneous) — whence all agents are faced with the same optimization problems to which they derive identical solutions $[\mathbf{W}^{(k)}]_i$ — then, in fact, $\mathbf{b}^k(1) = \mathbf{b}^k(\infty)$ since $\mathbf{W}^{(k)}$, for each k, is identical in each row and is row-stochastic (see Lemma 2.A.1 in Appendix 2.A). So, under this prerequisite, agents could also be thought of as having 'perfect foresight', knowing that $b_i^k(1)$ will equal $b_i^k(\infty)$ anyways. Assumption 2.4.2 is innocuous, while Assumption 2.4.3 is the bona fides assumption discussed in the introduction, which we thought of as being based on egocentric biases. Next, Assumption 2.4.4, that agents' initial beliefs are independent, is highly implausible, of course: individuals go to the same or similar schools, are influenced by the same or similar media, etc., all of which may induce correlation in individuals beliefs (possibly even if we think of these beliefs as prior to social communication); we make this assumption for technical ease, as otherwise deriving closed-form solutions to the optimization problems in question may be quite challenging. Finally, Assumption 2.4.5 demands that topics are of the same general 'area', as we have indicated in Section 2.3, whence one may expect individuals' reliability (for this field of human expertise) to be predictable across a multitude of topics. Forfeiting the assumption would mean to present agents with a problem where nothing can be learned, in terms of adjusting the network structure $\mathbf{W}$, across various topics.

We also mention that our above analysis has assumed that $\delta \cdot T$ is strictly positive (for all or at least infinitely many topics X_k), for, e.g., otherwise $\tilde{W}_{ij}^{(k)}$ would not be proportional to $R_j^k(\eta)$. In Section 2.6, we also consider the case when $\delta \cdot T$ is zero for all but finitely many topics. We treat this case, which allows us to derive wisdom results in certain circumstances even under the presence of biased agents, as a special (or, 'extreme') case of our model that differs, however, from the choice a rational agent would pursue, as we have sketched.

Illustration

To illustrate the relationship between $W_{ij}^{(k)}$ as set by a (boundedly) rational agent and as set via (heuristic) weight adjustment rule (2.3.2), consider the following exemplary situation. Let there be n agents, all of whose initial beliefs $b_i^k(0)$ are *normally* distributed around μ_k, for all topics X_k. Assume that there are two types of agents, L and H, with variances σ_L^2 and σ_H^2, respectively, such that $\sigma_L^2 < \sigma_H^2$. Let there be n_L agents of type L and n_H agents of type H such that $n_L + n_H = n$. In other words, for each L-type agent i_L, we have, for all $k = 1, 2, 3, \ldots$,

$$b_{i_L}^k(0) \sim \mathtt{N}(\mu_k, \sigma_L^2),$$

and, accordingly, for each H-type agent i_H, we have

$$b_{i_H}^k(0) \sim \mathtt{N}(\mu_k, \sigma_H^2).$$

Thus, under Assumptions 2.4.1 to 2.4.5, a rational agent i would assign weights,

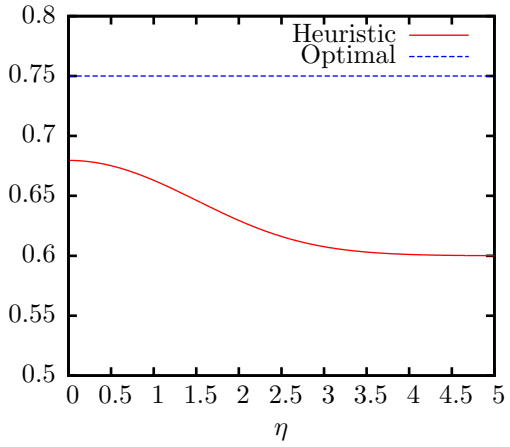
$$W_{ij}^{(k)} = \frac{1}{\sigma_T^2} \cdot \frac{1}{C}, \tag{2.4.7}$$

where $T \in \{L, H\}$, depending on whether j is of type L or H, and where C is the constant $C = \frac{n_L}{\sigma_L^2} + \frac{n_H}{\sigma_H^2}$. In contrast, an agent who sets weights according to the rule (2.3.2), would set

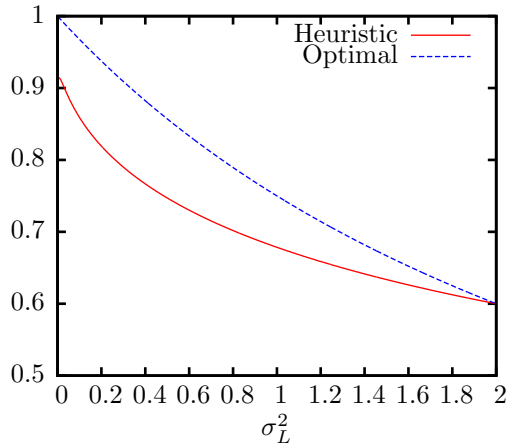
$$\tilde{W}_{ij}^{(k)} \propto \Pr[\left\|b_j^k(0) - \mu_k\right\| < \eta] = \int_{-\eta}^{\eta} \frac{1}{\sqrt{2\pi\sigma_T^2}} \exp(-\frac{x^2}{2\sigma_T^2})\, dx = 2\int_0^{\eta} \frac{1}{\sqrt{2\pi\sigma_T^2}} \exp(-\frac{x^2}{2\sigma_T^2})\, dx, \tag{2.4.8}$$

depending on whether j is of type $T = L$ or $T = H$. In Figure 2.3, we plot the behavior of (2.4.7) vs. (2.4.8) for specific values of σ_L^2 and σ_H^2, namely $\sigma_L^2 = 1$ and $\sigma_H^2 = 2$. For the values of σ_L^2 and σ_H^2

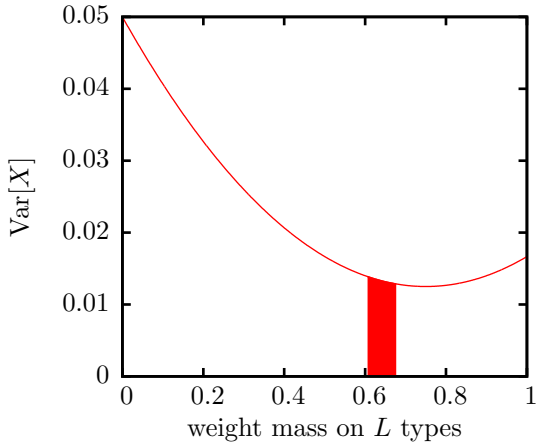
discussed, the optimal rule under Assumptions 2.4.1 to 2.4.5 would accord total weight mass for L-types of $n_L \cdot W_{ij}^{(k)} = \frac{3}{4}$ (for j of type L), and total weight mass for H-types of $n_H W_{ij}^{(k)} = \frac{1}{4}$ (for j of type H). In contrast, as the graphs show, if weights are set according to (2.4.8), then, $\tilde{W}_{ij}^{(k)}$ is, for the L types,



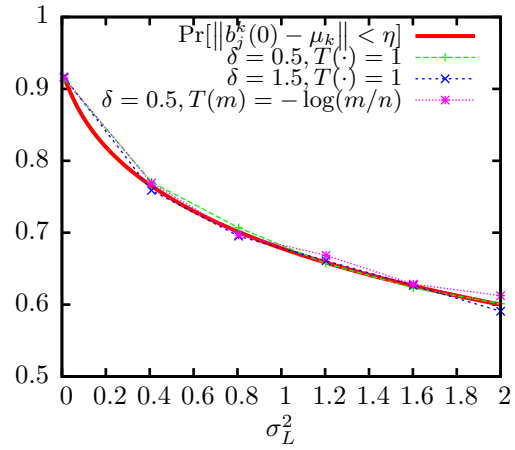
(a) Weight mass for L types; optimal vs. heuristic, $\tilde{W}_{ij}^{(k)}$, as a function of η; $\sigma_L^2 = 1$ and $\sigma_H^2 = 2$ fixed.



(b) Weight mass for L types; optimal vs. heuristic, $\tilde{W}_{ij}^{(k)}$, as a function of σ_L^2; $\eta = 0.25$ fixed.



(c) Variance of X as defined in (2.4.2) as a function of weight mass assigned to L types. The colored area gives the range of $\tilde{W}_{ij}^{(k)}$ as illustrated in (a).



(d) Weight mass for L types as theoretically computed according to (2.4.8) (as in (b)) as a function of σ_L^2, and, as a comparison, as given by sample runs, for two different values of $\delta = 0.5, 1.5$, and functions T ($T \equiv 1$ and $T(m) = -\log(m/n)$); averages; $\eta = 0.25$ fixed.

Figure 2.3: Throughout $\sigma_H^2 = 2$ and $n = 100 = 60 + 40 = n_L + n_H$.

always lower than $\frac{3}{4}$, as the optimal rule would prescribe. Depending on η, $\tilde{W}_{ij}^{(k)}$ ranges from 0.60, if η is large, to about 0.68, for small η. The value for η large is obvious since if η is sufficiently large in size, then each agent will receive identical weight $\tilde{W}_{ij}^{(k)}$, $\frac{1}{n}$, and, hence, total weight mass for T-types is $\frac{n_T}{n}$, for $T \in \{L, H\}$. The figure also shows the inverse relationship between $\tilde{W}_{ij}^{(k)}$ and σ_T^2 (for $T = L$, in this case; Figure 2.3 (b)), the closeness of $\tilde{W}_{ij}^{(k)}$ to 'optimality' (Figure 2.3 (c)), and a comparison between the theoretic value $\tilde{W}_{ij}^{(k)}$ is proportional to, $\Pr[\|b_j^k(0) - \mu_k\| < \eta]$, and actual realizations of $\tilde{W}_{ij}^{(k)}$ as a function of δ and T (Figure 2.3 (d); cf. Equation (2.3.2)).

2.5 Notation and definitions

We introduce the following helpful notation and definitions.

Definition 2.5.1. Let any $\epsilon \geq 0$ be fixed. We call an agent i *ϵ-intelligent for (topic) X_k* if i's initial belief on X_k is (ϵ) 'close to truth', i.e., $\left\|b_i^k(0) - \mu_k\right\| < \epsilon$. We call i *ϵ-intelligent*, if i is intelligent for all topics X_k.

This definition captures the idea that an agent's initial beliefs, which we think of as not influenced by peers (or their beliefs), express something *innate* to agent i, his *hidden ability* or, simply, *intelligence*. However, we say nothing here on how i has arrived at his initial beliefs, e.g., whether it was through hidden ability in a proper sense or, for instance, 'merely' through guessing. We also remark that the concept of ϵ-intelligence (or ϵ-wisdom, as we define below) is clearly related to our weight adjustment rule; in particular, for given tolerance η, agents increase their weight for an agent j if this agent is ϵ-intelligent (or ϵ-wise) for a topic X_k and for all $\epsilon \leq \eta$.

When i is 'close to truth' in the limit of the DeGroot learning process, we call i *wise*.

Definition 2.5.2. We call an agent i *ϵ-wise for (topic) X_k* if i's limit belief on X_k is 'close to truth', i.e., $\left\|b_i^k(\infty) - \mu_k\right\| < \epsilon$. We call i *ϵ-wise*, if i is wise for all topics X_k.

We also introduce stochastic analogues of the above definitions. If an agent has initial beliefs stochastically centered around truth for a topic, we call the agent *stochastically intelligent for this topic*.

Definition 2.5.3. We call an agent i *stochastically intelligent for (topic) X_k* if i's initial belief on X_k is 'stochastically centered around truth', i.e., $b_i^k(0) = \mu_k + \sigma_{ik}$, where σ_{ik} is some individual and topic-specific white-noise variable. We call i *stochastically intelligent*, if i is stochastically intelligent for all topics X_k.

We omit the corresponding definition for wisdom since we rarely make use of a concept of 'stochastic wisdom' in the remainder of this work.

Next, fix a level of intelligence or wisdom $\epsilon \geq 0$. For convenience, let us denote the open ϵ-interval around truth, within with agents are considered ϵ-intelligent (or ϵ-wise), by $B_{k,\epsilon}$ and its complement by $B_{k,\epsilon}^c$. Formally, we have:

Definition 2.5.4.

$$B_{k,\epsilon} := (\mu_k - \epsilon, \mu_k + \epsilon),$$
$$B_{k,\epsilon}^c = S \backslash B_{k,\epsilon}.$$

Below, in the main sections of our work, our principal modeling perspective — although we may occasionally deviate from or slightly generalize this perspective — is the notion of two groups of agents, $\mathcal{N}_1$ and $\mathcal{N}_2$ with $\mathcal{N}_1 \cup \mathcal{N}_2 = [n]$ and $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$, one of whose initial beliefs are *unbiased* — group $\mathcal{N}_1$'s — and the other's initial beliefs are *biased*, whereby we define bias as

$$\beta_{i,k} = \left\|\mathbb{E}[b_i^k(0)] - \mu_k\right\|.$$

Hence, for members i of $\mathcal{N}_1$, we assume that $\beta_{i,k} = 0$ and for members i of $\mathcal{N}_2$, we assume that $\beta_{i,k} > 0$ for all topics X_k. In addition, we think of the two groups of agents as having independent and identical distributions of initial beliefs, with distribution functions $F_{\mathcal{N}_l,k}(A) = \Pr[b_i^k(0) \in A]$, for $l = 1, 2$ and $A \subseteq S$, where, of course, identical distribution refers to within group and independence refers to both within and across group relations. Finally, for fixed level of tolerance $\eta \geq 0$, we assume that $F_{\mathcal{N}_l,k}(B_{k,\eta})$ does not depend upon k, that is, $F_{\mathcal{N}_l,k}(B_{k,\eta}) = F_{\mathcal{N}_l,k'}(B_{k',\eta})$, for all k, k'. This means that agents' probability of being within an η-interval around truth — for initial beliefs — is the same across topics. This assumption is very similar, in spirit, to Assumption 2.4.5 and captures predictability of agents. We also think of this invariant probability as denoting an agent's ability or reliability.

To conclude this section, we introduce notation regarding convergence (and consensus) of our endogenous opinion dynamics paradigm.

Definition 2.5.5. Let $k \geq 1$ be arbitrary. We say that $\mathbf{W}^{(k)}$ is *convergent for opinion vector* $\mathbf{b}(0) \in S^n$ if $\lim_{t\to\infty}(\mathbf{W}^{(k)})^t\mathbf{b}(0)$ exists. Moreover, we say that $\mathbf{W}^{(k)}$ *induces a consensus for opinion vector* $\mathbf{b}(0)$ if $\mathbf{W}^{(k)}$ is convergent for $\mathbf{b}(0)$ and $\lim_{t\to\infty}(\mathbf{W}^{(k)})^t\mathbf{b}(0)$ is a *consensus*, that is, a vector $\mathbf{c} \in S^n$ with all entries identical.

Rather than saying that $\mathbf{W}^{(k)}$ converges, we may occasionally also say that beliefs converge (under $\mathbf{W}^{(k)}$) or that our DeGroot learning / opinion dynamics paradigm converges. We also mention that we typically assume matrix $\mathbf{W}^{(1)}$ to be the $n\times n$ identity matrix (in the absence of further information, agents follow their own signals), which sometimes facilitates analytical derivations, but we also consider more general forms of the matrix $\mathbf{W}^{(1)}$, where we find that such a generalization is worthwhile mentioning.

Throughout our work, we assume that weight matrices $\mathbf{W}^{(k)}$ are *row-stochastic*, that is,

$$\sum_{j=1}^{n}[\mathbf{W}^{(k)}]_{ij} = 1,$$

for all $i \in [n]$. We denote the entries of an arbitrary matrix $\mathbf{A}$ by A_{ij} or $[\mathbf{A}]_{ij}$. We denote by $\mathbf{I}_n$ the $n \times n$ identity matrix and by $\mathbb{1}_n$ the vector of n 1's, i.e., $\mathbb{1}_n = (1, \ldots, 1)^\intercal$. We may omit the dimensionality if it is clear from the context.

2.6 The standard DeGroot model

In the subsequent sections, we derive a few results regarding the standard DeGroot learning model under our endogenous weight formation paradigm. First, we show that, in our setup, agents almost always reach a consensus (Proposition 2.6.1 and the subsequent remark), that is, for almost all topics X_k, under very mild conditions. Then, in Section 2.6.1, we show that if agents are unbiased and receive initial belief signals that are centered around truth, then agents' beliefs converge to truth for topics X_k, as $n, k \to \infty$, irrespective of whether agents adjust weights based on limiting or on initial beliefs. Next, in Section 2.6.2, we illustrate that agents may be arbitrarily far off from truth as the number of biased agents involved in the opinion dynamics process becomes large, thus demonstrating that crowd wisdom may fail under these circumstances. For the situation when $T(n) = 0$, we also give sufficient conditions on when crowd wisdom does not fail, even under the presence of biased agents. In Section 2.6.3, we discuss weights on own beliefs as a (simple) extension of the classical DeGroot learning paradigm and as discussed by DeMarzo, Vayanos, and Zwiebel (2003).

We start our discussion with a theorem given in the original DeGroot paper (DeGroot, 1974), which helps us determining when our endogenous opinion dynamics process leads agents to a consensus.

Theorem 2.6.1. If there exists a positive integer t such that every element in at least one column of the matrix $\mathbf{W}^t$ is positive, then $\mathbf{W}$ induces a consensus for any vector $\mathbf{b}(0) \in S^n$.

Theorem 2.6.1 can be used in a straightforward manner to derive conditions, in our setup, under which agents reach a consensus. Namely, during the course of dicussing issues $X_1, X_2, X_3, \ldots$, as long as no agent has been η-intelligent (resp. η-wise), agents do not adjust their weights to other agents, and, consequently, agents reach a consensus if and only if $\mathbf{W}^{(1)}$ induces a consensus. At the first time point that some agent has been η-intelligent (resp. η-wise), all agents subsequently adjust weights for this agent, and, hence, (at least) one column of the respective weight matrix is strictly positive for the subsequent topic. Hence, for this topic, all agents reach a consensus. But note that this column remains positive for *all* weight matrices corresponding to discussion topics discussed thereafter (as can easily be shown inductively) because even redistribution of weight mass to other agents, via weight normalization, cannot make a matrix entry zero once it has been positive. Now, we formalize these simple ideas. Then, we generalize to the setting when agents have individualized tolerances η_i.

Let A_i be the set of time points agent i is η-intelligent (resp. η-wise) for some topic X_k,

$$A_i = \{k \in \mathbb{N} \mid \left\|b_i^k(\tau) - \mu_k\right\| < \eta\} \subseteq \mathbb{N},$$

where $\tau = 0$ (resp. $\tau = \infty$) and let a_i be the first time that i is η-intelligent for some topic X_k,

$$a_i = \min A_i.$$

Then, we have the following proposition, for which we assume that $T(\cdot) > 0$ on its whole domain. This assumption is innocuous here; if it does not hold, the proposition may easily be adjusted to account for the different setup.

Proposition 2.6.1. Let $\eta \geq 0$ be fixed. Let $\tau = 0$ (resp. $\tau = \infty$). Let $r = \min_{i\in[n]} a_i$ be the earliest time point that some agent is η-intelligent (resp. η-wise) for topic X_r. (a) Then agents reach a consensus for all topics X_k with $k > r$, independent of their initial beliefs. (b) For topics $1, \ldots, r$, agents reach a consensus if and only if $\mathbf{W}^{(1)}$ induces a consensus.

Proof. (a) By the proposition, we know that some agent i is η-intelligent (resp. η-wise) for topic X_r. Accordingly, agents increase their weight to i by $\delta \cdot T(\cdot) > 0$ at time $r+1$. Hence, weight matrix $\mathbf{W}^{(r+1)}$ has a strictly positive column and so do, in general, have all matrices $\mathbf{W}^{(k)}$, for $k > r$. By Theorem 2.6.1, agents thus reach a consensus for all issues X_k, with $k > r$.

(b) For issues $X_1, \ldots, X_r$, no weight adjustments are made, whence $\mathbf{W}^{(1)} = \cdots = \mathbf{W}^{(r)}$ and a consensus is reached if and only if $\mathbf{W}^{(1)}$ induces a consensus. □

Remark 2.6.1. Assume, for the moment, that agents have individualized tolerances η_i. Then part (a) of Proposition 2.6.1 is true if we replace A_i as above by

$$A_i = \{k \in \mathbb{N} \mid \left\|b_i^k(\tau) - \mu_k\right\| < \min_{j\in[n]} \eta_j\},$$

and we define a_i as above as $a_i = \min A_i$.

Remark 2.6.2. Consider $\tau = 0$ for this remark. If initial beliefs are random variables, then r, as specified in Proposition 2.6.1, is a random variable (which we could consider a 'stopping time'). Accordingly, its distribution might be of interest. Assuming agent i's initial opinions for each topic X_k to be distributed with distribution function $F_{i,k}$, that is, $P[b_i^k(0) \in A] = F_{i,k}(A)$, for $A \subseteq S$, we have that the probability that at least one agent i is η-intelligent for topic X_k is given by $p_{k,\eta} = 1 - \prod_{i\in[n]} F_{i,k}(B_{k,\eta}^c)$, due to independence of agents' initial beliefs. Then, if $F_{i,k}(B_{k,\eta}^c)$ does not depend on X_k but only on η, we have that r has a geometric distribution with probability p_η (where we omit, in the notation, the dependence on k due to our assumption), that is,

$$P[r = \nu] = (1 - p_\eta)^{\nu-1} p_\eta, \quad \text{for } \nu = 1, 2, 3, \ldots$$

From the specification of p_η, we thus see that if $F_{i,k}(B_{k,\eta}^c) < 1$ for all i, then $p_\eta \to 1$ as $n \to \infty$. Accordingly, the distribution of r converges to the degenerate distribution with $P[r = 1] = 1$ and $P[r \neq 1] = 0$. Thus, in this situation, agents 'almost always' — that is, with possibly only finitely many, namely, one, exceptions, topic X_1 — reach a consensus for topics X_k, for $k = 1, 2, 3, \ldots$.

We also find the next simple result which states that if *all* agents start with initial beliefs within a precision of ϵ around truth, then agents will also end up with limiting beliefs with level of wisdom of ϵ, provided that agents' beliefs convergence at all, as time goes to infinity.

Proposition 2.6.2. Let level of intelligence $\epsilon \geq 0$ be fixed. If all agents are ϵ-intelligent for X_k and the DeGroot learning process (2.3.1) converges, then all are ϵ-wise for topic X_k.

Proof. This simply follows from the fact that the interval $B_{k,\epsilon} = (\mu_k - \epsilon, \mu_k + \epsilon)$ is a convex set and weights are always row-stochastic in our model setup. Thus, if all agents start their beliefs in $B_{k,\epsilon}$, limit beliefs will also be in $B_{k,\epsilon}$, provided that they converge. □

As we have seen in Proposition 2.6.1, whether or not the DeGroot learning process (2.3.1) converges on the first r topics depends on the initial weight matrix $\mathbf{W}^{(1)}$. Thereafter, convergence (even to consensus) is guaranteed. Hence, using Proposition 2.6.2, we obtain:

Corollary 2.6.1. Let level of intelligence $\epsilon \geq 0$ be fixed. If all agents are ϵ-intelligent (i.e., for all topics X_k), then all agents are ϵ-wise for all topics X_k, with $k > r$, where r is defined as in Proposition 2.6.1.

2.6.1 Unbiased agents

In this setup, we assume that *all* agents receive initial signals

$$b_i^k(0) = \mu_k + \epsilon_{ik}, \tag{2.6.1}$$

where μ_k is truth for issue X_k and ϵ_{ik} is white noise (i.e., with mean zero and independent of other variables) with variance $\sigma_i^2 = \text{Var}[\epsilon_{ik}]$ (note that we assume the variance to be independent of the issue X_k). As throughout, we assume agents' initial signals to be independent.

We consider first the situation when agents adjust weights based on limiting beliefs, i.e., $\tau = \infty$. In the next proposition, we show that agents become ϵ-wise in this situation (for any $\epsilon > 0$), in the limit as both n, population size, and k, which indexes topics, go to infinity. The intuition behind this result is simple: since, in our setup, agents tend toward a consensus (see Proposition 2.6.1), agents will generally all be *jointly* η-wise (where η is agents' tolerance) or not. Then, since agents adjust based on limiting beliefs, agents receive the same increments (or not) to their weight structure, so that, as k becomes large, $\mathbf{W}^{(k)}$ is the matrix with entries $\frac{1}{n}$, approximately. Then, the law of large number implies convergence to truth, as n becomes large, since initial beliefs are stochastically centered around truth by (2.6.1).

Proposition 2.6.3. Let $\eta \geq 0$ be fixed. Assume that agents' initial beliefs are centered around truth in the form (2.6.1). Moreover, assume that agents initially follow their own beliefs, that is, $\mathbf{W}^{(1)}$ is the $n \times n$ identity matrix $\mathbf{I}_n$. Finally, assume that agents adjust weights based on limiting beliefs, i.e., $\tau = \infty$. Let $T(\cdot) > 0$. Then, as $k, n \to \infty$, all agents become ϵ-wise for topics X_k, for all $\epsilon > 0$, almost surely.

Proof. As before, let $r = \min_{i \in [n]} a_i$ be the first time point that one agent is η-intelligent for topic X_r (which is the same as η-wise for topic X_r, as $\mathbf{W}^{(1)}$ is the $n \times n$ identity matrix, by assumption). For simplicity, assume first that, for topic X_r, *all* agents are η-intelligent (and hence, η-wise); we then treat the more general case where only some agents are η-intelligent for X_r as an analogous situation. In this case, $\mathbf{W}^{(r+1)}$ looks as follows, after weight adjustments,

$$\mathbf{W}^{(r+1)} = \frac{1}{1 + n\tilde{\delta}} \begin{pmatrix} 1+\tilde{\delta} & \tilde{\delta} & \dots & \tilde{\delta} \\ \tilde{\delta} & 1+\tilde{\delta} & \dots & \tilde{\delta} \\ \vdots & \dots & \ddots & \vdots \\ \tilde{\delta} & \tilde{\delta} & \dots & 1+\tilde{\delta} \end{pmatrix},$$

where we let $\tilde{\delta} = \delta \cdot T(\cdot)$. Consider any matrix $\mathbf{A}$ of the form

$$\mathbf{A} = \begin{pmatrix} \beta & \alpha & \dots & \alpha \\ \alpha & \beta & \dots & \alpha \\ \vdots & \dots & \ddots & \vdots \\ \alpha & \alpha & \dots & \beta \end{pmatrix} \tag{2.6.2}$$

such that $\beta + (n-1)\alpha = 1$ (that is, $\mathbf{A}$ is row-stochastic), with $0 < \alpha, \beta < 1$. In Appendix 2.A, we show that matrix $\mathbf{A}$ has one eigenvalue $\lambda = 1$, to which corresponds an eigenvector $\mathbf{c} = (c, \dots, c)^\intercal$, and $(n-1)$ identical eigenvalues of absolute size smaller than 1. Moreover, since $\mathbf{A}$ is symmetric, it is diagonalizable of the form $\mathbf{A} = \mathbf{U}\mathbf{V}\mathbf{U}^\intercal$, where $\mathbf{V}$ is a diagonal matrix that contains the eigenvalues of $\mathbf{A}$ on the diagonal and $\mathbf{U}$ is orthonormal, that is, $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_n$; without loss of generality, assume that the eigenvalues in $\mathbf{V}$ are arranged by size, i.e., $V_{11} = 1 > V_{22} = \dots = V_{nn}$ and the corresponding eigenvectors are located in the respective columns of $\mathbf{U}$, i.e., the first column of $\mathbf{U}$ is the vector $\mathbf{c}$. We have

$$\mathbf{A}^t = \mathbf{U}\mathbf{V}^t\mathbf{U}^\intercal.$$

As $t \to \infty$, $\mathbf{V}$ converges to the matrix with one entry equal to 1 and all other entries equal to zero (due to the eigenvalue structure of $\mathbf{A}$). Thus, we then have

$$\lim_{t\to\infty} \mathbf{A}^t = \begin{bmatrix} \mathbf{c} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} \mathbf{U}^\intercal = \begin{bmatrix} \mathbf{c} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}^\intercal \\ \mathbf{c}_2^\intercal \\ \vdots \\ \mathbf{c}_n^\intercal \end{bmatrix} = c^2 \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix},$$

where $\mathbf{c}_2, \ldots, \mathbf{c}_n$ are the eigenvectors corresponding to eigenvalues λ_2 to λ_n. Moreover, since $\mathbf{A}$ is row-stochastic, $\mathbf{A}^t$ is row-stochastic for every t, and, accordingly, $\lim_t A^t$ is row-stochastic. Therefore $c^2 = \frac{1}{n}$. In other words, if each agent is η-wise for topic X_r, then for topic X_{r+1}, we have

$$\mathbf{b}^{r+1}(\infty) = \lim_{t\to\infty}(\mathbf{W}^{(r+1)})^t \mathbf{b}^{r+1}(0) = \begin{pmatrix} \sum_{j=1}^{n} \frac{1}{n} b_j^{r+1}(0) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{\sum_{j=1}^{n} b_j^{r+1}(0)}{n} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{2.6.3}$$

Now, for all topics X_k, with $k > r$, agents reach a consensus by Proposition 2.6.1. Hence, agents are either all jointly η-wise or none of them is, for all $k > r$. Therefore, all weight matrices $\mathbf{W}^{(k)}$, for $k > r$, have the form (2.6.2) (either all entries receive an increment of $\tilde{\delta}$ and are then renormalized, or none receives an increment). Hence, agents' limiting beliefs are always weighted averages of their initial beliefs, where the weights are $\frac{1}{n}$. Applying the law of large numbers then implies that agents become ϵ-wise as $n \to \infty$ for any $\epsilon > 0$ almost surely, for all $k > r$.

For the more general case when not all agents are η-wise for topic X_r, one can show that agents' limiting beliefs for topic X_{r+1} are (uniform) averages of the initial beliefs of the agents who were η-wise for X_r, rather than averages of all agents' initial beliefs. As topics progress, either all agents are jointly η-wise or they are not (since agents always reach a consensus for topics X_k, with $k > r$). Hence, since agents adjust weights based on limiting beliefs, the entries in the weight matrices $\mathbf{W}^{(k)}$ all either receive jointly an increment of $\tilde{\delta}$ or not (in fact, increments of $\tilde{\delta}$ are added infinitely often, almost surely, as $k \to \infty$ since initial beliefs are centered around truth). Hence, $\mathbf{W}^{(k)}$ tends toward a matrix with all entries $\frac{1}{n}$ as $k \to \infty$ and the law of large numbers takes care for almost sure convergence. □

Next, we state that Proposition (2.6.3) holds true also if agents adjust weights based on initial beliefs. This is understandable: if agents adjust weights based on limiting beliefs, weights converge to $\frac{1}{n}$ as k increases. However, this weighting structure is not optimal, as it ignores the different variances of agents' initial beliefs, but agents' final beliefs still converge to truth in the limit. Hence, if agents set weights 'closer to optimality' as they do when they adjust based on initial beliefs (cf. Section 2.4), they should certainly also converge to truth. We prove the proposition more formally by referring, in Appendix 2.A, to results developed in Golub and Jackson (2010), which generalize the 'ordinary' law of large numbers.

Proposition 2.6.4. Let $\eta \geq 0$ be fixed. Assume that agents' initial beliefs are centered around truth in the form (2.6.1). Moreover, assume that agents initially follow their own beliefs, that is, $\mathbf{W}^{(1)}$ is the $n \times n$ identity matrix $\mathbf{I}_n$. Finally, assume that agents adjust weights based on initial beliefs, i.e., $\tau = 0$. Let $T(\cdot) > 0$. Then, as $k, n \to \infty$, all agents become ϵ-wise for topics X_k, for all $\epsilon > 0$, almost surely.

2.6.2 Biased agents

The case $T(n) = 0$

In the biased agent setup, we start with the following conditions. Fix a level of wisdom $\epsilon > 0$, with $\epsilon \leq \eta$, agents' tolerance. Let there be $n = n_1 + n_2$ agents, and denote by $\mathcal{N}_1$ and $\mathcal{N}_2$ the respective agent sets such that $[n] = \mathcal{N}_1 \cup \mathcal{N}_2$. The agents in $\mathcal{N}_1$ are ϵ-intelligent and we think of them as having unbiased initial beliefs about any topic X_k; in particular, we think of their initial beliefs as distributed according to $\mu_k + \epsilon_{ik}$, where ϵ_{ik} is white noise, appropriately restricted such that $\mu_k + \epsilon_{ik} \in B_{k,\epsilon}$. Conversely, let the n_2 agents in $\mathcal{N}_2$ have initial beliefs distributed according to a random variable Z_k (that depends on topic X_k) with distribution function $F_{Z_k}(A) = P[b_i^k(0) \in A]$, for $A \subseteq S$ (in particular, agents in $\mathcal{N}_2$ all have the same distribution of initial beliefs). Assume that $F_{Z_k}(A) > 0$ for all non-empty intervals $A \subseteq S$. We think of the agents in $\mathcal{N}_2$ as biased in that it holds that $\beta_k = \|\mathbb{E}[Z_k] - \mu_k\| > 0$ for all topics X_k. Finally, assume that $T(m) > 0$ for all $m < n$ and $T(n) = 0$ and let $\mathbf{W}^{(1)}$ be the $n \times n$ identity matrix. For short, we will also refer to the n_2 agents in $\mathcal{N}_2$ as 'biased' agents.

Our first result, concerning weight adjustment at $\tau = \infty$, states that agents' limiting beliefs, in expectation, in this context will be a mixture of truth μ_k and $\mathbb{E}[Z_k]$ unless no biased agent 'guesses' truth for topic X_1, the first topic to be discussed, in which case all agents reach level of wisdom ϵ for all topics X_k. In other words, if a biased agent is true for the initial topic X_1, then agents will always

mix truth with a biased variable. That agents do not mix when no biased agent is true for X_1 crucially depends on the condition $T(n) = 0$. Namely, if no biased agent is close enough to truth for topic X_1, only the ϵ-intelligent agents will be, such that, for topic X_2, agents only increment weights to agents in $\mathcal{N}_1$; consequently, as we show, for topic X_2, limiting consensus beliefs will be uniform means of these agents' beliefs so that all agents are ϵ-wise for X_2; but, since $T(n) = 0$, no more weight adjustments occur whatsoever, so that all agents are ϵ-wise for all topics X_k to come. We also remark that if agents' limiting beliefs are mixtures of truth and a biased variable, this does not mean that agents would not be ϵ-wise for a certain topic (which depends both on the biased agents' bias and on ϵ); it solely means that agents mix truth with something that distracts them away from truth.

For the proof of the result, we make use of the insight that if someone is wise (or intelligent) at a more refined level, he is also wise (or intelligent) at a coarser level; the following lemma, which restates this, is self-explanatory and needs no proof.

Lemma 2.6.1. Let $0 \leq \epsilon_1 \leq \epsilon_2$. If an agent i is ϵ_1-wise (ϵ_1-intelligent) for some topic X_k, then she is also ϵ_2-wise (ϵ_2-intelligent) for X_k.

In the following proposition, ϵ_1 will be ϵ, the level of wisdom to be obtained, and ϵ_2 will be η, agents' tolerance.

Proposition 2.6.5. Let the weight adjustment time point be $\tau = \infty$. Let tolerance $\eta \geq 0$ be fixed and fix a level $\epsilon \geq 0$ of wisdom, with $\epsilon \leq \eta$.

Under the outlined conditions, if $N_\eta(\mathbf{b}^1(\tau), \mu_k)$ contains only unbiased ϵ-intelligent agents — that is, $N_\eta(\mathbf{b}^1(\tau), \mu_k) = \mathcal{N}_1$ — then all agents become ϵ-wise for all topics X_k, with $k > 1$. If $N_\eta(\mathbf{b}^1(\tau), \mu_k)$ contains also agents from the set $\mathcal{N}_2$,[13] then agents' limiting beliefs, in expectation, are given by $\lambda_Z \, \mathbb{E}[Z_k] + \lambda_\mu \mu_k$, for all topics $k > 1$, where λ_Z and λ_μ are coefficients such that $\lambda_\mu = \frac{n_1}{|N_\eta(\mathbf{b}^1(\tau), \mu_k)|}$ and $\lambda_Z = \frac{\left|N_\eta(\mathbf{b}^1(\tau), \mu_k) \cap \mathcal{N}_2\right|}{|N_\eta(\mathbf{b}^1(\tau), \mu_k)|}$ so that $\lambda_\mu + \lambda_Z = 1$.

Proof. For convenience, we consider the situation when only one agent, $i = 1$, is ϵ-intelligent. The more general case is a straightforward extension of our arguments. We also assume that agent $i = 1$ holds beliefs $b_i^k(0) = \mu_k$, for all $k \geq 1$.

Let $N_\eta(\mathbf{b}^1(\tau), \mu_k)$ contain only ϵ-intelligent agents. Since $\mathbf{W}^{(1)}$ is the identity matrix, the limiting beliefs of agents $1, \ldots, n$ on topic X_1 are as follows:

$$b_1^1(\infty) = \mu_1, \; b_2^1(\infty) = b_2^1(0), \; \ldots, \; b_n^1(\infty) = b_n^1(0).$$

Moreover, since initial beliefs of the agents in $\mathcal{N}_2$ are in $B_{k,\eta}^c$, the agents in $\mathcal{N}_2$ are, consequently, also not η-wise for topic X_1, in contrast to the ϵ-intelligent agent, who is η-wise for topic X_1. Thus, the weight structure at the beginnning of discussion of topic X_2 looks as follows, after weight adjustment and renormalization

$$\mathbf{W}^{(2)} = \frac{1}{1+\tilde{\delta}} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \tilde{\delta} & 1 & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & & \\ \tilde{\delta} & 0 & 0 & \cdots & 1 \end{pmatrix};$$

recall our convention that $\tilde{\delta} = \delta \cdot T(\cdot)$. Limiting beliefs for topic X_2 are thus given by

$$\mathbf{b}^2(\infty) = \lim_{t \to \infty} (\mathbf{W}^{(2)})^t \mathbf{b}^2(0),$$

where the initial belief vector $\mathbf{b}^2(0)$ is $(\mu_2, b_2^2(0), \ldots, b_n^2(0))^\intercal$. It is not difficult to see that powers of any

[13] But not all of them. If $N_\eta(\mathbf{b}^1(\tau), \mu_k) = [n]$, then the set $N_\eta(\mathbf{b}^1(\tau), \mu_k)$ should be replaced by $N_\eta(\mathbf{b}^2(\tau), \mu_k)$, etc.

matrix with structure $\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha & 1-\alpha & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & & \\ \alpha & 0 & 0 & \cdots & 1-\alpha \end{pmatrix}$ have the form

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha & 1-\alpha & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & & \\ \alpha & 0 & 0 & \cdots & 1-\alpha \end{pmatrix}^t = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha\sum_{i=1}^{t-1}(1-\alpha)^i & (1-\alpha)^t & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & & \\ \alpha\sum_{i=1}^{t-1}(1-\alpha)^i & 0 & 0 & \cdots & (1-\alpha)^t \end{pmatrix}.$$

For $0 < \alpha \leq 1$, the right-hand side of the last equation obviously converges to the matrix with all entries identical to zero, except for the first column, which consists of n entries 1. Hence, by this fact, $\mathbf{b}^2(\infty)$ is the vector with all entries μ_2 and all agents are, consequently, ϵ-wise for topic X_2, and, thus, also η-wise (by Lemma 2.6.1). Since in this case, it holds that $\left|N_\eta(\mathbf{b}^2(\infty), \mu_2)\right| = n$, we have $T(\left|N_\eta(\mathbf{b}^2(\infty), \mu_2)\right|) = 0$ by assumption, so that agents do not adjust weights for topic X_3 (more precisely, the adjustment increment is zero). Hence, $\mathbf{W}^{(3)} = \mathbf{W}^{(2)}$, and agents will also be ϵ-wise for topic X_3 since agent $i = 1$ is ϵ-intelligent for X_3. Inductively, this holds for all X_k, with $k > 1$.

Now, assume that at least one agent in $\mathcal{N}_2$ happens to know truth for topic X_1 (that is, his initial belief is within an η radius of truth), which may always occur since $F_{Z_k}(A) > 0$ for all intervals $A \subseteq S$ by assumption. For convenience, we assume that exactly one agent in $\mathcal{N}_2$, say, agent 2, happens to know truth for topic X_1. Then, at the beginning of the discussion of topic X_2, agents increase their weights for agents 1 and 2, resulting in the following structure:

$$\mathbf{W}^{(2)} = \frac{1}{1+2\tilde{\delta}} \begin{pmatrix} 1+\tilde{\delta} & \tilde{\delta} & 0 & 0 & \cdots & 0 \\ \tilde{\delta} & 1+\tilde{\delta} & 0 & 0 & \cdots & 0 \\ \tilde{\delta} & \tilde{\delta} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \ddots & & \\ \tilde{\delta} & \tilde{\delta} & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Again, limiting beliefs for topic X_2 are then given by

$$\mathbf{b}^2(\infty) = \lim_{t\to\infty} \left(\mathbf{W}^{(2)}\right)^t \mathbf{b}^2(0).$$

It is not difficult to see that powers of matrices with structures as in the given $\mathbf{W}^{(2)}$ converge to the matrix with the first two columns being $\frac{1}{2}\mathbb{1}_n$ and the remaining columns are zero vectors. Thus, limiting beliefs of all agents are just the average of the first two agents' initial beliefs. This implies a limiting consensus such that all agents are either jointly η-wise or not η-wise for topic X_2. If all are η-wise, no weight adjustments occur for topic X_3 (since $T(n) = 0$), but if they are not η-wise, no weight adjustments occur as well (no one was right). Thus, as before, $\mathbf{W}^{(2)} = \mathbf{W}^{(3)} = \mathbf{W}^{(4)} = \cdots$, such that for all topics to come, limiting beliefs of all agents will always be averages of the first agent's (who is ϵ-intelligent) and the second agent's (who was just lucky for topic X_1) initial beliefs. Hence, in expectation, agents' limiting (consensus) beliefs will be

$$\frac{1}{2}\,\mathbb{E}[Z_k] + \frac{1}{2}\mu_k.$$

The more general forms of λ_Z and λ_μ can be straightforwardly derived in an analogous manner in the more general setting. □

Example 2.6.1. We illustrate Proposition 2.6.5 in Figure 2.4, where we let $S = [0, 1]$, $\mu_k = 0$ for all $k \geq 1$, $[n] = \{1, \ldots, 50\}$, $\mathcal{N}_1 = \{1\}$ and F_{Z_k} is the random uniform distribution on S, $\epsilon = 0$ and $\eta = 0.2$.

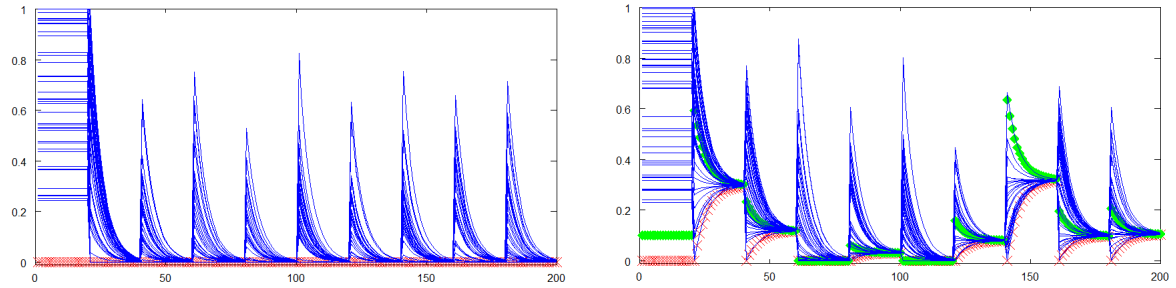


Figure 2.4: Description in text; also $\tau = \infty$, $\delta = 0.2$ and $T(m) = 1$ for $m < n$ and $T(n) = 0$. For each topic, we plot discussion rounds $t = 0, 1, 2, \ldots, 20$. Left: $N_\eta(\mathbf{b}^1(\infty), \mu_1) = \{1\} = \mathcal{N}_1$ (ϵ-intelligent agent in red) such that all agents are ϵ-wise for all topics X_k, with $k > 1$. Right: $N_\eta(\mathbf{b}^1(\infty), \mu_1) = \{1, 2\}$ contains also one biased agent (in green) such that limiting beliefs of agents are mixtures of truth and Z_k.

Remark 2.6.3. As an application of Proposition 2.6.5, consider the situation when the number n of agents goes to infinity. Then, if the fraction $\frac{n_2}{n}$ of agents in $\mathcal{N}_2$ converges to zero, agents become ϵ-wise, in the limit, as $n \to \infty$. Namely, first, the coefficient λ_Z converges to zero in this case since $\lambda_Z \leq \frac{n_2}{n_1} = \frac{n_2}{n-n_2}$, so that agents' expected consensus is indeed μ_k as $n \to \infty$. Moreover, not only do agents' limiting beliefs converge to μ_k *in expectation*, but agents become indeed ϵ-wise in the limit, as the support of the distribution of the ϵ-intelligent agents' initial beliefs is $B_{k,\epsilon}$.

As examples of $\frac{n_2}{n}$ converging to zero, of course, if the number n_2 remains constant as $n \to 0$, then $\frac{n_2}{n}$ goes to zero. But even if, for example, n_2 grows as in $\sqrt{n}$, all agents finally become ϵ-wise.

Proposition 2.6.5 may be restated in the following way; agents' limiting beliefs, in expectation, are given by $\lambda_Z \mathbb{E}[Z_k] + \lambda_\mu \mu_k$, where $\lambda_Z = 0$ if $N_\eta(\mathbf{b}^1(0), \mu_k) = \mathcal{N}_1$. We can then determine the probability that $\lambda_Z = 0$.

Corollary 2.6.2. Under the conditions of Proposition 2.6.5, with probability exactly $F_{Z_k}(B^c_{k,\eta})^{n_2} > 0$, we have $\lambda_Z = 0$.

Proof. The event that the n_2 biased agents' initial beliefs $b^1_i(0)$ are in $B^c_{k,\eta}$ is, by the iid property, $F_{Z_k}(B^c_{k,\eta})^{n_2}$. □

Remark 2.6.4. According to the corollary, the probability that $\lambda_Z = 0$ is strictly positive but decreasing as the number of biased agents increases. Hence, as n_2 becomes large, agents' limiting beliefs are very likely mixtures of truth μ_k and $\mathbb{E}[Z_k]$, a value that is different from truth.

Remark 2.6.5. We may consider the setup of Proposition 2.6.5 as a 'type inference' problem. What the proposition says and shows is that, since agents adjust their weights based on limiting beliefs, they cannot infer the intelligent agents once a biased non-intelligent agent has guessed truth because agents always reach a consensus in our situation (cf. also Proposition 2.6.1). Thus, the intelligent agents cannot properly signal their type in this case because all agents' limiting beliefs are indistinguishable.

Now, consider the exact same situation as in Proposition 2.6.5, except that agents adjust weights based on initial beliefs, i.e., $\tau = 0$. In this situation, a sufficient condition for wisdom is that agents find truth sufficiently valuable, i.e., δ is sufficiently large. In this case, wisdom, in the limit as $k \to \infty$, obtains almost surely, namely, all that is required is that only the ϵ-intelligent agents in $\mathcal{N}_1$ are initially true for some topic X_k.

Proposition 2.6.6. Let the weight adjustment time point be $\tau = 0$.

Under the conditions as in Proposition 2.6.5, if δ is sufficiently large, then, almost surely, there exists a (time point) $M \in \mathbb{N}$ such that all agents are ϵ-wise for all topics X_k, with $k > M$.

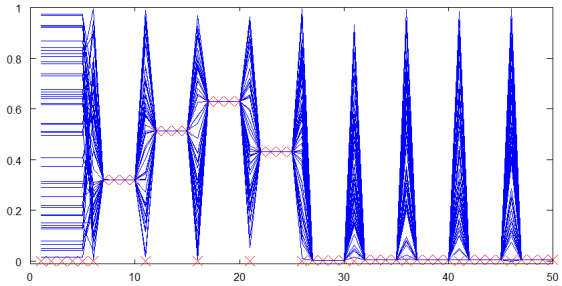


Figure 2.5: Description in text; also $\tau = 0$, $\delta = 100$ and $T(m) = 1$ for $m < n$ and $T(n) = 0$. For each topic, we plot discussion rounds $t = 0, 1, 2, \ldots, 5$. For topic $M = 5$, we have $N_\eta(\mathbf{b}^M(0), \mu_M) = \mathcal{N}_1$ such that all agents are ϵ-wise for all topics X_k, with $k > M$.

Proof. Let M be the first time point that (1) only ϵ-intelligent agents in $\mathcal{N}_1$ happen to know truth, initially, for topic X_M, that is, $b_i^M(0) \in B_{k,\eta}$ for all $i \in \mathcal{N}_1$ and no $i \in \mathcal{N}_2$; and, (2) not all agents are η-wise for X_M (such that $T(\cdot) > 0$). Then, weight adjustment at $M + 1$ will add $\tilde{\delta} > 0$ to the weights of the ϵ-intelligent agents in $\mathcal{N}_1$. If $\tilde{\delta}$ is sufficiently large, after normalization, weights for the non-intelligent agents become arbitrarily small and (arbitrarily close to) uniform for the ϵ-intelligent agents. In particular, δ may be so large that all agents' beliefs $b_i^{M+1}(1)$ lie in $B_{k,\epsilon}$. Since this is a convex set and weight matrices are row-stochastic, beliefs will remain in $B_{k,\epsilon}$ for all time periods t; hence, agents will be ϵ-wise in the limit for topic X_{M+1}, and, consequently, also η-wise. Since $T(n) = 0$, no more adjustments will occur after time point $M + 1$ and all agents become ϵ-wise for all topics X_k, with $k > M$, since their weights are now (sufficiently close to) uniform for the ϵ-intelligent agents in $\mathcal{N}_1$. □

Example 2.6.2. We illustrate Proposition 2.6.6 in Figure 2.5, where we let $S = [0, 1]$, $\mu_k = 0$ for all $k \geq 1$, $[n] = \{1, \ldots, 50\}$, $\mathcal{N}_1 = \{1\}$ and F_{Z_k} is the random uniform distribution on S, $\epsilon = 0$ and $\eta = 0.05$.

Remark 2.6.6. To summarize, the intelligent agents in $\mathcal{N}_1$ can now correctly signal their type. All that is required is that only ϵ-intelligent agents in $\mathcal{N}_1$ happen to know truth for some topic, in which case they will receive such a large weight increment that they lead society to ϵ-wisdom; then, no more weight adjustments occur because the 'right guys' have been identified.

Remark 2.6.7. In our current setup, the difference between weight adjustment at $\tau = 0$ vs. at $\tau = \infty$ is as follows. While adjusting at $\tau = 0$ leads agents to ϵ-wisdom almost surely provided that they find truth sufficiently valuable, that is, δ is large enough; updating at $\tau = \infty$ leads agents to ϵ-wisdom provided that biased agents do not know (or, perhaps, 'guess') truth for topic X_1. The latter condition is difficult to satisfy if we assume that the number of biased agents becomes large, while the condition of sufficiently large δ also depends on population size n and, in particular, on n_2, the population size of the biased agents. In other words, if $T(n) = 0$, we can specify sufficient conditions for wisdom even under the presence of biased agents, but these are rather challenging.

The case $T(\cdot) > 0$

Now, we consider the same setup as in the last subsection, except that we assume that $T(\cdot) > 0$ on its whole domain. In this case, agents continuously adjust their weights to other agents, which is also the rational behavior of an agent who assumes the conditions outlined in Section 2.4; recall our previous discussion.

We consider a slightly more general situation here than in the last subsection in that we allow each agent to have initial beliefs distributed according to individualized distribution functions, rather than to assume groups with identical distribution functions; the more restrictive setting is then a special case of our generalization. Accordingly, assume that agent i's initial belief for topic X_k is distributed according to random variable $Z_{i,k}$ with distribution function $F_{i,k}(A) = \Pr[Z_{i,k} \in A]$ for all $A \subseteq S$ and all topics

X_k, for $k \in \mathbb{N}$. We assume that $F_{i,k}(B_{k,\eta})$, which gives the probability that agent i is within an η-radius around truth μ_k, does not depend on topic X_k, that is, $F_{i,k}(B_{k,\eta}) = F_{i,k'}(B_{\eta,k'})$ for all k, k', which means that the probability that agent i is truthful is the same across topics. We then have the following proposition.

Proposition 2.6.7. Let tolerance $\eta \geq 0$ be fixed. Assume that agents adjust weights based on initial beliefs, i.e., $\tau = 0$, and assume that $T(\cdot) > 0$. Then, as $k \to \infty$, agents' limiting consensus beliefs on issue X_k are distributed according to

$$b_i^k(\infty) \sim \sum_{j=1}^{n} \lambda_j Z_{j,k}$$

where

$$\lambda_j \propto F_{j,k}(B_{k,\eta})$$

with $\sum_{j=1}^{n} \lambda_j = 1$ (note that λ_j does not depend upon k by assumption). In particular, we have

$$\mathbb{E}[b_i^k(\infty)] = \sum_{j=1}^{n} \lambda_j \, \mathbb{E}[Z_{j,k}].$$

Proof. Our proof is not rigorous.

Since agents are homogenous with respect to tolerance η, they will all *jointly* increase their weight to a particular agent j (or they will *jointly* not do so). Therefore, as k increases, rows of $\mathbf{W}^{(k)}$ become more and more similar, independent of the initial conditions $\mathbf{W}^{(1)}$ (if weight matrix $\mathbf{W}^{(1)}$ is identical in each row, this will propagate to any $\mathbf{W}^{(k)}$ with $k > 1$, but even if not, rows will become more and more similar by the homogeneity of agents). The weight mass that any particular agent i assigns to any particular agent j is clearly proportional to $F_{j,k}(B_{k,\eta})$ (cf. Figure 2.1) since this value indicates how frequently agent j is truthful. Hence, since rows of $\mathbf{W}^{(k)}$ are (approximately) identical, as k becomes large, with each entry $[\mathbf{W}^{(k)}]_{ij}$ being proportional to $F_{j,k}(B_{k,\eta})$, limiting beliefs of agents are given by,

$$b_i^k(\infty) \approxeq b_i^k(1) = \sum_{j=1}^{n} \lambda_j b_j^k(0),$$

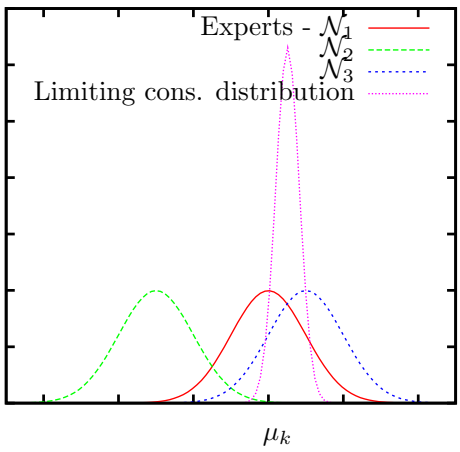
where $\lambda_j \propto F_{j,k}(B_{k,\eta})$. This completes the proof. □

Remark 2.6.8. The coefficients λ_j have a very intuitive interpretation. Since they indicate how limiting consensus beliefs are formed in terms of initial beliefs, their standard interpretation is that of *social influence weights* (cf., e.g., Golub and Jackson, 2010). Clearly, in our endogenous weight formation model, with weight sizes dependent upon 'past performance', an agent's social influence is intuitively given by his likelihood of being close to truth.

Example 2.6.3. Considering the distribution of limiting consensus beliefs, we note that if two $Z_{j,k}$, for $j = x, y$, are normally distributed with parameters $(\mu_x^k, \sigma_{x,k}^2)$ and $(\mu_y^k, \sigma_{y,k}^2)$, then both $\lambda_j Z_{j,k}$ as well as $\sum_j \lambda_j Z_{j,k}$ are normally distributed; the latter sum has normal distribution with parameters $(\lambda_x \mu_x^k + \lambda_y \mu_y^k, \sigma_{x,k}^2 + \sigma_{y,k}^2)$. Hence, if all agents' initial beliefs are normally distributed, their limiting beliefs are also normally distributed.

Moreover, if there are several 'types' or 'groups' of agents, $\mathcal{N}_1, \ldots, \mathcal{N}_m$, of sizes $n_1, \ldots, n_m$, where each group has identical and independent initial distribution (within groups), then agents in each group receive about the same weight mass, which is proportional to (see example below) $\lambda_{\mathcal{N}_l} \frac{1}{n_l}$, for $l \in \{1, \ldots, m\}$, so that if sizes $n_1, \ldots, n_m$ of groups become large, then, by the central limit theorem, $\sum_{j \in \mathcal{N}_l} \lambda_{\mathcal{N}_l} \frac{1}{n_l} Z_{j,k} =$ is approximately normally distributed. Thus, by our above remark, $\sum_{j \in [n]} \lambda_j Z_{j,k}$ is also approximately normally distributed. In other words, we would generally expect agents' limiting beliefs to be normally distributed, in this setup.

Example 2.6.4. Consider three groups of agents, $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3 \subseteq [n]$ with $\mathcal{N}_1 \cup \mathcal{N}_2 \cup \mathcal{N}_3 = [n]$ and where the $\mathcal{N}_l$'s are pairwise mutually disjoint. The first group, which we call experts, has initial beliefs distributed according to $\mathtt{N}(\mu_k, \sigma_1^2)$, where $\sigma_1^2 > 0$ is fixed (that is, each member in $\mathcal{N}_1$ has the given distribution function, and we assume members' initial beliefs to be independent). The second and third groups are biased. Assume, for illustration, that group two has distribution $\mathtt{N}(\mu_k - a, \sigma_2^2)$ and group three has $\mathtt{N}(\mu_k + b, \sigma_3^2)$. Assume the groups have sizes $n_1 = \frac{1}{5}n$, and $n_2 = n_3 = \frac{2}{5}n$, that is, the group of experts is smallest in size (but still growing in n). Moreover, let, for instance, $a = 3$, $b = 1$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, and let $\eta = 0.25$. Then, each expert has λ_j of about $\lambda_j \propto 0.19741$, members of group two have $\lambda_j \propto 0.0024$ and members of group three have $\lambda_j \propto 0.0278$. Since the λ's must sum to one, we have about $\lambda_{\mathcal{N}_1} \approxeq \frac{0.19751 n_1}{C_0}$ for experts, and $\lambda_{\mathcal{N}_2} \approxeq \frac{0.0024 n_2}{C_0}$ and $\lambda_{\mathcal{N}_3} \approxeq \frac{0.0278 n_3}{C_0}$ for groups two and three, respectively, and where $C_0 = \lambda_{\mathcal{N}_1} + \lambda_{\mathcal{N}_2} + \lambda_{\mathcal{N}_3}$ and $\lambda_{\mathcal{N}_l} = \sum_{j \in \mathcal{N}_l} \lambda_j$ for $l = 1, 2, 3$. For $n = 100$, this is about $\lambda_{\mathcal{N}_1} \approxeq 0.44$, $\lambda_{\mathcal{N}_2} \approxeq 0.01$, and $\lambda_{\mathcal{N}_3} \approxeq 0.55$, which is also, approximately, the limiting structure of the distribution of λ as $n \to \infty$. Hence, in the limit, as $n \to \infty$, these agents beliefs' would converge to a consensus that is off by about $\lambda_{\mathcal{N}_1} \cdot 0 + \lambda_{\mathcal{N}_2} \cdot (-3) + \lambda_{\mathcal{N}_3} \cdot 1 = 0.51227$ from truths μ_k. More precisely, the agents' limiting consensus values are distributed according to a normal distribution with mean $\mu_k + 0.51227$ and variance that converges to zero in n; in particular, variance of limiting consensus values is given by $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + \frac{\sigma_3^2}{n_3}$, which is $\frac{1}{10}$ for our example. We plot the (predicted and theoretical) limiting distribution of $b_i^k(\infty)$ and a sample histogram from an actual simulation in Figure 2.6.



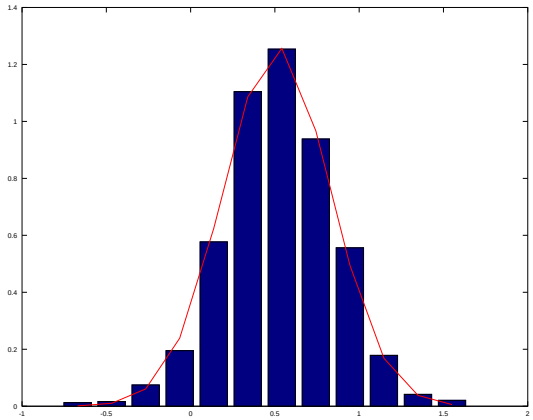


Figure 2.6: Left: The distribution functions of beliefs of three groups of agents as discussed in Example 2.6.4; experts' initial beliefs are always centered around truth, for all topics X_k, while there are two biased groups, one which underestimates truth and one which overestimates truth. If the relative sizes of groups are as described in the text, agents distribution of limiting beliefs, as k becomes large, is given by the high-peak normal distribution indicated, whose mean is off from truth by about 0.5. Right: Sample distribution from a simulation vs. predicted distribution according to Proposition 2.6.7.

In the next proposition, we discuss weight adjustment based on limiting beliefs. We assume that $F_{i,k}(A) > 0$ for all non-empty intervals $A \subseteq S$.

Proposition 2.6.8. Let tolerance $\eta \geq 0$ be fixed. Assume that agents adjust weights based on limiting beliefs, i.e., $\tau = \infty$, and assume that $T(\cdot) > 0$. Then, as $k \to \infty$, agents' limiting consensus beliefs on issue X_k are distributed according to

$$b_i^k(\infty) \sim \sum_{j=1}^{n} \frac{1}{n} Z_{j,k}$$

In particular, we have

$$\mathbb{E}[b_i^k(\infty)] = \sum_{j=1}^{n} \frac{1}{n} \mathbb{E}[Z_{j,k}].$$

Proof. Since agents reach a consensus for topics X_k, with $k > r$, and agents adjust weights based on limiting beliefs, weight matrix entries for all agents converge to $\frac{1}{n}$. Convergence to $\frac{1}{n}$ is assured since all agents have $F_{i,k}(B_{k,\eta}) > 0$ by assumption such that the probability that agents' limiting consensus is within an η-interval around truth is at least $F_{i,k}(B_{k,\eta})^n > 0$, from which it follows that agents adjust weights infinitely often (which each time entails an increment of $\tilde{\delta}$ and, thus, implies convergence of weight matrix entries to $\frac{1}{n}$) with probability 1. □

Remark 2.6.9. We see here, again, that adjusting based on limiting beliefs is 'worse' than adjusting based on initial beliefs, since limiting beliefs are formed through social interaction and may thus not indicate the inherent 'intelligence' of an agent.

To quantify the difference by way of illustration, in Example 2.6.4, agents' beliefs would now converge to a consensus, as $k \to \infty$, that is off from truths by about $\frac{n_1}{n} \cdot 0 + \frac{n_2}{n} \cdot (-3) + \frac{n_3}{n} \cdot 1 = -\frac{4}{5}$, which is further away than the value of about 0.51 given in the situation when agents adjust weights based on initial beliefs. In particular, agents in group $\mathcal{N}_2$, who are very poor at estimating truth, now receive much larger social influence than in the situation where agents adjust based on initial beliefs.

However, qualitatively, the results do not change (by much): in both circumstances, $\tau = 0$ and $\tau = \infty$, agents' limiting beliefs, under our endogenous weight adjustment process, are given by convex combinations of all agents' initial beliefs, whereby adjusting based on initial beliefs captures, in the social influence weights λ_j, the intelligence of agents while adjusting based on limiting beliefs leads agents to uniform social influence weights λ_j.

Now, consider, again, the setup where there are two groups of agents, which we denote by $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively; the first groups' initial beliefs are unbiased while the second groups' initial beliefs are biased, where we assume that agents within each group have independently and identically distributed initial beliefs. Assume, furthermore, that $F_{i,k}(B_{k,\eta}) > 0$ for all agents $i = 1, \ldots, n$.

Corollary 2.6.3. Let $\tau = 0$ or $\tau = \infty$ and let $\eta \geq 0$, the radius within which agents are considered to be truthful, be fixed. Then, if the group of biased agents $\mathcal{N}_2$ is 'large enough' (e.g., relative to $\mathcal{N}_1$), agents will not become ϵ-wise almost surely as $n, k \to \infty$, for any $\epsilon \in (0, \|\mu_k - \mathbb{E}[Z_{\mathcal{N}_2,k}]\|)$, whereby $Z_{\mathcal{N}_2,k}$ denotes a random variable that represents the distribution of initial beliefs of any agent from group $\mathcal{N}_2$.

Proof. By Proposition 2.6.7 and its proof, if $\tau = 0$, $\lambda_{\mathcal{N}_l} = \sum_{j \in \mathcal{N}_l} \lambda_j \approxeq \frac{F_{\mathcal{N}_l,k}(B_{k,\eta}) n_l}{C_0}$ as $k \to \infty$ (by $F_{\mathcal{N}_l,k}$, we denote the distribution function of an agent from group $\mathcal{N}_l$; also note that $F_{N_l,k}(B_{k,\eta})$ does not depend on k by assumption), where $l = 1, 2$ and $C_0 = \lambda_{\mathcal{N}_1} + \lambda_{\mathcal{N}_2}$. Thus, if n_2 is large enough (relative to n_1), $\lambda_{\mathcal{N}_2}$ be may arbitrarily close to 1 such that, in expectation, agents' limiting consensus belief will be arbitrarily close to $\mathbb{E}[Z_{\mathcal{N}_2,k}]$, whereby, by assumption, $Z_{\mathcal{N}_2,k}$ is a biased variable. As $n \to \infty$, limiting beliefs will converge to $\mathbb{E}[Z_{\mathcal{N}_2,k}]$ almost surely, in this case, by the law of large numbers.

If $\tau = \infty$, Proposition 2.6.8 leads to the same conclusion. □

Remark 2.6.10. What Corollary 2.6.3 shows is that agents may not become *infinitely* wise under our endogenous weight adjustment process if the group of agents with biased initial beliefs becomes large, as, in this case, this group's social influence will become arbitrarily large. But the corollary shows more: agents may not become ϵ-wise for any $\epsilon \in (0, \|\mu_k - \mathbb{E}[Z_{\mathcal{N}_2,k}]\|)$, which may be an arbitrarily large interval, depending on the bias of the agents in $\mathcal{N}_2$. In other words, if the number of biased agents is large (relative to the number of intelligent agents), the wisdom that society as a whole can attain is limited by the latter agents' bias.

2.6.3 Varying weights on own beliefs

DeMarzo, Vayanos, and Zwiebel (2003) consider a slight generalization of belief updating process (2.3.1) where agents may place varying weights on their own beliefs such that (2.3.1) reads as

$$\mathbf{b}^k(t+1) = \Big((1-\lambda_t)\mathbf{I}_n + \lambda_t \mathbf{W}^{(k)}\Big)\mathbf{b}^k(t) \tag{2.6.4}$$

whereby $0 < \lambda_t \leq 1$ (note that we treat λ_t as an exogenous variable). Such a weighting scheme may be empirically plausible, as it has been found (cf., e.g., Mannes, 2009) that people often tend to overweight their own beliefs relative to that of outsiders, probably because individuals have access to their own motivations for beliefs while they do not have such justification for others' beliefs. This reasoning would imply that λ_t is 'relatively small'. However, as long as weights on others' beliefs do not drop to zero too quickly, belief updating rule (2.6.4) leads to the same limiting beliefs as the original DeGroot updating rule (2.3.1) where $\lambda_t = 1$, for all t, provided that the latter converges; convergence may take sufficiently longer, however. Hence, under these circumstances, all our previous results remain valid. The following proposition is a straightforward generalization of the corresponding theorem, Theorem 1, in DeMarzo, Vayanos, and Zwiebel (2003), which restates the lessons we have just mentioned.

Proposition 2.6.9. Assume that $\mathbf{W}^{(k)}$ converges (for all initial belief vectors $\mathbf{b}(0)$), then if, $\sum_{t=1}^{\infty} \lambda_t = \infty$, updating process (2.6.4) also converges (for all initial belief vectors $\mathbf{b}(0)$) and leads to the same limiting beliefs as (2.3.1) where $\lambda_t = 1$ for all t.

We list the proof in the appendix.

In all subsequent sections, we only discuss the situations when $\tau = 0$ and $T(\cdot) > 0$, as the other cases may be derived in a manner similar to what we have sketched in this section.

2.7 Opposition

In this section, we consider the situation when two subsets of agents 'oppose' each other. Such opposition may derive, for example, from in-group vs. out-group antagonisms, as is an important concept in psychology and sociology (cf. Brewer, 1979; Castano et al., 2002; Kitts, 2006) and as has also more recently been taken into account in economics models (cf., in an experimental context, e.g., Charness, Rigotti, and Rustichini, 2007; Fehrler and Kosfeld, 2013) and in social network theory (cf. Beasley and Kleinberg, 2010). Prime exemplars of opposition forces can be found in politics (e.g., democrats vs. republicans; opposition parties vs. governing party in charge), for example, or also on a more global societal level (e.g., punks or hippies/counterculture vs. mainstream culture). In the context of (DeGroot-like) opinion dynamics models, opposition has, prominently, been discussed in Eger (2013) (but see also our discussion in Section 2.2), whose modeling we relate to.

In the model of Eger (2013), there are *two* types of links between agents. One link type refers to whether agents follow or oppose each other and the other link type denotes the intensity of relationship and is given by a non-negative real number $W_{ij} \in \mathbb{R}$. Belief updating is then performed via the operation

$$\mathbf{b}^k(t+1) = (\mathbf{W}^{(k)} \circ \mathbf{F}^{(k)})\mathbf{b}^k(t), \tag{2.7.1}$$

whereby the operator $\mathbf{W}^{(k)} \circ \mathbf{F}^{(k)}$ is defined via

$$\big((\mathbf{W}^{(k)} \circ \mathbf{F}^{(k)})(\mathbf{b})\big)_i = \sum_{j=1}^{n} W_{ij}^{(k)} F_{ij}^{(k)}(b_j),$$

whereby $F_{ij}^{(k)} \in \{F, D\}$, where $F : S \to S$ is the identity function ('agent i follows agent j') and $D : S \to S$ is an opposition function ('agent i opposes/deviates from agent j'). In other words, in this model, agents form their current beliefs by inverting (via D) or not ('via F') the past belief signals of others and then taking a weighted arithmetic average, as in standard DeGroot learning, of the so modified (or not) belief

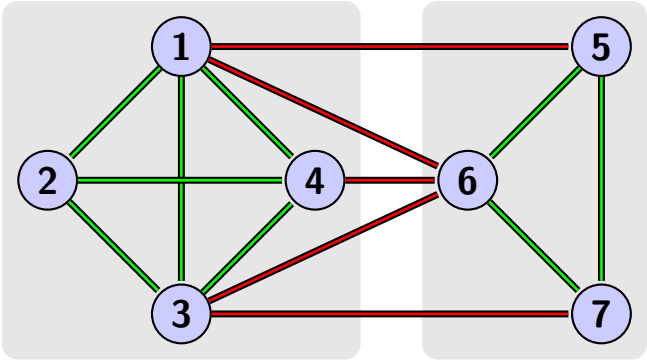


Figure 2.7: Illustration of an opposition bipartite operator $\mathbf{F}$ (agent nodes as blue circles). For convenience, D relationships are indicated in red, and F relationships in green. Here, we draw the network of agents as undirected, although we generally allow directed links between agents.

signals of their neighbors. As becomes evident, endogenizing this model would require endogenizing two variables, namely, $F_{ij}^{(k)}$ and $W_{ij}^{(k)}$, a task that is beyond the scope of this section. Therefore, we take $F_{ij}^{(k)}$ as exogenous and keep, as before, $W_{ij}^{(k)}$ as an endogenous variable, formed, in the case that $F_{ij} = F$, by reference to an agent's past performance.[14]

Hence, we consider the following situation. Denote by $\mathcal{A} \subseteq [n]$ and $\mathcal{B} \subseteq [n]$ the two groups of agents that oppose each other. We posit that $\mathbf{F}$ is *opposition bipartite* (cf. Figure 2.7): for all agents $i, i' \in \mathcal{A}$ it holds that $F_{ii'} = F_{i'i} = F$ (analogously for $\mathcal{B}$) and for all agents $i \in \mathcal{A}$ and $i' \in \mathcal{B}$ it holds that $F_{ii'} = F_{i'i} = D$, which simply means that agents within the two groups follow each other whereas agents across the two groups oppose each other.[15] Next, we assume that, regarding weight adjustments, members of both groups *ignore* members of the other group (a sin or bias of omission), taking into account only members of their own group, that is,

$$W_{ij}^{(k+1)} = \begin{cases} W_{ij}^{(k)} + \delta \cdot T(\left|N(\mathbf{b}^k(\tau), \mu_k)\right|) & \text{if } \left\|b_j^k(\tau) - \mu_k\right\| < \eta \text{ and } G(i) = G(j), \\ W_{ij}^{(k)} & \text{otherwise,} \end{cases} \tag{2.7.2}$$

where $G(i)$ denotes the group of agent i, which is either $\mathcal{A}$ or $\mathcal{B}$; for simplicity, assume $T(\cdot) = 1$, here and in the remainder of this section. Finally, we assume that agents i of group $\mathcal{A}$ initially assign uniform intensity of relationship $W_{ij}^{(1)} = b$ to each member j of group $\mathcal{B}$ and members of group $\mathcal{B}$ do analogously, assigning $W_{ij}^{(1)} = c$, where b and c are positive constants. We also assume that these levels stay fixed over topics, that is, $W_{ij}^{(k)} = W_{ij}^{(1)}$ whenever $G(i) \neq G(j)$. When $G(i) = G(j)$, as said, we let weights be formed according to (2.7.2). Finally, we always assume that weight matrices $\mathbf{W}^{(k)}$ are row-stochastic. We now discuss the so specified model, with endogenous weight (or intensity) formation for at least a subset of agents, in the following example. For opposition function D, we let D be soft opposition on $S = \mathbb{R}$ (see Eger, 2013, for details) with the functional form $D(x) = -x$. We first outline the following proposition from Eger (2013), which gives conditions for convergence of $\mathbf{W} \circ \mathbf{F}$, where we omit, here and in the following, reference to topics X_k for notational convenience.

Proposition 2.7.1. Let D be soft opposition on $S = \mathbb{R}$. Then, $\mathbf{W} \circ \mathbf{F}$ is affine-linear with representation $(\mathbf{A}, \mathbf{0})$. Then, if $\mathbf{F}$ is opposition bipartite, $\lambda = 1$ is an eigenvalue of $\mathbf{A}$. If $\lambda = 1$ is the only eigenvalue of $\mathbf{A}$ on the unit circle and if $\lambda = 1$ has algebraic multiplicity of 1, then $\lim_{t\to\infty}(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0) = \mathbf{p}$ for some polarization opinion vector $\mathbf{p}$ (that depends on $\mathbf{b}(0)$) and all initial opinion vectors $\mathbf{b}(0) \in S^n$.

In the proposition, a polarization opinion (or belief) vector is any vector $\mathbf{p}$ consisting of two beliefs $a, b \in S$ such that $D(a) = b$ and $D(b) = a$. For our specification of D, this would mean that $a = -b$.

[14] If $F_{ij} = D$, it would make no sense, or be at least problematic, to posit that an agent would increase his intensity of relationship, relating to opposition behavior, in proportion to another agent's accuracy of predicting truth.

[15] This specificiation is not self-evident; intra-group antagonisms, based, e.g., on personalized differences between members of the same group, might plausibly be allowable.

Hence, the proposition says that if D is soft opposition, then $\mathbf{W} \circ \mathbf{F}$ is representable by a matrix $\mathbf{A}$, and if in addition $\mathbf{F}$ is opposition bipartite (as we assume), then convergence of $\mathbf{W} \circ \mathbf{F}$ depends on the eigenvalues of $\mathbf{A}$. In the following example, we will make reference to the proposition.

Example 2.7.1. Let $n_1 = |\mathcal{A}|$ and $n_2 = |\mathcal{B}|$ with $n_1 + n_2 = n$. Before (partly) endogenizing $\mathbf{W}$, assume first that $\mathbf{W}$ and $\mathbf{F}$ have the following form,

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{\mathcal{A},\mathcal{A}} & \mathbf{W}_{\mathcal{A},\mathcal{B}} \\ \mathbf{W}_{\mathcal{B},\mathcal{A}} & \mathbf{W}_{\mathcal{B},\mathcal{B}} \end{pmatrix} \quad \mathbf{F} = \begin{pmatrix} \mathbf{F}_{\mathcal{A},\mathcal{A}} & \mathbf{F}_{\mathcal{A},\mathcal{B}} \\ \mathbf{F}_{\mathcal{B},\mathcal{A}} & \mathbf{F}_{\mathcal{B},\mathcal{B}} \end{pmatrix} \tag{2.7.3}$$

where $[\mathbf{W}_{\mathcal{A},\mathcal{A}}]_{ij} = a$, $[\mathbf{W}_{\mathcal{A},\mathcal{B}}]_{ij} = b$, $[\mathbf{W}_{\mathcal{B},\mathcal{A}}]_{ij} = c$, $[\mathbf{W}_{\mathcal{B},\mathcal{B}}]_{ij} = d$, and $[\mathbf{F}_{\mathcal{A},\mathcal{A}}]_{ij} = [\mathbf{F}_{\mathcal{B},\mathcal{B}}]_{ij} = F$, $[\mathbf{F}_{\mathcal{A},\mathcal{B}}]_{ij} = [\mathbf{F}_{\mathcal{B},\mathcal{A}}]_{ij} = D$; matrices $\mathbf{W}_{\mathcal{A},\mathcal{A}}$ and $\mathbf{F}_{\mathcal{A},\mathcal{A}}$ are of size $n_1 \times n_1$, $\mathbf{W}_{\mathcal{A},\mathcal{B}}$ and $\mathbf{F}_{\mathcal{A},\mathcal{B}}$ of size $n_1 \times n_2$, etc. Hence, agents in $\mathcal{A}$ follow each other, assigning weight a to each other, and agents in $\mathcal{B}$ assign weight d to each other; across the two sets, agents oppose each other, with weights b and c, respectively, as already indicated above. Moreover, for simplicity, as the given specification posits, we assume that weights are uniform within groups and opposition weights are also uniformly distributed. Since, as Proposition 2.7.1 tells, the so defined $\mathbf{W} \circ \mathbf{F}$ allows an (affine-)linear representation, this is given by, in this setup,

$$\mathbf{A} = \begin{pmatrix} \mathbf{W}_{\mathcal{A},\mathcal{A}} & -\mathbf{W}_{\mathcal{A},\mathcal{B}} \\ -\mathbf{W}_{\mathcal{B},\mathcal{A}} & \mathbf{W}_{\mathcal{B},\mathcal{B}} \end{pmatrix}, \tag{2.7.4}$$

as one can verify (cf. Eger, 2013). Furthermore, if $(\mathbf{W} \circ \mathbf{F})^t \mathbf{b}(0) = \mathbf{A}^t \mathbf{b}(0)$ converges to a polarization vector (e.g., under the conditions of Proposition 2.7.1), then the one limiting belief is given by $\sum_{j=1}^n s_j b_j(0)$ and the other is given by $-\sum_{j=1}^n s_j b_j(0)$, where $\mathbf{s} = (s_1, \dots, s_n)^\intercal$ is the unique eigenvector of $\mathbf{A}^\intercal$ satisfying $\sum_{j=1}^n |s_j| = 1$ and corresponding to eigenvalue $\lambda = 1$ of $\mathbf{A}^\intercal$ (cf. Eger, 2013, Remark 6.4). The vector $\mathbf{s}$ is then a (generalized) social influence vector (cf. the concept of eigenvector centrality, e.g., Bonacich, 1972), with $|s_i|$ denoting the social influence (proper) of agent i and $\operatorname{sgn}(s_i)$ his group membership. Since, by our specification of $\mathbf{W} \circ \mathbf{F}$, agents in group $\mathcal{A}$ must have the same social influence (by homogeneity of these agents due to the uniform weight structure) as well as agents in group $\mathcal{B}$, we must accordingly have that $\mathbf{s} = (\underbrace{x, \dots, x}_{n_1}, \underbrace{y, \dots, y}_{n_2})^\intercal$ for some $x, y \in \mathbb{R}$. Then, y (or $|y|$) measures social influence of members of group $\mathcal{B}$ and accordingly for $\mathcal{A}$. Hence, from $\mathbf{A}^\intercal \mathbf{s} = \mathbf{s}$, we find (1) $n_1 a x - n_2 c y = x$, (2) $-n_1 b x + n_2 d y = y$, and (3) $n_1 x - n_2 y = 1$ (from the unit condition on $\mathbf{s}$). From this, it follows that

$$y = \frac{b}{n_2(d-b) - 1}, \quad \text{and} \quad x = (1 + n_2 y)a - n_2 c y. \tag{2.7.5}$$

The case of y may serve as an illustration. Computing the comparative statics of $|y|$ with respect to b and d, we first find that since $n_2(d-b) \leq n_2 d \leq 1$, it holds that $y \leq 0$. Hence, $|y| = \frac{b}{1 - n_2(d-b)}$ and then,

$$\frac{\partial |y|}{\partial b} = \frac{1 - n_2 d}{\big(1 - n_2(d-b)\big)^2} \geq 0, \quad \frac{\partial |y|}{\partial d} = \frac{n_2 b}{\big(1 - n_2(d-b)\big)^2} > 0$$

such that an increase in d leads to an increase in the absolute value of y, as we would expect — if the weight that members of group $\mathcal{B}$ place upon each other increases, their social influence, measured in absolute value, increases. Moreover, $|y|$ also increases in b — the more members of group $\mathcal{A}$ want to disassociate from members of group $\mathcal{B}$, the more does group $\mathcal{B}$'s social influence increase, in absolute value. We exemplify in Figure 2.8 (left).

Hence, under our current assumptions, limiting polarization beliefs of agents are given by $\sum_{j\in[n]} s_j b_j(0) = \sum_{j\in\mathcal{A}} x b_j(0) + \sum_{j\in\mathcal{B}} y b_j(0)$ and $-\sum_{j\in[n]} s_j b_j(0) = -\sum_{j\in\mathcal{A}} x b_j(0) - \sum_{j\in\mathcal{B}} y b_j(0)$, respectively. Let us, for the moment, assume that all agents are ϵ-intelligent with $\epsilon = 0$, that is, all agents precisely receive truths for topics, as initial beliefs. Then, limiting beliefs are, thus,

$$b^k_{\mathcal{A}}(\infty) = \sum_{j\in[n]} s_j b_j(0) = \mu_k \underbrace{\big(n_1 x + n_2 y\big)}_{=\mathsf{c}}, \quad \text{and} \quad b^k_{\mathcal{B}}(\infty) = -\sum_{j\in[n]} s_j b_j(0) = \mu_k \big(\underbrace{-\big(n_1 x + n_2 y\big)}_{=-\mathsf{c}}\big),$$

respectively, where closed-form solutions of x and y are given in Equation (2.7.5). In Figure 2.8 (right), we plot, for $c = \frac{1}{2n}$ and $d = \frac{1-n_1 c}{n_2}$ fixed, the coefficient $\mathsf{c} = n_1 x + n_2 y$ of limiting beliefs (and its negative, as coefficient of the other limiting belief), as a function of b (and, hence, also of a since $a = \frac{1-n_2 b}{n_1}$); note that this coefficient denotes the 'scaling' of truth in the limiting beliefs, whence, if it is equal to 1, (some) agents exactly reach truth. We observe the following: if b is very low (compared with c), i.e., agents in group $\mathcal{A}$ care little about agents in group $\mathcal{B}$ (at least relatively) — that is, opposition from $\mathcal{A}$ to $\mathcal{B}$ is (relatively) low — then c is very close to 1, which means that agents in group $\mathcal{A}$ have limiting beliefs very close to truth, while agents in group $\mathcal{B}$ hold limiting beliefs that are very close to $-\mu_k$, the 'opposite' of truth. As b increases, c becomes smaller, approaching zero as $b = c$. In other words, if opposition 'force' is equal between groups $\mathcal{A}$ and $\mathcal{B}$ — in the sense that $b = c$ — then both groups reach limiting beliefs of zero, no matter what truth is. As group $\mathcal{A}$ begins to oppose group $\mathcal{B}$ more heavily than $\mathcal{B}$ opposes $\mathcal{A}$, that is, $b > c$, group $\mathcal{A}$ goes further away from truth, toward opposite levels of truth in that c becomes negative, while group $\mathcal{B}$ begins to approach truth.

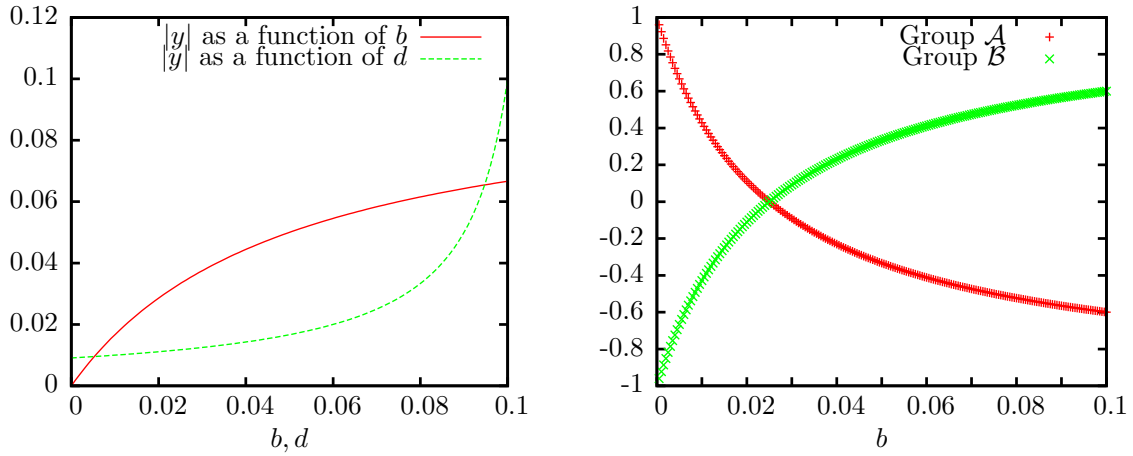


Figure 2.8: Both graphs: $n = n_1 + n_2 = 10 + 10 = 20$ agents. Left: Social influence, $|y|$, of agent i of group $\mathcal{B}$ increases both as a function of d (b fixed), 'within-group trust' of members of group $\mathcal{B}$, and b (d fixed), the importance assigned to members of group $\mathcal{B}$ via agents of group $\mathcal{A}$. Note that $n_2 b \leq 1$ and $n_2 d \leq 1$ (by row-stochasticity of weight matrices $\mathbf{W}$), which implies, in our case, $b, d \leq \frac{1}{10}$. Right: $c = \frac{1}{2n} = 0.025$, $d = \frac{3}{2n}$ fixed. Coefficient $\mathsf{c} = (n_1 x + n_2 y)$ of μ_k (red) and $-\mathsf{c}$ (green) as a function of b. Description in text.

Now, concerning the question whether the conditions on the eigenvalues of matrix $\mathbf{A}$, stated in Proposition 2.7.1, are satisfied — that is, do agents in fact converge to a polarization? — we mention that exactly determining the spectrum of $\mathbf{A}$ is difficult in the current situation, for general n_1 and n_2, and a, b, c, d. For $n_1 = 1$ and n_2 arbitrary (and, by symmetry hence also for $n_2 = 1$ and n_1 arbitrary), we find, in the appendix, that $\mathbf{A}$ has exactly one eigenvalue, namely $\lambda = 1$, on the unit circle and whose algebraic multiplicity is 1. Thus, in this case, by Proposition 2.7.1, beliefs under $\mathbf{W} \circ \mathbf{F}$ indeed converge to a polarization, as we have sketched it, and limiting beliefs have the indicated form. We strongly suspect that this is true for arbitrary n_1 and n_2, but leave the derivation open.

Finally, when would we expect $\mathbf{W} \circ \mathbf{F}$ to have the form (2.7.3), taking the form of $\mathbf{F}$ as exogenous? The structure of $\mathbf{W}$ holds, for instance, when $\mathbf{W}^{(1)}$ has the form indicated in (2.7.3), agents adjust weights (to members of their own group) based on initial beliefs, $\tau = 0$, and, e.g., $\epsilon = 0$ (agents' initial beliefs are exactly μ_k); then all $\mathbf{W}^{(k)}$ have the form as given in (2.7.3). Form (2.7.3) also arises, in the limit, as k becomes large, when $\tau = 0$ and initial beliefs are stochastically centered around truth and each agent has the same variance (namely, agents then tend to assign uniform weights to those they take into consideration in adjusting weights; uniform weights for outgroup members have been assumed exogenous by us, anyways). In fact, the simulations shown in Figure 2.9, for the latter case, show good agreement with the analytical predictions for the situation when all agents are ϵ-intelligent, for $\epsilon = 0$,

even for small k, indicating that $\mathbf{W}^{(k)}$ has a form close to (2.7.3) quickly, when agents are stochastically intelligent with identical variances.

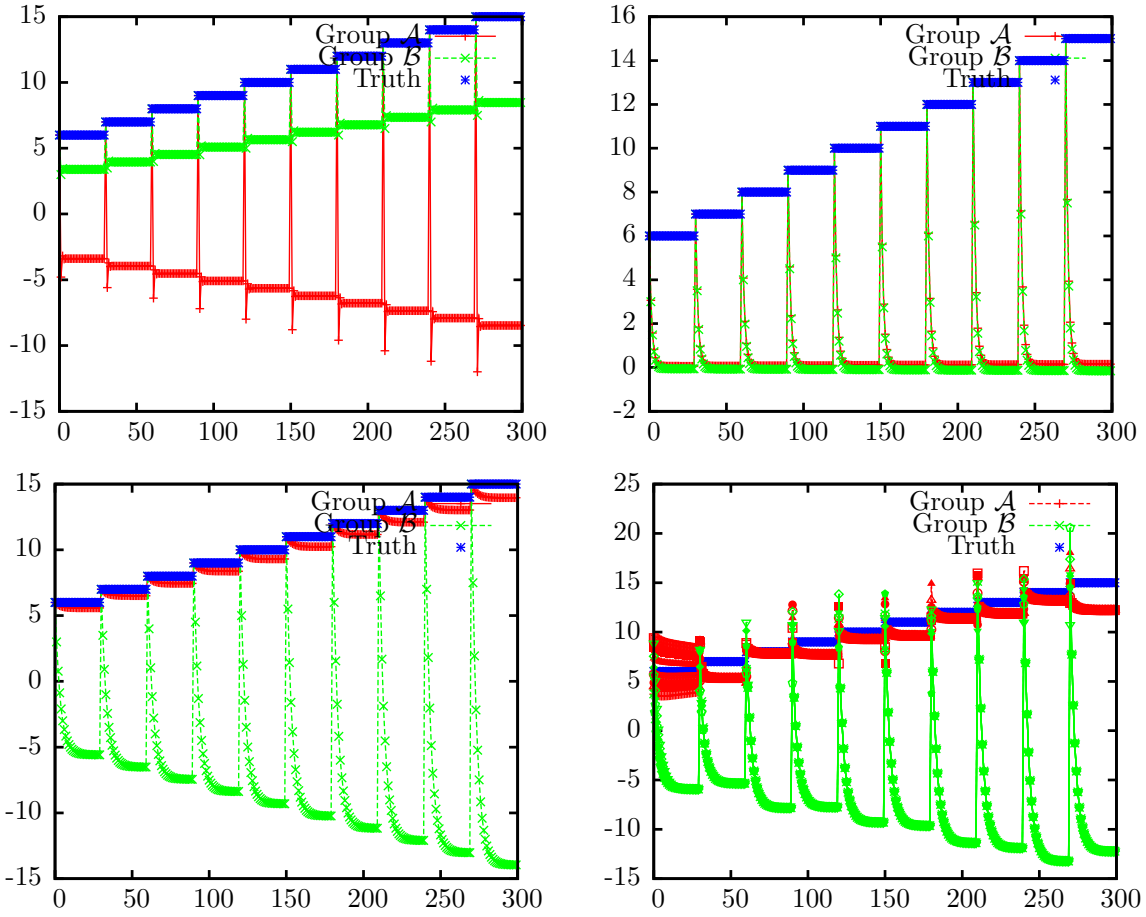


Figure 2.9: Throughout: $n = n_1 + n_2 = 10 + 10 = 20$ agents, $c = \frac{1}{2n} = 0.025$, $d = \frac{3}{2n}$ fixed. Discussion of 10 topics; $t = 0, 1, 2, \ldots, 20$ discussion periods shown, for each topic. Truth $\mu_k = k + 5$, for $k = 1, \ldots, 10$ (blue lines). In red and green: members of groups $\mathcal{A}$ and $\mathcal{B}$, respectively. Top left: All agents receive initial beliefs $b_i^k(0) = \mu_k$, $b = 0.09$ (cf. Figure 2.8 (right)). Top right: $b_i^k(0) = \mu_k$, $b = 0.0245$. Bottom left: $b_i^k(0) = \mu_k$, $b = 0.0009$. Bottom right: $b_i^k(0) = \mathtt{N}(\mu_k, 4)$, $b = 0.0009$.

Remark 2.7.1. In Example 2.7.1, we have outlined conditions under which we expect, due to polarization, at most one group of agents to become wise for topics. The conditions that we have highlighted — e.g., ϵ-intelligence, for $\epsilon = 0$, or initial beliefs stochastically centered around truth whereby all agents have identical variances — might appear quite special. We believe that similar polarization results hold for much more general conditions, but those outlined have the benefit of being analytically tractable more easily while still indicating results, as we think, of a general nature.

Remark 2.7.2. To summarize the importance of results indicated in Example 2.7.1, note that in this section, agents have been influenced by two 'polar' forces. On the one hand, they were attracted by truth by their adherence to principles that potentially lead them closer to truth — e.g., weight adjustment to those agents in their in-group that have been truthful in the past. On the other hand, agents had — exogenously — specified antagonisms to members of another group (a sin or bias of commission), their outgroup, which drew them toward beliefs that are different from those held by their adversaries. The message from Example 2.7.1 is clear in this context: the group that has (relatively) stronger incentives to

disassociate from negative referents tendentially will drift away from truth considerably, while the group with (relatively) weaker such incentives may still become wise (under appropriate initial conditions on beliefs), which is an intuitive result since, for the former group, disassociation seems to be (relatively) more critical than truth.

2.8 Conformity

Buechel, Hellmann, and Klößner (2013) and Buechel, Hellmann, and Klößner (2012)[16] study a DeGroot-like opinion dynamics model under *conformity*, that is, where individuals are not only *informationally* socially influenced by others but also *normatively* in that they are motivated to state opinions that tend to fit the group norm, possibly, in order to get "utility gain[s] by simply making the same choice as one's reference group" (cf. Zafar, 2011, p. 774). A classical example of such conforming behavior is documented in the famous study of Asch (1955) where subjects wrongly judged the length of a stick after some other, supposedly neutral, participants had given the same wrong judgment. More examples and relevant theoretical and empirical literature, e.g., Deutsch and Gerard (1955), Jones (1984), etc., on conforming behavior among human subjects are directly provided in Buechel, Hellmann, and Klößner (2013). As we have indicated in the introduction, we may perceive of conformity to a reference opinion, in our context, as a bias toward the beliefs of one's reference group.

Mathematically, agents in the named model update their beliefs *informationally* according to the following rule,

$$\mathbf{b}(t+1) = \mathbf{D}\mathbf{b}(t) + (\mathbf{W} - \mathbf{D})\mathbf{s}(t), \tag{2.8.1}$$

where $\mathbf{s}(t) \in S^n$ denotes the vector of *stated opinions* or beliefs (whose formation, as assumed, underlies normative social pressure, as we indicate below), $\mathbf{b}(t) \in S^n$ denotes the vector of *true beliefs*, $\mathbf{W}$ is the social network (or, 'learning matrix') as in the standard DeGroot model, and $\mathbf{D}$ denotes its diagonal. Updating rule (2.8.1) says that agents form their current beliefs by taking a weighted arithmetic average of their past true beliefs and others' stated beliefs. Then, as concerns *normative social influence*, agents are assumed to choose stated beliefs $s_i(t)$ by reference to the utility maximization problem

$$u_i(\mathbf{s}; \mathbf{b}) = -(1-\delta_i)(s_i - b_i)^2 - \delta_i(s_i - q_i)^2, \tag{2.8.2}$$

whereby the term $(s_i - b_i)^2$ represents an agent's preference for honesty (misrepresenting true opinions may cause cognitive discomfort, cf. Festinger, 1957) and the term $(s_i - q_i)^2$ represents preference for conforming to a reference opinion q_i. The parameter $\delta_i \in (-1, +1)$ displays the relative importance of the preference for conformity in relation to the preference for honesty. If $\delta_i < 0$, then agents have preference for *counter-conformity* in that their reference group serves as a negative referent. Now, consider that at the end of each (opinion updating) round $t = 0, 1, 2, \ldots$, agents play a normal form game $([n], S^n, u_i(\cdot; b_i(t)))$. Let $\mathbf{q} = (q_1, \ldots, q_n)^\intercal$ and let $\mathbf{q}(t) = \mathbf{Q}\mathbf{s}(t)$ where $\mathbf{Q}$ is an $n \times n$ matrix that indicates how reference opinions are formed from stated opinions; we assume that $Q_{ii} = 0$ for all $i \in [n]$ such that agents do not take into account their own stated opinion in reference opinion formation[17] and we also assume that $\mathbf{Q}$ is row-stochastic. The next proposition says that the normal form game has a unique Nash equilibrium.

Proposition 2.8.1. Denote by $\mathbf{\Delta}$ the diagonal matrix with $\Delta_{ii} = \delta_i$. Then the normal form game $([n], S^n, u(\cdot; \mathbf{b}(t)))$, for $u(\cdot; \mathbf{b}(t)) = \big(u_1(\cdot; \mathbf{b}(t)), \ldots, u_n(\cdot; \mathbf{b}(t))\big)$, has a unique Nash equilibrium, which is given by

$$\mathbf{s}^*(t) = (\mathbf{I}_n - \mathbf{\Delta}\mathbf{Q})^{-1}(\mathbf{I}_n - \mathbf{\Delta})\mathbf{b}(t) = \tilde{\mathbf{Q}}\mathbf{b}(t).$$

We prove Proposition 2.8.1 in the appendix. The proposition is a (straightforward) extension of the corresponding proposition, Proposition 1, in Buechel, Hellmann, and Klößner (2012) in that they choose the particular $\mathbf{Q}$ with $Q_{ij} = \frac{W_{ij}}{1-W_{ii}}$ $(Q_{ii} = 0)$. In the revised version of their paper, the named authors

[16] Henceforth, we only relate to the more recent version of their paper, unless the difference becomes important.

[17] They know better anyway, by knowledge of their true opinions.

also specify an iterative process that explains how agents reach the Nash equilibrium $\mathbf{s}^*(t)$ but we omit the recapitulation of this idea, because it is rather technical and does not provide further insight at this point.

Hence, simply assuming that agents play the Nash equilibrium $\mathbf{s}^*(t)$ at the end of each period t (such that, for $t+1$, $\mathbf{b}(t)$ *and* $\mathbf{s}(t)$ are available), beliefs evolve according to, combining (2.8.1) with $\mathbf{s}^*(t)$,

$$\mathbf{b}^k(t+1) = \mathbf{M}^{(k)}\mathbf{b}^k(t),$$

where

$$\mathbf{M}^{(k)} = \mathbf{D}^{(k)} + (\mathbf{W}^{(k)} - \mathbf{D}^{(k)})\tilde{\mathbf{Q}}^{(k)} = \mathbf{D}^{(k)} + (\mathbf{W}^{(k)} - \mathbf{D}^{(k)})(\mathbf{I} - \mathbf{\Delta}^{(k)}\mathbf{Q}^{(k)})^{-1}(\mathbf{I} - \mathbf{\Delta}^{(k)}), \tag{2.8.3}$$

and where we also index matrices by topic indices. As becomes obvious, this model has now many variables that can potentially be endogenized, namely, $\mathbf{W}^{(k)}$, $\mathbf{\Delta}^{(k)}$, which summarizes the conformity parameters, and $\mathbf{Q}^{(k)}$, which summarizes how agents form reference opinions. In the following, we take $\mathbf{\Delta}^{(k)}$ as exogenously given and constant across topics; the elements $[\mathbf{\Delta}^{(k)}]_{ii} = \delta_i$ may then be perceived as 'personality traits' of individuals. For $\mathbf{W}^{(k)}$, we assume the same endogenous weight formation as before, where weight increments depend upon past performance. The matrix $\mathbf{Q}^{(k)}$, we take as arbitrary exogenous variable first, satisfying row-stochasticity and $Q_{ii} = 0$ as above, and specialize then in the examples.

Our first proposition paths the way for a convergence result in our situation. It says that the property of having a positive column propagates from $\mathbf{W}^{(k)}$ to $\mathbf{M}^{(k)}$ if no agent is counter-conforming.

Proposition 2.8.2. Let $k \geq 1$ be arbitrary. Let $\delta_i \geq 0$ for all $i \in [n]$ such that agents never counter-conform. Then, if $\mathbf{W}^{(k)}$ has a positive column, then so does $\mathbf{M}^{(k)}$.

Proof. By the proof of Proposition 2.8.1, given in the appendix, the inverse of $\mathbf{I}_n - \mathbf{\Delta}^{(k)}\mathbf{Q}^{(k)}$ always exists (as long as $|\delta_i| < 1$, which we assume throughout) and is given by $\sum_{r=0}^{\infty}\left(\mathbf{\Delta}^{(k)}\mathbf{Q}^{(k)}\right)^r$. Since $\delta_i \geq 0$ and since $\mathbf{Q}^{(k)}$ is assumed to be row-stochastic, the latter sum is a sum of non-negative matrices and therefore, the infinite sum yields a matrix with non-negative entries. Moreover, since $\mathbf{A}^0 = \mathbf{I}_n$ for any arbitrary matrix $\mathbf{A}$, all diagonals of $\sum_{r=0}^{\infty}\left(\mathbf{\Delta}^{(k)}\mathbf{Q}^{(k)}\right)^r$ are hence strictly positive ($\mathbf{I}_n$ has strictly positive diagonals and the remaining summands are all non-negative). Moreover, since $\mathbf{P} := \mathbf{I}_n - \mathbf{\Delta}^{(k)}$ is a diagonal matrix with each entry $P_{ii} \in (0, 1]$,

$$\tilde{\mathbf{P}} = \left(\mathbf{I}_n - \mathbf{\Delta}^{(k)}\mathbf{Q}^{(k)}\right)^{-1}\left(\mathbf{I}_n - \mathbf{\Delta}^{(k)}\right)$$

accordingly also has diagonal entries that are all strictly positive. Next, since $\mathbf{W}^{(k)}$ has a strictly positive column j by assumption, $\mathbf{W}^{(k)} - \mathbf{D}^{(k)}$ has a strictly positive column j, except for element j of that column, which is zero. Hence, multiplying, $\mathbf{W}^{(k)} - \mathbf{D}^{(k)}$, a non-negative matrix by assumption, with $\tilde{\mathbf{P}}$ results in a matrix that also has a strictly positive column j, except possibly for its diagonal. But since $\mathbf{D}^{(k)}$ has a positive entry $[\mathbf{D}^{(k)}]_{jj}$ (since column j of $\mathbf{W}^{(k)}$ is positive by assumption),

$$\mathbf{D}^{(k)} + (\mathbf{W}^{(k)} - \mathbf{D}^{(k)})(\mathbf{I} - \mathbf{\Delta}^{(k)}\mathbf{Q}^{(k)})^{-1}(\mathbf{I} - \mathbf{\Delta}^{(k)})$$

has a positive column j. The latter matrix is, by definition, Eq. (2.8.3), precisely the matrix $\mathbf{M}^{(k)}$. □

As a corollary, we have our first convergence (to consensus) result, which provides both an alternative to the convergence result provided in Buechel, Hellmann, and Klößner (2013), and a generalization as well as a strengthening. It is more general since it considers arbitrary $\mathbf{Q}$ rather than the peculiar choice that the named authors consider. It provides an alternative since it says that conformity and a positive column of $\mathbf{W}^{(k)}$ are sufficient conditions for convergence, while the proposition in Buechel, Hellmann, and Klößner (2013) states that conformity and a positive *diagonal* of $\mathbf{W}^{(k)}$ are sufficient conditions for convergence. Finally, it is a strengthening because it states convergence to *consensus* rather than merely convergence. Before proving the theorem, we need the following lemma which says that the rows of $\mathbf{M}^{(k)}$ sum to 1 and which we prove in the appendix.

Lemma 2.8.1. The matrix $\mathbf{M}^{(k)}$ defined in (2.8.3) satisfies

$$\mathbf{M}^{(k)}\mathbb{1} = \mathbb{1}$$

for any row-stochastic $\mathbf{Q}$.

Corollary 2.8.1. Let $k \geq 1$ be arbitrary. Assume that $\delta_i \geq 0$ for all $i \in [n]$. Then, if $\mathbf{W}^{(k)}$ has a positive column, then $\mathbf{M}^{(k)}$ induces a consensus.

Proof. First, if $\delta_i \geq 0$, then $\mathbf{M}^{(k)}$ is a non-negative matrix by the proof of Proposition 2.8.2. Moreover, by Lemma 2.8.1, $\mathbf{M}^{(k)}$ is then also row-stochastic. Finally, by Proposition 2.8.2, if $\mathbf{W}^{(k)}$ has a positive column, then so does $\mathbf{M}^{(k)}$. A row-stochastic matrix with positive column induces a consensus by Theorem 2.6.1. □

It might be worthwhile, in future considerations, to study in more depth which properties $\mathbf{M}^{(k)}$ inherits from $\mathbf{W}^{(k)}$, and under which conditions. As mentioned, Buechel, Hellmann, and Klößner (2013) demonstrate that $\mathbf{M}^{(k)}$ inherits a positive diagonal from $\mathbf{W}^{(k)}$ (under their particular choice of $\mathbf{Q}$ and under conformity) as well as the general social network structure (see their discussion in their Section 4), while we show that the property of positive columns also propagates, for arbitrary $\mathbf{Q}$.

For now, we contend ourselves with the fact that Corollary 2.8.1 implies that, as in the standard DeGroot model, agents almost always reach a consensus — that is, for almost all topics — even under conformity ($\delta_i \geq 0$), under very mild conditions.

Proposition 2.8.3. Let $\eta \geq 0$, agents' tolerance, be fixed. Let $\tau = 0$ (resp. $\tau = \infty$). Let $T(\cdot) > 0$. As in Proposition 2.6.1, let r be the earliest time point that some agent is η-intelligent (resp. η-wise) for topic X_r. Then, under the conformity model presented above, with $\delta_i \geq 0$ for all $i \in [n]$, (a) agents reach a consensus for all topics X_k with $k > r$, independent of their initial beliefs. (b) For topics $1, \ldots, r$, agents' reaching a consensus depends on $\mathbf{W}^{(1)}$ and on $\mathbf{Q}^{(1)}$ to $\mathbf{Q}^{(r)}$ as well as on $\mathbf{\Delta}^{(1)}$ to $\mathbf{\Delta}^{(r)}$.

Proof. (a) As in the corresponding proof of Proposition 2.6.1, $\mathbf{W}^{(r+1)}$ has a positive column and so do, in general, have all matrices $\mathbf{W}^{(k)}$, for $k > r$. Hence, by Corollary 2.8.1, agents reach a consensus for topics X_k, for $k > r$, under the conformity model, as long as $\delta_i \geq 0$ for all $i \in [n]$. (b) Of course, whether or not agents reach a consensus for X_1 to X_r depends on the parameters of the model. □

As mentioned before, if there exist agents whose initial beliefs have positive probability of lying within an η-interval around truths, then r, as defined in Proposition 2.8.3, is a finite number almost surely. For standard parametrizations (e.g., all agents have positive probability of being truthful, for all topics), r is very low — typically $r = 1$ — with probability that goes to 1 in n, population size (cf. Remark 2.6.2).

Example 2.8.1. If agents are counter-conforming, Proposition 2.8.3 may be false in that beliefs may even diverge, rather than lead to a consensus. Consider, for instance,

$$\mathbf{W}^{(r+1)} = \frac{1}{1+2\delta}\begin{pmatrix} 1+\delta & \delta & 0 \\ \delta & 1+\delta & 0 \\ \delta & \delta & 1 \end{pmatrix},$$

which would be the resulting weight matrix if $\tau = 0$ and agent 1's and 2's initial beliefs were in an η-radius around truth, for the first time, for topic X_r. For convenience, assume that

$$\mathbf{Q}^{(r+1)} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \quad \mathbf{\Delta} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix},$$

where $-1 < a, b, c < 1$. Then, sample belief dynamics for this setting are sketched in Figure 2.10. As the graphs illustrate, under counter-conformity, agents may want to diassociate from others in a manner strong enough to induce divergence.

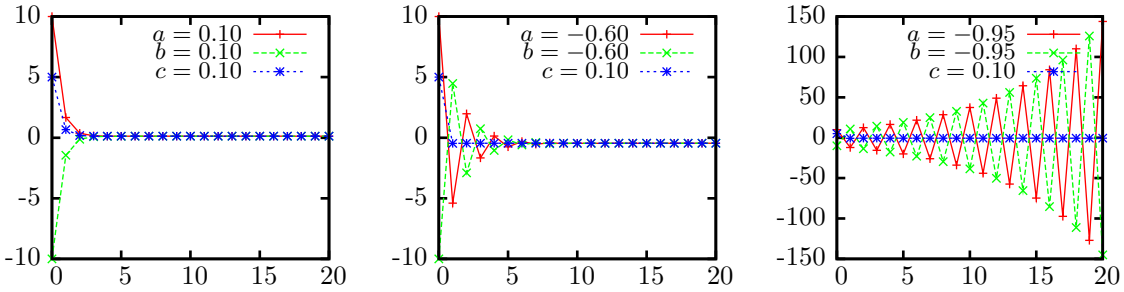


Figure 2.10: Belief dynamics for topic X_{r+1} as sketched in Example 2.8.1. We use $\delta = 5$ for the weight adjustment increment.

Next, we discuss social influence of agents as a function of conformity parameters and the structure of $\mathbf{W}^{(k)}$. In particular, we show that even agents with an 'empty record of past successes' can be influential in the endogenous conformity model if conformity parameters δ_i and matrix $\mathbf{Q}$ are appropriately specified.

Proposition 2.8.4. Let $\delta_i > 0$ for all $i \in [n]$. Assume that $\mathbf{Q}$ is strictly positive on all off-diagonal entries. Then, if $\mathbf{W}$ has at least two positive columns, $\mathbf{M}$ is strictly positive everywhere.

Proof. Again, $\mathbf{M}$ is

$$\mathbf{M} = \mathbf{D} + (\mathbf{W} - \mathbf{D})(\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q})^{-1}(\mathbf{I}_n - \boldsymbol{\Delta}).$$

We have $\mathbf{R} := (\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q})^{-1} = \sum_{r=0}^{\infty}(\boldsymbol{\Delta}\mathbf{Q})^r$ is strictly positive in each entry since $\mathbf{Q}$ is positive on all off-diagonals and since $\delta_i > 0$, whence $(\boldsymbol{\Delta}\mathbf{Q})^1$ is positive on all off-diagonals and note that $(\boldsymbol{\Delta}\mathbf{Q})^0 = \mathbf{I}_n$ has positive diagonals; the remaining terms $(\boldsymbol{\Delta}\mathbf{Q})^r$, for $r \geq 2$, are non-negative. Hence, also $\tilde{\mathbf{R}} := \mathbf{R}(\mathbf{I}_n - \boldsymbol{\Delta}) > 0$ entrywise, since the diagonals of $(\mathbf{I}_n - \boldsymbol{\Delta})$ are positive. Hence, since $\mathbf{W}$ has at least two positive columns, $(\mathbf{W} - \mathbf{D})\tilde{\mathbf{R}}$ is also positive and then also $\mathbf{M} = \mathbf{D} + (\mathbf{W} - \mathbf{D})\tilde{\mathbf{R}}$. □

Remark 2.8.1. Thus, if $\mathbf{Q}$ is strictly positive in each entry (other than the diagonals) and all agents are (strictly) conforming and $\mathbf{W}$ has at least two positive columns, then $\mathbf{M}$ is strictly positive in each entry. This means that $\mathbf{M}^t$ is strictly positive in each entry, for all powers $t \geq 1$. This also means that $\lim_{t\to\infty} \mathbf{M}^t$, which exists for a row-stochastic $\mathbf{M}$ that is strongly connected and aperiodic (as an $\mathbf{M}$ with strictly positive entries is), is positive in each entry. But this means that, under conformity, if agents form their reference opinions $\mathbf{q}$, to which they strive to conform, with respect to all other agents (that is, $\mathbf{Q}$ is strictly positive, except for the diagonals), then each agent i has strictly positive social influence on the limiting beliefs, even if i has never been truthful in the past. The amount of influence (even non-truthful) agents have on limiting beliefs depends then both on past performance *and* on the conformity parameters δ_i. We illustrate in the next example.

Example 2.8.2. Assume the following situation. There are $n = 3$ agents and $\mathbf{W}^{(k)}$ has the form

$$\mathbf{W}^{(k)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

Such a $\mathbf{W}^{(k)}$ may arise, for instance, for large k, when $\tau = 0$ and when agents 1 and 2 have initial beliefs centered around truth, with identical variances, and agent 3 has never been truthful, for instance, because $\Pr[b_3^k(0) \in B_{k,\eta}] = 0$, for all $k \geq 1$. Assume that

$$\mathbf{Q}^{(k)} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix},$$

which means that everyone weighs everyone else uniformly in forming reference opinions, and assume that $\mathbf{Q}^{(k)}$ is constant across topics. Finally, let $\delta_1 = \delta_2 = a$ and $\delta_3 = b$, where $0 < a, b < 1$. Consider

the social influence of agents — that is, their influence on limiting beliefs as a function of initial beliefs. Since agents 1 and 2 are identical, they must have the same social influence, which we denote by $x \geq 0$, and let agent 3 have influence $y \geq 0$, such that $2x + y = 1$. In the appendix, we show that y has the form

$$y = \frac{a(1-b)}{4 - ab - 3a}.$$

Computing the comparative statics with respect to a and b, we find

$$\frac{\partial y}{\partial a} = \frac{(1-b)\Big(4 - ab + 3b\Big)}{\Big(4 - ab - 3a\Big)^2} > 0, \quad \frac{\partial y}{\partial b} = \frac{a(a-7)}{\Big(4 - ab - 3a\Big)^2} < 0.$$

Thus, influence of agent 3 decreases in own conformity, b, and increases in conformity of the stochastically intelligent agents, a. Moreover, we find

$$\lim_{a \to 1} y = 1, \quad \lim_{b \to 1} y = 0,$$

such that agent 3, who has zero probability of knowing truth, may have arbitrarily large social influence on limiting beliefs, as long as agents 1 and 2 exhibit arbitrarily large conformity, and agent 3's social influence may also vanish, as his own conformity becomes arbitrarily large. In Figure 2.11, we plot y as a function of a (for fixed levels of b) and b (for fixed levels of a). Note that this result, namely, that

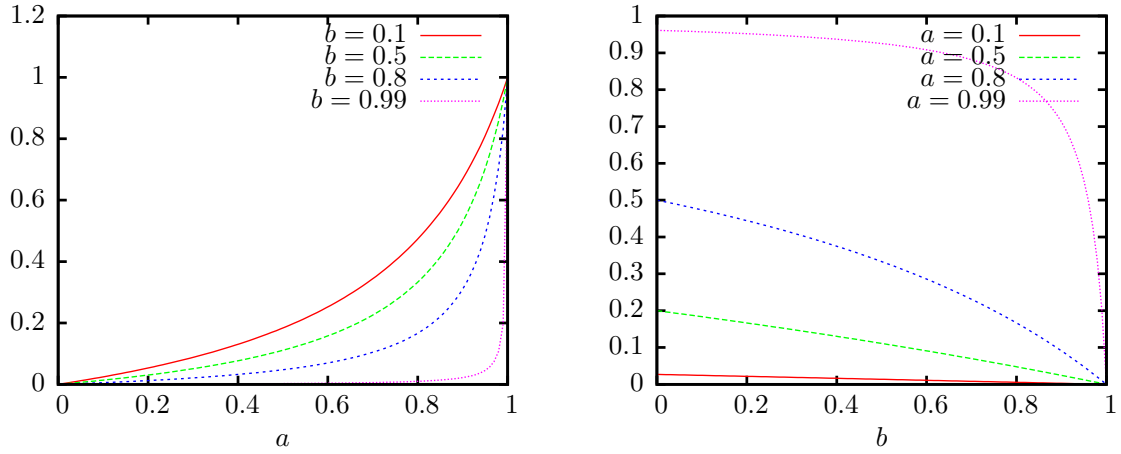


Figure 2.11: Social influence y of agent 3 as a function of a, conformity of agents 1 and 2, and b, own conformity.

an agent with zero probability of being truthful may become arbitrarily influential, could not have been possible in the standard model with endogenous weight formation as we have sketched, since such an agent's social influence would always be zero (or converge to zero) by the results developed in Section 2.6, as k becomes large. This result would also not have been possible had there been only one positive column of matrix $\mathbf{W}^{(k)}$ (and the others all zero) since in this case the corresponding agent, even if he were excessively conforming and would thus state an opinion arbitrarily close to that of his reference group, would both ignore his own stated opinion (because he knows his true opinion) and that of others (because all other columns are zero, by assumption) such that the agent corresponding to the positive column would always have social influence of 1. In other words, we require at least two conforming intelligent agents in order for a non-intelligent to be able to become influential, under our current model specification.

To summarize, our current example shows that if conformity parameters are 'adequately' specified, an agent with zero probability of being close to truth may become arbitrarily influential, which, again,

means that society may be arbitrarily far off from truth in our setup. In particular, under conformity, society may be drawn away from truth by agents without any past successes.[18]

But, of course, a crucial aspect in the current example has also been matrix $\mathbf{Q}$, which determines how agents form reference opinions, and which agents a particular agent strives to conform to, and which we have assumed strictly positive. Of course, it might not be implausible to assume that $\mathbf{Q}$ depends, in particular, on past performance. This is our final example.

Example 2.8.3. Let $[\mathbf{Q}^{(k)}]_{ij} = \frac{[\mathbf{W}^{(k)}]_{ij}}{1-[\mathbf{W}^{(k)}]_{ij}}$ for $i \neq j$ and let $[\mathbf{Q}^{(k)}]_{ii} = 0$ such that $\mathbf{Q}$ is formed in an analogous way as $\mathbf{W}$.[19] In particular, in this setup, agents want to conform to those who have performed well in the past. Assume that there are two types of agents, whereby one type has initial beliefs centered around truth as in (2.6.1) and the other type has zero probability of ever being truthful, $\Pr[b_i^k(0) \in B_{k,\eta}] = 0$ for all k and all $i \in \mathcal{N}_2$ where $\mathcal{N}_2 \subseteq [n]$ is the set of agents of type two, and by $\mathcal{N}_1$ we denote the set of agents of type one. Then, if $\tau = 0$ and as k becomes large, $\mathbf{W}^{(k)}$ has the structure where in each row i, $W_{ij}^{(k)} \approx \frac{1}{C\sigma_j^2}$ for $j \in \mathcal{N}_1$ (C is a normalization constant) and $W_{ij}^{(k)} \approx 0$ for $j \in \mathcal{N}_2$, by the results developed in Section 2.6, and where σ_j^2 is the variance of agent j's initial belief. The informational social influence of agent i is then also given by, roughly, $w_i = \frac{1}{C\sigma_i^2}$, for $i \in \mathcal{N}_1$, and $w_i = 0$, for $i \in \mathcal{N}_2$, respectively. Moreover, the combined informational as well as normative social influence of agents is given by

$$v_i = \frac{(1-\delta_i)\cdot w_i}{\sum_{j\in\mathcal{N}_1}(1-\delta_j)w_j}$$

for agents $i \in \mathcal{N}_1$ and $v_i = 0$ for agents $i \in \mathcal{N}_2$. These results follow directly from Theorem 1 and Corollary 1 given in Buechel, Hellmann, and Klößner (2013), which precisely state that closed and strongly connected groups (which $\mathcal{N}_1$ is, at least for large k) have social influence v_i as given and the 'rest of the world', which group $\mathcal{N}_2$ forms, has $v_i = 0$.

This example shows that if $\mathbf{Q}$ follows the structure of $\mathbf{W}$, then, unlike in the previous example where $\mathbf{Q}$ was uniform (or at least strictly positive on the off-diagonals), agents that never know truth cannot be influential. It moreover shows that social influence decreases in conformity (for the agents in $\mathcal{N}_1$), as we have already observed in the previous example.

Investigating social influence in the general case, for arbitrary $\mathbf{Q}, \mathbf{W}$, and $\mathbf{\Delta}$, would be highly interesting, as it indicates to which degree agents who are never truthful can still be influential, and scope for future work.[20]

2.9 Homophily

In this section, we extend the standard endogenous weight adjustment opinion dynamics model discussed in Section 2.6 by introduction of the concept of *homophily*, according to which, as McPherson, Smith-Lovin, and Cook (2001) phrase it, 'similarity breeds connection', and which is a majorly accepted standard concept in modern socio-economic research. In the opinion dynamics literature, homophily has been modeled by positing that weights (social ties) between any two agents are functionally dependent on the agents' current belief distance (cf. the Hegselmann and Krause models, Deffuant et al., 2000; Pan, 2010, etc.), that is, agents with more similar current beliefs place greater (current) weight upon each other. Opinion updating is then performed as in standard DeGroot learning, via weighted averages of peers' past beliefs, where the weights are now endogenously formed by the homophily principle. As indicated in the introduction, we think of homophily, in our context, as arising from biased reasoning, where individuals overrate beliefs that are similar to their own (cf. Kunda, 1990).

[18] Which might be an illustration of why even 'blatantly' and repetitively false propaganda may work.

[19] As mentioned, this is the specification discussed in Buechel, Hellmann, and Klößner (2013).

[20] Proposition 2.8.4 is an important step in this direction already, since it says that if $\mathbf{Q}^{(k)}$ is strictly positive on all off-diagonals, $\mathbf{\Delta}^{(k)}$ is strictly positive in all diagonals, and $\mathbf{W}^{(k)}$ has two positive columns, then *all* agents are influential.

In the Hegselmann and Krause models, to which we relate, agents set *time-varying* weights according to the following rule,[21]

$$W_{ij}(t) = \begin{cases} \frac{1}{|I_i(\mathbf{b}(t))|} & \text{if } j \in I_i(\mathbf{b}(t)), \\ 0 & \text{else,} \end{cases}$$

where $I_i(\mathbf{b}(t))$ denotes the set of agents within an η_H-radius, for $\eta_H \geq 0$, around agent i's belief $b_i(t)$ at time t, that is, $I_i(\mathbf{b}(t)) = \{j \in [n] \mid \|b_j(t) - b_i(t)\| < \eta_H\}$, and where η_H is an external parameter. One plausible integration of this setup in our framework is to let agents *increment* weights to other agents whenever distance between their current beliefs is sufficiently small, that is,[22,23]

$$W_{ij}^{(k)}(t+1) = \begin{cases} W_{ij}^{(k)}(t) + \delta_H & \text{if } j \in I_i(\mathbf{b}(t)), \\ W_{ij}^{(k)}(t). \end{cases} \tag{2.9.1}$$

Then, after truth is revealed, agents again adjust weights according to the 'truth related' principles outlined in Section 2.3. In particular, we would now have,

$$W_{ij}^{(k+1)}(0) = \begin{cases} \lim_{t\to\infty} W_{ij}^{(k)}(t) + \delta_T \cdot T(\left|N(\mathbf{b}^k(\tau), \mu_k)\right|) & \text{if } \left\|b_j^k(\tau) - \mu_k\right\| < \eta_T, \\ \lim_{t\to\infty} W_{ij}^{(k)}(t) & \text{otherwise,} \end{cases} \tag{2.9.2}$$

for all $k \geq 1$, where we need to consider, for next topic's initial weights, the limit, as time goes to infinity, of the time-varying weights for the previous topic, since weights are now also adjusted within topic periods. Note that, here, we also subscript η — the radius within which weights are adjusted — and δ — the weight increment — by T and H, respectively, depending on whether we relate to adjusting/incrementing based on truth or based on homophily.

Also observe that (2.9.2) is well-defined only if $\lim_{t\to\infty} W_{ij}^{(k)}(t)$ exists, which we naïvely assume in the following but the formal proof of which we leave open. It is worthwhile mentioning that adjusting weights based on truth may microeconomically be justified precisely as we did in Section 2.4 — namely, it may follow from the tenet that agents have disutility from not knowing truth, whence, by incrementing weights to agents who have been truthful in the past, they increase their likelihood of eventually becoming close to truth, provided that the assumptions they make (*bona fides*, etc.) are satisfied. In contrast, we offer no explicit microeconomic foundation — that is, based on utility functions and their explicit maximization — here of why agents would increment weights to other agents based on the homophily relation, taking this behavior simply as a form of (exogenous) bias. Moreover, similar as in the opposition model, it is appropriate, in the current setting, to think of agents as motivated by two contrarian forces — truth and homophily —[24] which may possibly act to the 'same ends', but which we generally think of as of antipodal origin and direction.

Concerning updating of beliefs, beliefs evolve according to

$$b_i^k(t+1) = \sum_{j=1}^{n} W_{ij}^{(k)}(t) b_j^k(t), \tag{2.9.3}$$

as outlined in Section 2.3, with the addition that we now let weights $W_{ij}^{(k)}$ vary within discussion periods. Due to the slightly greater complexity involved in belief dynamics, we summarize the belief evolution process in the below schematic form.

Providing general results for the belief dynamics process currently under consideration is not so easy since weights do not only vary by time now, but, in particular, by the current belief state vector $\mathbf{b}^{(k)}(t)$. Lorenz (2005) gives convergence results for this general setup, which we list in Appendix 2.A, whose

[21] Note that, thus far, we have assumed weights to be only varying across *topics* and not in addition across discussion *rounds* (time) within a given topic.

[22] Otherwise, if weights were not incremented but rather set in an 'absolute manner', the continuity of weight relationships across topics could not be maintained.

[23] After each round t, we *renormalize* weights in order for them to satisfy the row-stochasticity condition.

[24] Note that in the opposition model, the two forces were an agent's ingroup and his outgroup.

let $\mathbf{W}^{(1)}(0)$ be (exogenously) given
for $k = 1, 2, 3, \ldots$ **do**
  let $b_i^k(0)$ denote initial beliefs for topic X_k for all agents $i = 1, \ldots, n$
  for $t = 0, 1, 2, 3, 4, \ldots$ **do**
    adjust weights $\mathbf{W}^{(k)}(t)$ based on homophily, Eq. (2.9.1); normalize weights
    update beliefs $b_i^k(t+1)$ for all agents i via Eq. (2.9.3)
  end for
  adjust weights $\mathbf{W}^{(k+1)}(0)$ based on truth, Eq. (2.9.2); normalize weights
end for

assumptions, however, do not apply to our situation. As a first illustration, still, we show that, unlike in the standard model and under conformity, even after some agent has been truthful for a topic X_r, agents need not converge to a consensus, but may hold distinct limiting beliefs about X_{r+1}; whether consensus obtains or not may depend on the relative sizes of δ_T, which we may think of as 'importance of truth', versus δ_H, which we may think of as 'importance of homophily'.

Example 2.9.1. Let there be $n = 4$ agents. Assume that $\mathbf{W}^{(1)}(0)$ is the $n \times n$ identity matrix. Moreover, let $\tau = 0$ and assume that agent 1 receives initial signal $b_1^1(0) = \mu_1$ and that agents 2, 3 and 4 are not within an η_T-radius around truth, for topic X_1. Finally, assume that any two agents' initial beliefs $b_i^1(0)$ and $b_j^1(0)$ are at least a distance of η_H away from each other for topic X_1 so that homophily plays no role for topic X_1. Then, at the beginning of topic X_2, agents adjust weights based on truth such that $\mathbf{W}^{(2)}(0)$ looks as follows

$$\mathbf{W}^{(2)}(0) = \frac{1}{1+\delta_T} \begin{pmatrix} 1+\delta_T & 0 & 0 & 0 \\ \delta_T & 1 & 0 & 0 \\ \delta_T & 0 & 1 & 0 \\ \delta_T & 0 & 0 & 1 \end{pmatrix}.$$

Assume that initial beliefs of agents for topic X_2 are $b_1^2(0) = b_2^2(0)$ and $b_3^2(0) = b_4^2(0)$, whereby initial beliefs of agents 1 and 2, on the one hand, and 3 and 4, on the other hand, are at a distance of at least η_H. This specification means that agents 1 and 2, on the one hand, and 3 and 4, on the other, form distinct 'homophily clusters', at least at time $t = 0$, for topic X_2. In Figure 2.12, we sketch belief dynamics for topic X_2 for different values of δ_H, with $\delta_T = 0.1$ and $\eta_H = \eta_T = 0.2$ fixed. We see that, unlike in the standard DeGroot learning case in this setup (and also in the conformity model) and as already indicated, agents do not necessarily reach a consensus. If homophily is 'too strong', that is, δ_H is 'too large', agents polarize in this setting. As homophily becomes weaker, that is, δ_H becomes smaller, the beliefs of agents 3 and 4 move closer to the beliefs of agents 1 and 2, the former of which has been truthful for topic X_1. As δ_H falls below a certain threshold, the agents reach a consensus.

To sketch one (simple) result of a general nature, here, however, consider, similarly as before, the situation when there are two groups $\mathcal{N}_1$ and $\mathcal{N}_2$ of agents, whereby agents in $\mathcal{N}_1$ are ϵ-intelligent, for a fixed $\epsilon \geq 0$, and agents in $\mathcal{N}_2$ have initial beliefs $b_i^k(0)$ with $\Pr[b_i^k(0) \in B_{k,\eta_T}] = 0$, that is, agents in $\mathcal{N}_2$ have initial beliefs such that the probability that they 'correspond to' truth is zero. In the next proposition, we show that *all* agents *may* become ϵ-wise, even if $\delta_H > 0$, in this situation provided that agents value truth 'sufficiently much' *and* value relations based on homophily sufficiently little. This result is not entirely trivial because, for instance, for our opposition model, arbitrarily small 'opposition force' could induce (at least some) agents to not converge to truth. The result says that homophily does not *always* need to interfere with wisdom.

Proposition 2.9.1. Let $\eta_T \geq 0$, $\epsilon \in [0, \eta_T]$, and topic X_k be fixed, for $k \geq 2$.[25] Then there exist $\delta_T > 0$ large enough and $\delta_H > 0$ small enough such that all agents become ϵ-wise for X_k.[26]

[25] For topic X_1, there are no weight adjustments based on truth, so we exclude this situation.

[26] Of course, if δ_H were zero, any positive δ_T would satisfy the conditions of the proposition. In our setup, we assume, however, that homophily always plays a role, that is, $\delta_H > 0$ for all 'homophily increments' δ_H.

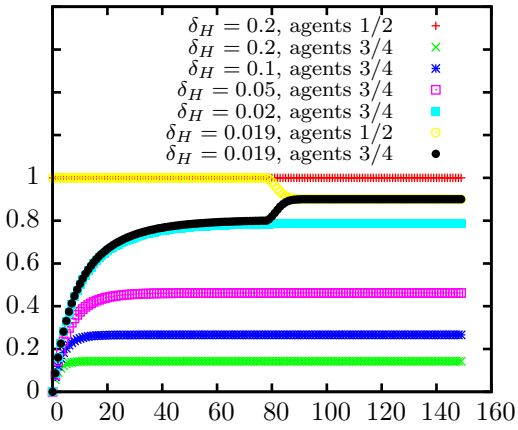


Figure 2.12: Belief dynamics for topic X_2 with setup as sketched in Example 2.9.1. Initial beliefs are $b_1^2(0) = b_2^2(0) = 1$, $b_3^2(0) = b_4^2(0) = 0$. As long as beliefs of agents 3 and 4 do not come within distance of $\eta_H = 0.2$ to the beliefs of agents 1 and 2, the latters' beliefs evolve according to $b_1^2(t) = b_2^2(t) = 1$ since both agents have the same initial beliefs and are not 'disturbed' by agents 3 and 4. In contrast, beliefs of agents 3 and 4 are affected by agent 1's beliefs since agent 1 has been truthful for topic X_1, but their weight link to this agent vanishes as t becomes large if δ_H is 'large enough'. In general, we have $b_1^2(t) = b_2^2(t)$ and $b_3^2(t) = b_4^2(t)$ for all $t \geq 0$ so that it suffices to graph belief dynamics of agents 1, on the one hand, and 3, on the other.

Proof. Since the agents in $\mathcal{N}_1$ are ϵ-intelligent, with $\epsilon \leq \eta_T$, all agents adjust weights for these agents based on truth. By choosing δ_T large enough and δ_H small enough (but positive), it can be ensured that beliefs $[\mathbf{b}^k(1)]_i$, for all $i \in [n]$, are in the (open) ϵ-interval around truth μ_k (the weights for the ϵ-intelligent agents may become arbitrarily close to uniform provided δ_T is large enough and δ_H is small enough and the weights for the agents in $\mathcal{N}_2$ may become arbitrarily close to zero). Since $\mathbf{W}^{(k)}(t)$ is row-stochastic, for every $t \geq 0$, and since the (open) interval of radius ϵ around truth is a convex set, all belief vectors $\mathbf{b}^k(t)$, for $t \geq 1$, lie, component-wise, in $B_{k,\epsilon}$. □

Remark 2.9.1. In the last proposition, $\delta_T = \delta_T(k)$ and $\delta_H = \delta_H(k)$ may depend upon the topic X_k (e.g., in particular on the distribution of initial beliefs for this topic). If we let, $\delta_T := \max_{k\in\mathbb{N}} \delta_T(k)$ and $\delta_H := \min_{k\in\mathbb{N}} \delta_H(k)$,[27] then for this choice of δ_T and δ_H, all agents will be ϵ-wise for all topics X_k, for $k \geq 2$.

Example 2.9.2. We illustrate Proposition 2.9.1 in Figure 2.13, where we sketch belief dynamics for a sequence of topics for fixed parametrizations and various choices of δ_T and δ_H. In the figure, we simulate belief dynamics across topics for $n = 50$ agents, where $n_1 = 10$ agents are ϵ-intelligent and $n_2 = 40$ agents have initial distribution of beliefs such that their initial beliefs are never in an η_T interval around truth; for the sake of concreteness, we let $\mu_k = 0$, for all $k \geq 1$, $\epsilon = 0.25$, $\eta_T = 0.25$, and for the agents in $\mathcal{N}_2$, we let their initial beliefs be distributed according to the random uniform distribution on the interval $[1, 4]$.

The graphs illustrate, first, that truth attracts all agents since even the beliefs of the agents in $\mathcal{N}_2$ move in the direction of truth, as time progresses. However, as long as preference for homophily δ_H is not small enough and preference for truth δ_T is not large enough, the agents in $\mathcal{N}_2$ do not become ϵ-wise for topics. The graphs also illustrate the clustering of beliefs due to the homophily relationship, a circumstance well-known from the classical Hegselmann-Krause models.

Remark 2.9.2. The graphs in Figure 2.13 show much analogy with results of the original opinion dynamics model 'under homophily' as developed in the work of Hegselmann and Krause, on which our current modeling is based (recall that the difference is that we *increment* weights in case two agents'

[27] This would require to ensure that the so defined δ_H is strictly positive (rather than zero) and that $\delta_T < \infty$.

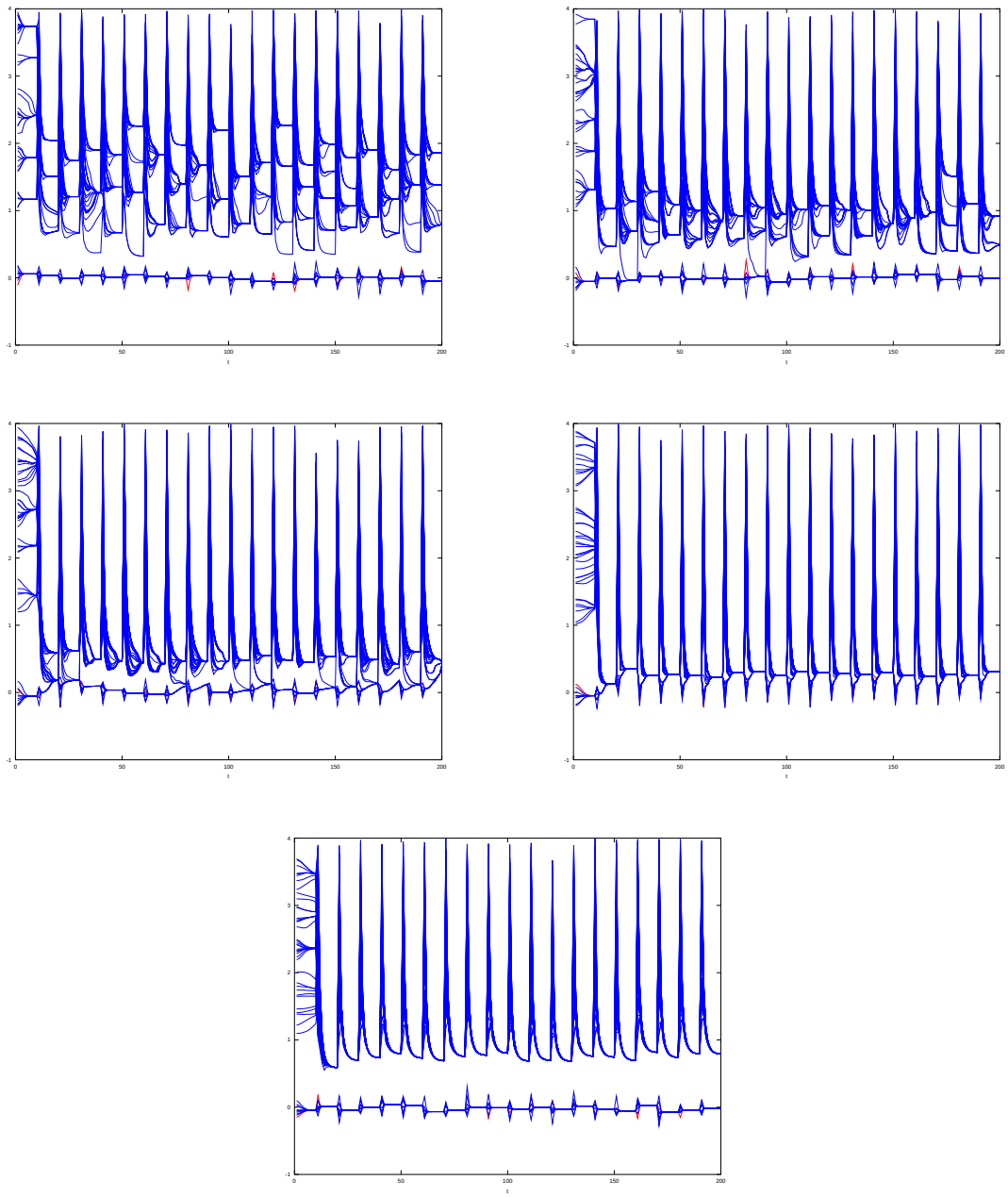


Figure 2.13: Parametrizations throughout: $\epsilon = 0.25$, $\mu_k = 0$ for all $k \geq 1$, $\eta_T = \eta_H = 0.25$, $n = 50$ agents, $|\mathcal{N}_1| = 10$, $|\mathcal{N}_2| = 40$. From top to bottom and left to right: $(\delta_H, \delta_T) = (0.2, 1.0)$; $(\delta_H, \delta_T) = (0.1, 1.0)$; $(\delta_H, \delta_T) = (0.05, 1.0)$; $(\delta_H, \delta_T) = (0.02, 1.0)$; $(\delta_H, \delta_T) = (0.02, 0.1)$. We show topics X_k, for $k = 1, 2, \ldots, 20$ and, for each topic, discussion rounds $t = 0, 1, \ldots, 10$.

beliefs are similar, while they set weights uniformly in this case, and that we in addition introduce truth as an influential factor). In particular, in the graphs, we find that

- the opinion dynamics process always converges, and that
- agents (or, rather, their beliefs) cluster into subgroups in which agents reach a consensus.

Proving these apparently generally true observations is beyond the scope of our investigation here, and

we leave it for future consideration.[28]

We close this section by presenting simulations on the role of the truth related radius η_T and the homophily related radius η_H, respectively. Concerning η_H, we find in Figure 2.14 that a smaller η_H (that is, based on homophily, agents trust/listen to only those with very similar beliefs) tends to produce a larger degree of fragmentation of limiting belief spectra while larger η_H (that is, based on homophily, agents even trust/listen to agents with rather distinct beliefs) tends to promote global agreement among agents. Interestingly, smaller η_H also leads agents closer to truth (since the homophily relation applies to fewer agents). Concerning η_T, in Figure 2.15, we find that, overall, an increase in η_T increases the average distance of limiting beliefs to truth since also beliefs that are remote from truth are taken into consideration in link weight adjustment.

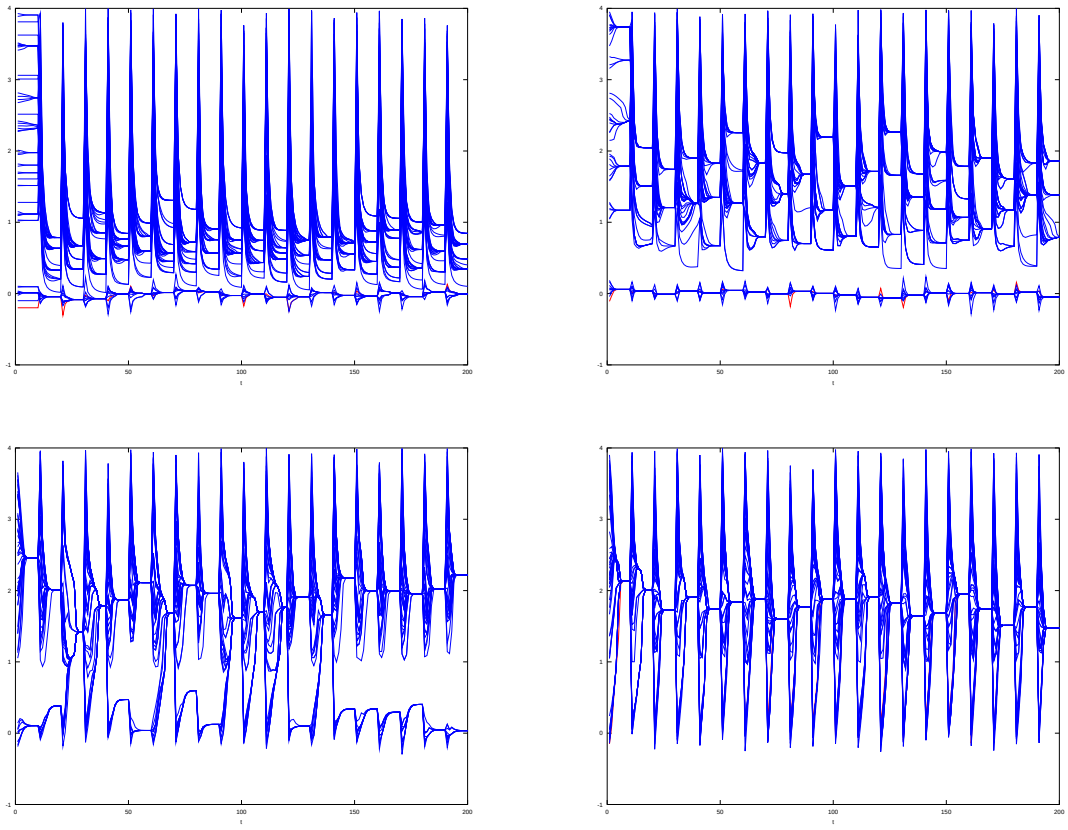


Figure 2.14: Parametrizations throughout: $\epsilon = 0.25$, $\mu_k = 0$ for all $k \geq 1$, $\eta_T = 0.25$, $\delta_H = 0.2$, $\delta_T = 1.0$, $n = 50$ agents, $|\mathcal{N}_1| = 10$, $|\mathcal{N}_2| = 40$. From top to bottom and left to right: $\eta_H = 0.05$, $\eta_H = 0.25$, $\eta_H = 1.10$, $\eta_H = 1.50$. We show topics X_k, for $k = 1, 2, \ldots, 20$ and, for each topic, discussion rounds $t = 0, 1, \ldots, 10$.

In sum, in this section, we have shown that, under the 'homophily bias' and under the presence of agents with biased initial beliefs, agents need neither become wise nor reach a consensus. If the homophily relation is sufficiently 'weak', wisdom may obtain (Proposition 2.9.1), but if it is sufficiently 'strong', agents' beliefs will generally cluster into distinct regions of the belief spectrum. As in the conformity model (and also as under opposition), even agents with zero probability of being close to truth may influence others.

[28] See Krause (2000) for a starting point on how to prove the results in question.

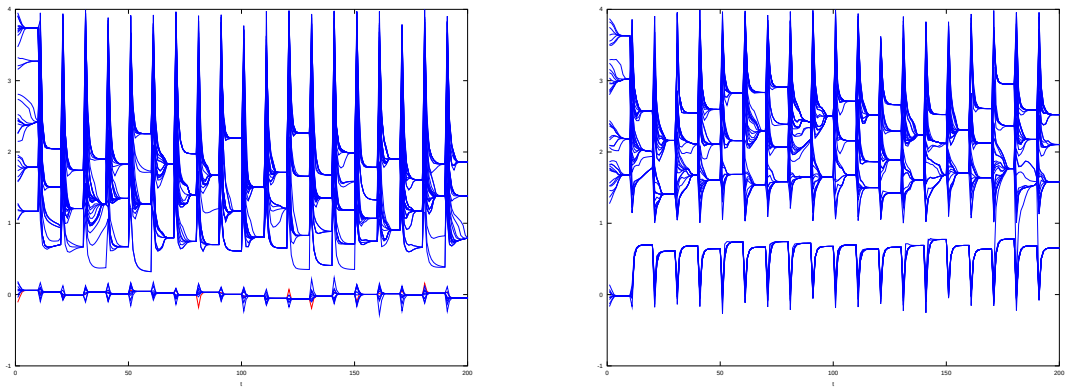


Figure 2.15: Parametrizations throughout: $\epsilon = 0.25$, $\mu_k = 0$ for all $k \geq 1$, $\eta_H = 0.25$, $\delta_H = 0.2$, $\delta_T = 1.0$, $n = 50$ agents, $|\mathcal{N}_1| = 10$, $|\mathcal{N}_2| = 40$. Left: $\eta_T = 0.25$. Right: $\eta_T = 2.50$. We show topics X_k, for $k = 1, 2, \ldots, 20$ and, for each topic, discussion rounds $t = 0, 1, \ldots, 10$.

2.10 Conclusion

As Acemoglu and Ozdaglar (2011), and many others, point out, the importance of the beliefs we hold cannot be overstated. For example, the demand for a product depends on consumers' opinions and beliefs about the quality of that product and majority opinions determine the political agenda. Thus, beliefs also shape (our) behavior in that they lead us to buy certain products and reject others or in that they are causal factors for the implementation of laws and policies. On a more abstract level, the set of norms and beliefs we hold determine, in the end, who we are and substantiate our cultural foundations. In modern microeconomic research, beliefs and opinions are thought to originate from *social learning* processes whereby individuals are situated in a network of peers and update their opinions, e.g., via communication with others. Rejecting the hypothesis that individuals are fully rational, much recent research has assumed that people learn from others via simple 'rules of thumb', simply averaging peers' past beliefs to arrive at new beliefs. Then, given that there exist 'true states' for the issues that individuals hold beliefs about, a natural question to ask is whether such agents, who commit the bias of not properly accounting for the repetition of information they hear, can, in fact, still learn these true states and, thus, become collectively 'wise' (cf. Surowiecki, 2004), successfully aggregating dispersed information.

In the current work, we have studied belief dynamics under an *endogenous* network formation process. In particular, we have assumed that agents strengthen their ties to other agents based on the criterion of 'past performance' such that agents increment their trust weights to whoever has been 'close enough' to truth for a current topic. We have, moreover, assumed that agents are *multiply biased* in that they are not only susceptible to persuasion bias — the simplifying DeGroot learning rule — but also have biased initial beliefs (the possibly non-social, 'intelligence-based' substrate of beliefs), and commit several other sins of reasoning, such as being biased toward members of their in-group and motivated to disassociate from members of their out-group, being motivated to conform with the beliefs of their reference group, or overrating beliefs that are close to their own. Our goal has been to outline situations under which collective failure (or at least, 'failure of wisdom') can obtain, even though the potential for wisdom — dispersed correct information — is assured. Thus, our work was also in part targeted at the recent 'optimism' concerning biased ('naïve') learning in social networks and crowd wisdom (e.g., Golub and Jackson, 2010), which has also been challenged by experimental research (cf., e.g., Lorenz et al., 2011).

As to our results, under the standard DeGroot learning model, we have seen that wisdom can fail if there are *sufficiently many* agents with biased initial beliefs such that they, still, have positive probability of being close to truth. The intuition behind this result is that even if the biased agents have small, but positive, probability of 'guessing' truth, then, if they are sufficiently many — such that many of them will still be close to truth — the biased agents can, in total, receive large enough weight mass from all agents,

whence they may become arbitrarily socially influential, leading all of society to the expected value of a biased variable, away from truth. This result may be thought of as based on the *bona fides* bias, which says that agents do not give up the assumption that their own (initial) beliefs are unbiased and that others' beliefs share this property with their own, even despite potential collective failure, from which it may be motivated that agents continuously apply their trust weight incrementing rule (to all agents). In the conformity model, wisdom may fail even when the biased agents have zero probability of being close to truth and when their number is small, provided that the unbiased agents are sufficiently conforming. We might take this as an argument for why even 'blatantly' and repetitively false propaganda could work. A necessary condition for this result is that agents want to conform to reference groups including even biased and completely unknowing agents (which might be justified on grounds/biases such as truth-unrelated prominence, e.g., due to political power or popularity). In the opposition model, wisdom can fail even if all agents' initial beliefs are unbiased and, in addition, arbitrarily close to truth, merely as a consequence of agents being attracted by contrarian forces — their in-groups, on the one hand, which attract them toward truth, and their out-groups, on the other, from which they want to disassociate. In the homophily model, wisdom can fail because agents are, again, influenced by antagonistic forces — truth, on the one hand, and agents with similar beliefs, on the other. Hence, biased agents' beliefs may cluster, if they form a homogenous group, and unbiased agents' beliefs may also cluster, so that some agents would become wise and others not.

Concerning future research directions within our context, of course, endogenizing several (more) of the parameters of the DeGroot learning models that we have discussed might be of interest. In the current work, we have solely endogenized the social network, without explaining, for example, where in-group/out-group antagonisms actually come from or how conformity may develop and how reference groups evolve. The endogenizing of such parameters would plausibly require psychological and socio-economic motivations that are independent of the criterion of 'past performance'. Moreover, in our model, we have generally assumed that agents are *homogenous* with respect to many dimensions of attributes such as their truth tolerances η, trust weight increments δ, etc., and a heterogenous setup may provide further insight. Finally, introducing strategic agents (cf., e.g., Anderlini, Gerardi, and Lagunoff, 2012), that potentially have incentives to deliberatively *mislead* others, might be a promising research direction to incorporate in our general setup of social learning and collective wisdom/failure.

Appendix 2.A Proofs

Standard model

Lemma 2.A.1. If matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has identical rows with row sum $s = \sum_{j=1}^{n} A_{ij}$, then $\mathbf{A}^t = s^{t-1}\mathbf{A}$ for any $t \geq 1$.

Proof. Follows by induction. □

Lemma 2.A.2. Consider any $n \times n$ matrix $\mathbf{A}$ of the form

$$\mathbf{A} = \begin{pmatrix} \beta & \alpha & \dots & \alpha \\ \alpha & \beta & \dots & \alpha \\ \vdots & \dots & \ddots & \vdots \\ \alpha & \alpha & \dots & \beta \end{pmatrix} \tag{2.A.1}$$

with $\alpha, \beta \in \mathbb{R}$. The eigenvalues of matrix $\mathbf{A}$ are given by $\lambda_1 = \beta + (n-1)\alpha$ and $\lambda_2 = \dots = \lambda_n = \beta - \alpha$.

Proof. We first consider the determinant of $\mathbf{A} = \mathbf{A}(n)$. Subtracting the second row from the first, we find $\det(\mathbf{A}(n)) = (\beta - \alpha)\det(\mathbf{A}(n-1)) + (\beta - \alpha)\det(\mathbf{B}(n-1))$, where $\mathbf{B}(n)$ is the $n \times n$ matrix with $[\mathbf{B}(n)]_{ij} = [\mathbf{A}(n)]_{ij}$ for all i, j with $(i, j) \neq (1, 1)$; for $(i, j) = (1, 1)$, we have $[\mathbf{B}(n)]_{ij} = \alpha$. Proceeding analogously as for $\mathbf{A}(n)$, we find $\det(\mathbf{B}(n)) = (\beta - \alpha)^{n-1}\alpha$. Therefore,

$$\det(\mathbf{A}(n)) = (\beta - \alpha)^{n-1}\big(\beta + (n-1)\alpha\big).$$

Now, consider the characteristic polynomial of $\mathbf{A}(n)$; it is $\chi(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}_n)$. Note that $\mathbf{A} - \lambda \mathbf{I}_n$ is a matrix of the form (2.A.1). Hence, its determinant is given by

$$\chi(\lambda) = \Big((\beta - \alpha) - \lambda\Big)^{n-1} \Big(\beta + \alpha(n-1) - \lambda\Big).$$

This concludes the proof. □

Wisdom of crowds under initial beliefs centered around truth

The following are results from Golub and Jackson (2010). They state conditions under which a growing population, parametrized by its size n, converges to truth μ under the assumption that agents receive initial belief signals that are centered around μ (as in (2.6.1)). The statement of the below results is that agents become (ϵ-)wise (for any $\epsilon > 0$) if and only if the *influence* of the most influential agent converges to zero as n increases, whereby an agent's influence is given by his *social influence*, as we have discussed above and as we define below. In undirected networks ($W_{ij} = W_{ji}$ for all $i, j \in [n]$) with uniform weights, this condition is tantamount to all agents' relative degrees (the number of links they have to other agents divided by the total number of links in the network) converging to zero as n becomes large. Hence, in this setup, an obstacle to wisdom would be the circumstance when each agent who newly enters society assigns, e.g., a constant fraction of his links to a particular agent, who would then become excessively influential.

Remark 2.A.1. If a social network $\mathbf{W}$ induces a consensus, then limiting beliefs can be represented as $\mathbf{b}(\infty) = \mathbf{s}^\intercal \mathbf{b}(0)$, for a non-negative vector $\mathbf{s}$ with $\sum_{i=1}^{n} s_i = 1$ which we call the *social influence vector* and s_i agent i's *influence*. The influence vector is given as the unique normalized unit-vector $\mathbf{s}$ which satisfies $\mathbf{s} = \mathbf{W}^\intercal \mathbf{s}$ (i.e., $\mathbf{s}$ is the normalized unit-eigenvector of $\mathbf{W}^\intercal$ corresponding to the eigenvalue $\lambda = 1$).

Now, as in Golub and Jackson (2010), we parametrize social networks $\mathbf{W}$ by their population size n, which we denote by $\mathbf{W}(n)$; we also parametrize other quantities such as limiting beliefs of a set of agents by population size n (here and in the following, we omit reference to topics k for notational convenience). Moreover, we denote a *society* by the sequence $\big(\mathbf{W}(n)\big)_{n\in\mathbb{N}}$. We restate the following lemma and the proposition from Golub and Jackson (2010), which they list as Lemma 1 and Proposition 2.

Lemma 2.A.3 (A law of large numbers)**.** If $\big(\mathbf{s}(n)\big)_{n\in\mathbb{N}}$ is any sequence of influence vectors, then

$$\mathbf{s}(n)^\intercal \mathbf{b}(0; n) \to \mu \quad \text{as } n \to \infty$$

(where convergence is in probability or almost surely) if and only if $s_1(n) \to 0$, where we assume, without loss of generality, that $s_1(n) \geq s_2(n) \geq \cdots \geq s_n(n)$.

Proposition 2.A.1. If $\big(\mathbf{W}(n)\big)_{n\in\mathbb{N}}$ is a sequence of networks, each inducing a consensus, then the underlying agents become (ϵ-)wise (for any $\epsilon > 0$) as $n \to \infty$ if and only if the associated influence vectors are such that $s_1(n) \to 0$ as $n \to \infty$.

We now argue informally that the proposition entails convergence to truth in the situation where agents' initial beliefs are centered around truth as in (2.6.1) and in our setup of endogenous weight formation.[29] Namely, we first argue that an agent's influence s_i is directly inversely proportional to his variance σ_i^2. Although a proof thereof would require technical sophistication, the claim appears very intuitive since influence s_i captures weight mass assigned to an agent by other agents (in addition to these agents' influence; cf. DeMarzo, Vayanos, and Zwiebel, 2003) and, in our setup, the weight mass that an agent receives is directly inversely proportional to his variance σ_i^2 (more intelligent agents receive weight increases more often). Next, consider networks $\big(\mathbf{W}(n)\big)_{n\in\mathbb{N}}$ where, for all $n \in \mathbb{N}$, agents' variances

[29] We assume that k is so large that each network $\mathbf{W}^{(k)}(n)$ always induces a consensus. Note that, if agents are stochastically intelligent, a consensus is reached quickly (and increasingly fast in the number of agents n), by the results developed in Section 2.6.

σ_i^2 satisfy $\sigma_i^2 \geq \bar{\sigma}^2 > 0$ for some lower bound $\bar{\sigma}^2 > 0$.[30] Then, as $n \to \infty$, the influence of the most influential agent certainly goes to zero since the number of agents increases (all of which are influential in the sense that they receive weight mass from others) while the expected weight mass that the most influential agent receives is bounded.

Varying weights on own beliefs

Proof of Proposition 2.6.9. Since $\mathbf{W} = \mathbf{W}^{(k)}$ converges for all initial belief vectors $\mathbf{b}(0)$, there exists a matrix $\mathbf{W}^\infty$ such that $\lim_{t\to\infty} \mathbf{W}^t = \mathbf{W}^\infty$. To prove the proposition, show that $\prod_{s=0}^{t-1} \mathbf{W}(\lambda_s)$ converges to $\mathbf{W}^\infty$ as $t \to \infty$, whereby $\mathbf{W}(\lambda) = \Big((1-\lambda)\mathbf{I} + \lambda\mathbf{W}\Big)$ and where $\mathbf{b}(t) = \Big(\prod_{s=0}^{t-1} \mathbf{W}(\lambda_s)\Big)\mathbf{b}(0)$ according to (2.6.4). Proceed exactly as in DeMarzo, Vayanos, and Zwiebel (2003).

Define the random variable Λ_t to be equal to 1 with probability λ_t and zero otherwise. Assume also that Λ_t are independent over time. Define the random matrix $\mathbf{Z}_t$ by $\mathbf{Z}_t = \prod_{s=0}^{t-1} \mathbf{W}(\Lambda_s) = \mathbf{W}^{\sum_{s=0}^{t-1} \Lambda_s}$. Then $\mathbb{E}[\mathbf{Z}_t] = \prod_{s=0}^{t-1} \mathbf{W}(\lambda_s)$. By the Borel-Cantelli lemma, if $\sum_{t=0}^{\infty} \Pr[\Lambda_t = 1] = \sum_{t=0}^{\infty} \lambda_t = \infty$, then

$$\Pr[\sum_{t=0}^{\infty} \Lambda_t = \infty] = \Pr[\Lambda_t = 1 \text{ infinitely often}] = 1.$$

Since the matrix $\mathbf{W}^t$ is bounded uniformly in t, the dominated convergence theorem implies that

$$\lim_{t\to\infty} \prod_{s=0}^{t-1} \mathbf{W}(\lambda_s) = \lim_{t\to\infty} \mathbb{E}[\mathbf{Z}_t] = \lim_{t\to\infty} \mathbb{E}[\mathbf{W}^{\sum_{s=0}^{t-1} \Lambda_s}] = \mathbf{W}^\infty.$$

□

Opposition

Lemma 2.A.4. Consider any matrix of the form (2.7.4) with $a, b, c, d > 0$ and such that $\sum_{j=1}^{n} |A_{ij}| = 1$ for all $i = 1, \ldots, n$. Let $n_1 = 1$. Then, the characteristic polynomial of $\mathbf{A}$ is given by

$$\chi(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}_n) = (-\lambda)^{n-2}\Big(\lambda^2 - (a + (n-1)d)\lambda + (n-1)(ad - bc)\Big) = (-\lambda)^{n-2}(\lambda - 1)(\lambda - q),$$

where $q = (n-1)(ad - bc) = a - 1 + (n-1)d$.

Proof. Expanding the determinant along the last row (and subtracting the second-to-last row from the last), we find that the determinant $\det(\mathbf{B}_n)$ of $\mathbf{B}_n = \mathbf{A}_n - \lambda\mathbf{I}_n$, with $\mathbf{A}_n = \mathbf{A}$, is given by

$$-\lambda \det(\mathbf{B}_{n-1}) - \lambda \det(\mathbf{C}_{n-1})$$

whereby $\mathbf{C}_n = \mathbf{A}_n - \lambda\mathbf{I}_n$, except for the entry in row n and column n, which is $[\mathbf{C}_n]_{nn} = A_{nn}$. The determinant of $\mathbf{C}_n$ can easily be found to be $(-\lambda)^{n-2} \cdot \Big((a-\lambda)d - bc\Big)$. Then solving $\det(\mathbf{A}_n)$ inductively leads to the required solution. Finally, the factorization of the quadratic polynomial results from the fact that $\mathbf{A}$ has one eigenvalue of 1, as can readily be checked. □

From Lemma 2.A.4, we can infer that matrix $\mathbf{A}$ from (2.7.4) has $n-2$ eigenvalues 0, one eigenvalue of 1, and one eigenvalue q, which is a real eigenvalue. Moreover, all eigenvalues of $\mathbf{A}$ are bounded from above by 1 (cf. Eger, 2013, Proposition 6.3). Assume that q were -1. Then

$$a - 1 + (n-1)d = q = -1 \quad \Longleftrightarrow \quad a + (n-1)d = 0 \quad \Longleftrightarrow \quad a = -(n-1)d,$$

whence a is negative, which contradicts $a > 0$. Thus, assume q were $+1$. Then

$$a + (n-1)d = 2,$$

which contradicts $a + (n-1)d = a + n_2 d < 1 + 1 = 2$ (since both b and c are positive and recall the row sum restrictions $n_1 a + n_2 b = 1$, etc.). Therefore, $\lambda = 1$ is the only eigenvalue of $\mathbf{A}$ on the unit circle and it has algebraic multiplicity of 1.

[30] Should there be no lower bound on the most intelligent agent's variance, then this agent may become excessively influential but his initial beliefs also become arbitrarily accurate, so that society becomes (ϵ-)wise simply because one agent is arbitrarily well-informed.

Conformity

Lemma 2.A.5. Consider $\mathbf{I}_n - \mathbf{A}$ for an $n \times n$ matrix $\mathbf{A}$. If $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$, then $\mathbf{I}_n - \mathbf{A}$ is invertible and its inverse is given by the *Neumann series*

$$(\mathbf{I}_n - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k.$$

Proof. See Meyer (2000), p.618. □

Proof of Proposition 2.8.1. Our proof follows along the lines of the proof of the corresponding proposition of Buechel, Hellmann, and Klößner (2012).

The best response s_i^* of player i to the strategies s_{-i} of the other players is given by the first order conditions,

$$\frac{\partial u_i(s_i, s_{-i}; b_i)}{\partial s_i}|_{s_i=s_i^*} = -2(1-\delta_i)(s_i^* - b_i) - 2\delta_i\Big(s_i^* - \sum_{j\neq i} Q_{ij}s_j\Big) = 0$$

for all $i \in [n]$. Note that the best response is unique. A strategy profile $\mathbf{s}^* \in S^n$ is a Nash equilibrium if and only if s_i^* is a best response to $\mathbf{s}_{-i}^*$. Thus, Nash equilibria $\mathbf{s}^* \in S^n$ satisfy:

$$(\mathbf{I}_n - \boldsymbol{\Delta})(\mathbf{s}^* - \mathbf{b}) + \boldsymbol{\Delta}(\mathbf{s}^* - \mathbf{Q}\mathbf{s}^*) = (\mathbf{I}_n - \boldsymbol{\Delta})(\mathbf{s}^* - \mathbf{b}) + \boldsymbol{\Delta}(\mathbf{I}_n - \mathbf{Q})\mathbf{s}^* = \mathbf{0}.$$

Rewriting leads to

$$\mathbf{s}^* = (\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q})^{-1}(\mathbf{I}_n - \boldsymbol{\Delta})\mathbf{b},$$

which is well-defined since $\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q}$ is invertible by Lemma 2.A.5. Namely, we have

$$\|\boldsymbol{\Delta}\mathbf{Q}\|^k \leq \|\boldsymbol{\Delta}\|^k \|\mathbf{Q}\|^k \leq \underbrace{\Big(\max_{i\in[n]} |\delta_i|\Big)^k}_{=:\delta_{\max}} \|\mathbf{Q}\|^k$$

for any matrix norm $\|\cdot\|$. Hence,

$$0 \leq \lim_{k\to\infty} \|\boldsymbol{\Delta}\mathbf{Q}\|^k \leq \lim_{k\to\infty} (\delta_{\max})^k \|\mathbf{Q}\|^k = 0,$$

since $|\delta_i| < 1$ by assumption, for all $i \in [n]$, and $\|\mathbf{Q}\|^k$ is bounded since $\mathbf{Q}$ is row-stochastic. Therefore, $\lim_{k\to\infty}(\boldsymbol{\Delta}\mathbf{Q})^k = \mathbf{0}$. □

Proof of Lemma 2.8.1. Consider $\mathbf{M}\mathbb{1}$ (which is $\mathbf{M} \cdot \mathbb{1}$), which is

$$\mathbf{D}\mathbb{1} + (\mathbf{W} - \mathbf{D})(\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q})^{-1}(\mathbf{I}_n - \boldsymbol{\Delta})\mathbb{1}.$$

It suffices to show that

$$\mathbf{R}\mathbb{1} := (\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q})^{-1}(\mathbf{I}_n - \boldsymbol{\Delta})\mathbb{1} = \mathbb{1}$$

because of row-stochasticity of $\mathbf{W}$, which entails that $\mathbf{W}\mathbb{1} = \mathbb{1}$.

Now, we have $(\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q})^{-1} = \sum_{r=0}^{\infty}(\boldsymbol{\Delta}\mathbf{Q})^r$ by row-stochasticity of $\mathbf{Q}$ and since $|\delta_i| < 1$. Hence

$$\begin{aligned}
\mathbf{R}\mathbb{1} &= \sum_{r=0}^{\infty}(\boldsymbol{\Delta}\mathbf{Q})^r(\mathbf{I}_n - \boldsymbol{\Delta})\mathbb{1} = (\mathbf{I}_n - \boldsymbol{\Delta})\mathbb{1} + \sum_{r=1}^{\infty}(\boldsymbol{\Delta}\mathbf{Q})^{r-1}[\boldsymbol{\Delta}\mathbf{Q}\mathbb{1} - \boldsymbol{\Delta}\mathbf{Q}\boldsymbol{\Delta}\mathbb{1}] \\
&= (\mathbf{I}_n - \boldsymbol{\Delta})\mathbb{1} + \sum_{r=1}^{\infty}(\boldsymbol{\Delta}\mathbf{Q})^{r-1}[\boldsymbol{\Delta}\mathbb{1} - \boldsymbol{\Delta}\mathbf{Q}\boldsymbol{\Delta}\mathbb{1}] = (\mathbf{I}_n - \boldsymbol{\Delta})\mathbb{1} + (\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q})^{-1}[\mathbf{I}_n - \boldsymbol{\Delta}\mathbf{Q}]\boldsymbol{\Delta}\mathbb{1} \\
&= (\mathbf{I}_n - \boldsymbol{\Delta})\mathbb{1} + \boldsymbol{\Delta}\mathbb{1} = \mathbb{1}.
\end{aligned}$$

□

Proposition 2.A.2. In the situation of Example 2.8.2, the social influence weights x, x and y of agents $1, 2$ and 3 are given by

$$x = \frac{2(1-a)}{4-ab-3a}, \quad \text{and} \quad y = \frac{a(1-b)}{4-ab-3a}.$$

Proof. The social influence weights can be found by computing $\mathbf{M}$ and then solving $\mathbf{M}^\intercal \mathbf{x} = \mathbf{x}$ where $\mathbf{x} = (x, x, y)^\intercal$. The computation, though cumbersome, is straightforward. □

Homophily

The following theorem is the 'stabilization theorem' of Lorenz (2005). It discusses convergence of the opinion dynamics process $\mathbf{b}(t+1) = \mathbf{W}(\mathbf{b}(t), t)\mathbf{b}(t)$, where weight matrix $\mathbf{W}$ may depend on time t and the current vector of beliefs $\mathbf{b}(t)$, as in the homophily model we have sketched. We abbreviate the theorem to fit our needs.

Theorem 2.A.1 (Lorenz, 2005)**.** Let $(\mathbf{W}(t))_{t\in\mathbb{N}}$ be a sequence of row-stochastic matrices. If each matrix $\mathbf{W}(t)$ satisfies

(1) $[\mathbf{W}(t)]_{ii} > 0$ for all $i \in [n]$ ('each agent has a little bit of self-confidence'),

(2) $[\mathbf{W}(t)]_{ij} > 0 \iff [\mathbf{W}(t)]_{ji} > 0$ ('confidence is mutual'),

(3) there exists $\kappa > 0$ (that is independent of t) such that the smallest positive entry of $\mathbf{W}(t)$ is greater than κ ('positive weights do not converge to zero'),

then $\lim_{t\to\infty} \mathbf{b}(t)$ exists, that is, the belief dynamics process converges.

While Theorem 2.A.1 applies, in particular, to the Hegselmann and Krause models, on which our homophily model rests, it does not apply to the latter. This is easy to see: while condition (1) in the theorem on $\mathbf{W}(t)$ is satisfied in our case (due to $\delta_H > 0$ and $\left\|b_i^k(t) - b_i^k(t)\right\| = 0 < \eta_H$ for any positive η_H), both conditions (2) and (3) may be violated in our modeling. Condition (2) may be violated because of truth related weight adjustment, which is generally asymmetric (agent i may have been true for a topic X_k, while j may not have been true so that j increases his weight for i while i does not increase his weight for j); and condition (3) may be violated because a positive link weight between two agents may converge to zero in our model, e.g., when an agent i has known truth for a topic, so that another agent j increases his link weight for i (based on truth), but i and j's beliefs are sufficiently distinct such that homophily, toward other agents, causes the link weight $[\mathbf{W}(t)]_{ji}$ to drop to zero, as $t \to \infty$.

Appendix 2.B Experiment

Below, we list details on the experiment indicated in the introduction. In total, $n = 119$ subjects, all from `Amazon Mechanical Turk`, participated in the experiment; not all subjects answered all questions.[31] We set a time limit for answering the 16 'common knowledge' questions of 3 minutes and reimbursed subjects with 60 US cents if they completed and submitted the questions (this required them to press the 'submit' button rather than to indeed answer all questions), which corresponds to an hourly wage of 12 USD. Obviously, this was an attractive wage, since all requested slots (119) were filled within approximately one hour. On average, individuals took 2 minutes and 25 seconds to answer all 16 questions, including reading the instructions and optionally providing feedback, although some subjects complained that time limits were too narrow. Below, we summarize the instructions, the questions, and give histograms of the distributions of answers (Figure 2.16) as well as of the 'logarithmically scaled' data — for some questions, individuals beliefs' seemed to be lognormally distributed, so we provide these histograms in Figure 2.17. We note that we — very slightly — adjusted the data when it very obviously seemed to be corrupted. For example, one person gave as average daily temperature in Miami in July the number 8856347, which

[31] The data set is available upon request.

cannot plausibly be correct; similarly, two people answered the question concerning the age of homo sapiens sapiens as 80 years, which constitutes most probably a misunderstanding of the question.

From the histograms in Figures 2.16 and 2.17, we observe that people's beliefs appear to be centered around truth only occasionally. In particular, for example, the histogram for the question concerning the average height of an adult male US American appears to be consistent with independent normal distributions, centered around truth, as underlying subjects' beliefs. For the question regarding the number of official languages of the European union, the population density of Beijing, and the distance from earth to moon, independent lognormal distributions appear as plausible. As we have already discussed in the introduction, neither the mean nor the median are very reliable quantities for the true values of questions, as Table 2.2 illustrates.

Instructions
Give truthful estimates on 16 questions such as "When did the first settlers arrive in America?". Don't look them up, don't google them. We're interested in your honest estimate/guess, not in your ability to use search engines. If you don't know the correct answer, please try to provide your best guess. Please answer all 16 questions.
A valid answer to the above question might be "in 1620" (if this is what you think the correct answer is). Certainly don't take longer than 20 seconds to answer any one question.
If your answer requires a unit such as "pounds", "miles", or "kilometers", please indicate it, for the sake of clarity.

(1)	the average daily temperature in Miami in July (in Fahrenheit or Celsius)?	87.8F
(2)	the population size of New York city, as of 2012?	$8,336,697$
(3)	the current level of the Dow Jones stock market index?	$15,658.36$
(4)	the number of official languages in the European Union?	24
(5)	the age of modern humans (homo sapiens sapiens) on earth?	
	In other words, since how long do (modern) humans exist on earth?	$200,000$
(6)	the year the first world war started?	1914
(7)	the number of McDonald's restaurants in the US?	$12,804$
(8)	the number of people per square mile (or square kilometer) in China's capital Beijing?	$3,300$/sqm
(9)	the how many-th US president was Bill Clinton?	42nd
(10)	the average height of an adult male in the US as of 2012? (in feet and inches or centimeters)	177.8cm
(11)	the distance from earth to the moon (in miles or kilometres)?	$238,610$m
(12)	the number of states the United States consists of?	50
(13)	the fraction of population in the US that is left-handed in percent?	10.5%
(14)	the average life-span of an African elephant (in years) in the wild?	56
(15)	$17 - 4 \times 2$?	9
(16)	the diameter of the sun in miles or kilometers?	$857,490$m

Table 2.1: Questions, to be preceded by 'What do you think is ...', and 'true' answers. Most 'true' answers are taken from `wikipedia` or similar resources.

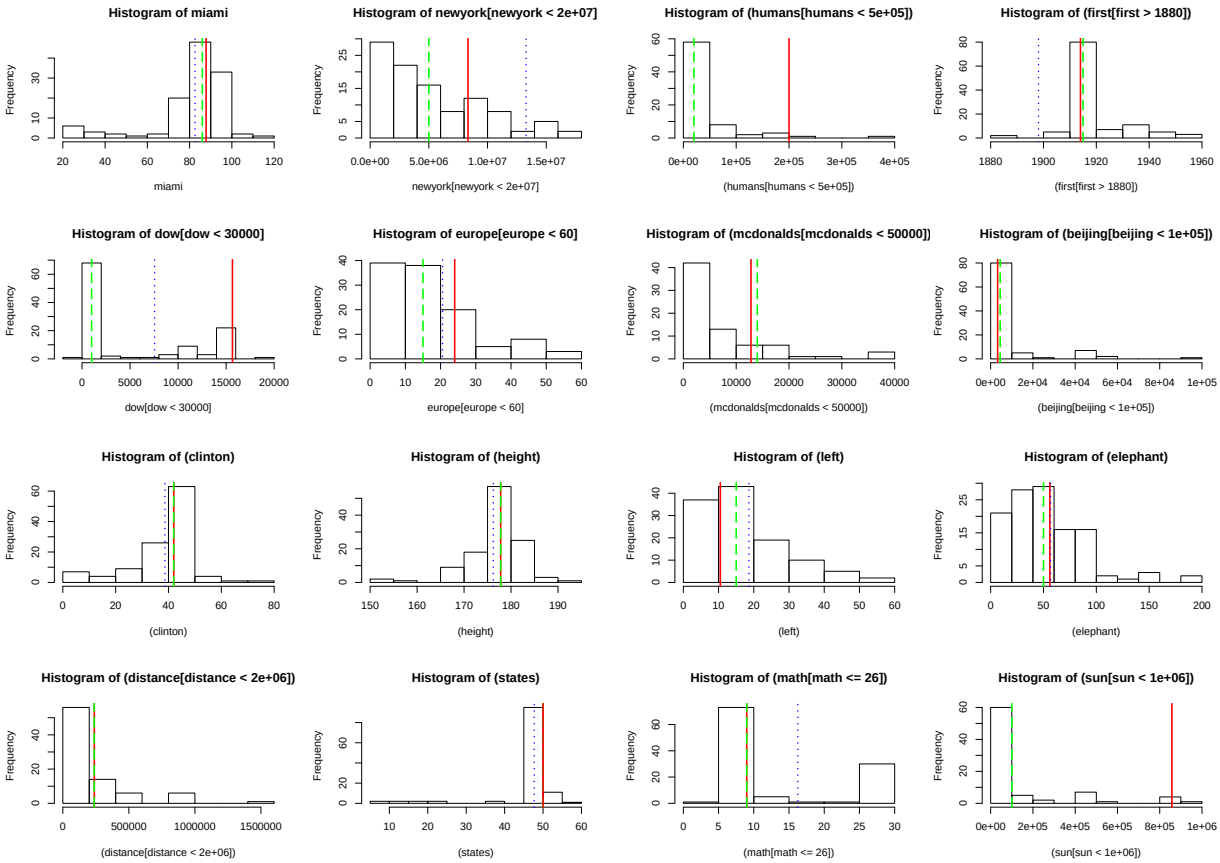


Figure 2.16: Histograms of answers to questions (1) to (16), top to bottom and left to right. Mean (dotted blue), median (dashed green), and truth (solid red) indicated.

	Median				Mean			
	1%	2%	5%	10%	1%	2%	5%	10%
(1)			x	x				x
(2)								
(3)								
(4)								
(5)								
(6)	x	x	x	x	x	x	x	x
(7)				x				
(8)								
(9)	x	x	x	x				x
(10)	x	x	x	x	x	x	x	x
(11)	x	x	x	x				
(12)	x	x	x	x			x	x
(13)								
(14)						x	x	x
(15)	x	x	x	x				
(16)								
	6	6	7	8	2	3	4	6

Table 2.2: Question numbers and indication whether (x) or not median or mean are within the indicated intervals around truth.

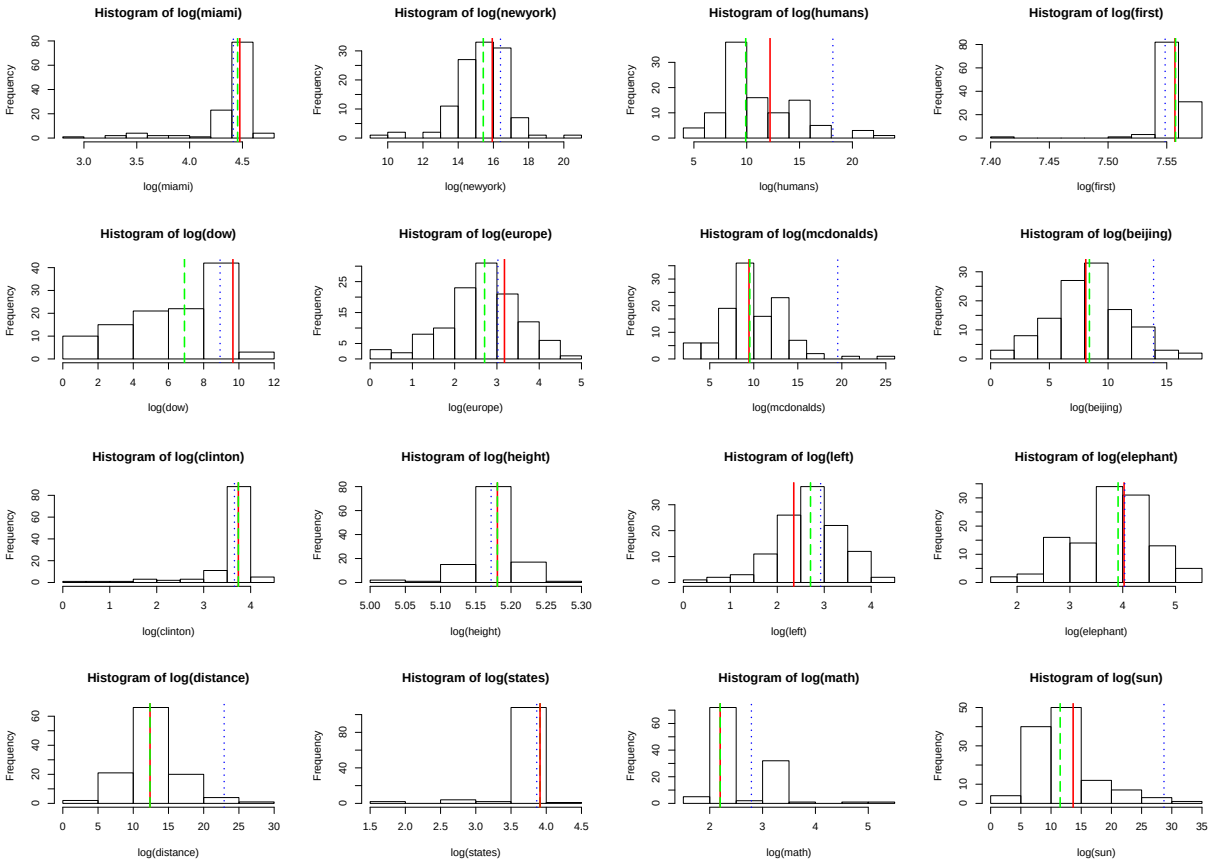


Figure 2.17: Histograms of log(answers) to questions (1) to (16), top to bottom and left to right. Mean (dotted blue), median (dashed green), and truth (solid red) indicated.

Bibliography

[1] Daron Acemoglu, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar. *Opinion Fluctuations and Disagreement in Social Networks*. LIDS report 2850. to appear in *Mathematics of Operations Research*. 2012. URL: `http://web.mit.edu/asuman/www/documents/disagreementsubmitted.pdf`.

[2] Daron Acemoglu, Munzer A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. "Bayesian Learning in Social Networks". In: *Review of Economic Studies* 78 (4 2011), pp. 1201–1236.

[3] Daron Acemoglu and Asuman Ozdaglar. "Opinion Dynamics and Learning in Social Networks". In: *Dynamic Games and Applications* 1 (1 2011), pp. 3–49.

[4] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. "Spread of (Mis)Information in Social Networks". In: *Games and Economic Behavior* 70 (2 2010), pp. 194–227.

[5] Luca Anderlini, Dino Gerardi, and Roger Lagunoff. "Communication and learning". In: *Review of Economic Studies* 79 (2 2012), pp. 419–450.

[6] Solomon Asch. "Opinions and social pressure". In: *Scientific American* 27 (5 1955), pp. 31–35.

[7] Bahador Bahrami, Karsten Olsen, Dan Bang, Andreas Roepstorff, Geraint Rees, and Chris Frith. "What failure in collective decision-making tells us about metacognition". In: *Philsophical Transactions of the Royal Society B* 367 (2012), pp. 1350–1365.

[8] Abhijit V. Banerjee. "A simple model of herd behavior". In: *Quarterly Journal of Economics* 107 (3 1992), pp. 797–817.

[9] Abhijit V. Banerjee and Drew Fudenberg. "Word-of-mouth learning". In: *Games and Economic Behavior* 46 (3 2004), pp. 1–22.

[10] David Beasley and Dan Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge, UK: Cambridge University Press, 2010.

[11] Philip Bonacich. "Factoring and weighting approaches to status scores and clique identification". In: *The Journal of Mathematical Sociology* 2 (1 1972), pp. 113–120.

[12] Marylinn B. Brewer. "In-Group Bias in the minimal intergroup situation: A cognitive-motivational analysis". In: *Psychological Bulletin* 86 (2 1979), pp. 307–324.

[13] David V. Budescu, Adrian K. Rantilla, Hsiu-Ting Yu, and Tzur M. Karelitz. "The effects of asymmetry among advisors on the aggregation of their opinions". In: *Organizational Behavior and Human Decision Processes* 90 (1 2003), pp. 178–194.

[14] Berno Buechel, Tim Hellmann, and Stefan Klößner. "Opinion Dynamics and Wisdom under Conformity". Working Paper. 2013.

[15] Berno Buechel, Tim Hellmann, and Stefan Klößner. "Opinion Dynamics under Conformity". Working Paper. 2012.

[16] Berno Buechel, Tim Hellmann, and Michael Pichler. "The Dynamics of Continuous Cultural Traits in Social Networks". Working Paper. 2012.

[17] Zhigang Cao, Mingmin Yang, Xinglong Qu, and Xiaoguang Yang. "Rebels Lead to the Doctrine of the Mean: Opinion Dynamic in a Heterogeneous DeGroot Model". In: *The 6th International Conference on Knowledge, Information and Creativity Support Systems*. Beijing, China, 2011, pp. 29–35.

[18] Emanuele Castano, Vincent Yzerbyt, David Bourguignon, and Eléonore Seron. "Who may Enter? The Impact of In-Group Identification on In-Group/Out-Group Categorization". In: *Journal of Experimental Social Psychology* 38 (2002), pp. 315–322.

[19] Arun G. Chandrasekhar, Horacio Larreguy, and Juan P. Xandri. *Testing models of Social Learning on Networks: Evidence from a lab experiment in the field*. Working Paper. 2012.

[20] Gary Charness, Luca Rigotti, and Aldo Rustichini. "Individual Behavior and group membership". In: *American Economic Review* 97 (4 2007), pp. 1340–1352.

[21] Marquis de Concordet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris, France: de l'Impr. Royale, 1785.

[22] Luca Corazzini, Filippo Pavesi, Beatrice Petrovich, and Luca Stanca. "Influential listeners: An experiment on persuasion bias in social networks". In: *European Economic Review* 56 (6 2012), pp. 1276–1288.

[23] Guillaume Deffuant, David Neau, Frederic Amblard, and Gerard Weisbuch. "Mixing beliefs among interacting agents". In: *Advances in Complex Systems* 3 (2000), pp. 87–98.

[24] Morris H. DeGroot. "Reaching a Consensus". English. In: *Journal of the American Statistical Association* 69.345 (1974), pp. 118–121. ISSN: 01621459. URL: `http://www.jstor.org/stable/2285509`.

[25] Peter M. DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. "Persuasion Bias, Social Influence, And Unidimensional Opinions". In: *The Quarterly Journal of Economics* 118.3 (Aug. 2003), pp. 909–968. URL: `http://ideas.repec.org/a/tpr/qjecon/v118y2003i3p909-968.html`.

[26] Morton Deutsch and Harold B. Gerard. "A study of normative and informational social influences upon individual judgement". In: *Journal of Abnormal Psychology* 51 (3 1955), pp. 629–636.

[27] Igor Douven and Alexander Riegler. "Extending the Hegselmann-Krause model I". In: *The Logic Journal of the IGPL* 18 (2 2010), pp. 323–335.

[28] Igor Douven and Alexander Riegler. "Extending the Hegselmann-Krause model II". In: *Proceedings of ECAP 2009* (2009).

[29] Igor Douven and Alexander Riegler. "Extending the Hegselmann-Krause model III: From single beliefs to complex belief states". In: *Episteme* 6 (2009), pp. 145–163.

[30] Steffen Eger. *Opinion dynamics under opposition*. http://arxiv.org/pdf/1306.3134v2.pdf. 2013.

[31] Hillel J. Einhorn, Robin M. Hogarth, and Eric Klempner. "Quality of Group Judgment". In: *Psychological Bulletin* 84 (1 1977), pp. 158–172.

[32] Sebastian Fehrler and Michael Kosfeld. "Can You Trust the Good Guys? Trust Within and Between Groups with Different Missions". Working Paper. 2013.

[33] Leon Festinger. *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson, 1957.

[34] John R.P. French. "A formal theory of social power". In: *Psychological Review* 63 (3 1956), pp. 181–194.

[35] Noah E. Friedkin and Eugene C. Johnsen. "Social influence and opinions". In: *Journal of Mathematical Sociology* 15 (3-4 1990), pp. 193–205.

[36] Noah E. Friedkin and Eugene C. Johnsen. "Social influence networks and opinion change". In: *Advances in Group Processes* 16 (1999), pp. 1–29.

[37] Douglas Gale and Shachar Kariv. "Bayesian Learning in Social Networks". In: *Games and Economic Behavior* 45 (2 2003), pp. 329–346.

[38] Francis Galton. "Vox populi". In: *Nature* 75 (1907), pp. 450–451.

[39] Benjamin Golub and Matthew O. Jackson. "How Homophily Affects the Speed of Learning and Best-Response Dynamics". In: *The Quarterly Journal of Economics* 127 (3 2012), pp. 1287–1338.

[40] Benjamin Golub and Matthew O. Jackson. "Naïve Learning in Social Networks and the Wisdom of Crowds". In: *American Economic Journal: Microeconomics* 2 (1 2010), pp. 112–149.

[41] Sajeev Goyal. "Learning in networks: a survey". In: *Group formation in economics; Networks, Clubs, and Coalitions*. Ed. by G. Demage and M. Wooders. Cambridge U.K.: Cambridge University Press, 2004. Chap. 4.

[42] Patrick Groeber, Jan Lorenz, and Frank Schweitzer. "Dissonance minimization as a microfoundation of social influence in models of opinion formation". In: *Journal of Mathematical Sociology* (2013).

[43] Frank Harary. "A criterion for unanimity in French's theory of social power". In: *Studies in social power* (1959).

[44] Rainer Hegselmann and Ulrich Krause. "Opinion dynamics and bounded confidence: models, analysis and simulation". In: *J. Artificial Societies and Social Simulation* 5.3 (2002).

[45] Rainer Hegselmann and Ulrich Krause. "Opinion Dynamics Driven by Various Ways of Averaging". In: *Computational Economics* 25.4 (2005), pp. 381–405. URL: `http://EconPapers.repec.org/RePEc:kap:compec:v:25:y:2005:i:4:p:381-405`.

[46] Rainer Hegselmann and Ulrich Krause. "Truth and Cognitive Division of Labour: First Steps Towards a Computer Aided Social Epistemology". In: *Journal of Artificial Societies and Social Simulation* 9.3 (2006), p. 10. ISSN: 1460-7425. URL: `http://jasss.soc.surrey.ac.uk/9/3/10.html`.

[47] Matthew O. Jackson and Alison Watts. "On the Formation of Interaction Networks in Social Coordination Games". In: *Games and Economic Behavior* 41 (2 2002), pp. 265–291.

[48] Irving L. Janis. *Victims of Groupthink*. Boston, MA: Houghton Mifflin, 1972.

[49] Stephen R.G. Jones. *The economics of conformism*. Oxford; New York, NY: B. Blackwell, 1984.

[50] Daniel Kahnemann and Amos Tversky. "Choices, Values, and Frames". In: *American Psychologist* 39 (4 1984), pp. 341–350.

[51] Norbert L. Kerr, Robert J. MacCoun, and Geoffrey P. Kramer. "Bias in Judgment: Comparing Individuals and Crowds". In: *Psychological Review* 103 (4 1996), pp. 687–719.

[52] Norbert L. Kerr and R. Scott Tindale. "Group Performance and Decision Making". In: *Annual Review of Psychology* 55 (2004), pp. 623–655.

[53] James A. Kitts. "Social influence and the emergence of norms amid ties of amity and enmity". In: *Simulation Modelling Practice and Theory* 14 (2006), pp. 407–422.

[54] Ulrich Krause. "A discrete nonlinear and non-autonomous model of consensus formation". In: *Communications in Difference Equations*. Ed. by S. Elaydi, G. Ladas, J. Popenda, and R. Rakowski. Amsterdam: Gordon and Breach Publ., 2000, pp. 227–236.

[55] Ziva Kunda. "The Case for Motivated Reasoning". In: *Psychological Bulletin* 108 (3 1990), pp. 480–498.

[56] Keith Lehrer. "Rationality as Weighted Averaging". In: *Synthese* 57 (1983), pp. 283–295.

[57] Keith Lehrer and Carl Wagner. "Rational consensus in science and society". In: *Reidel, Dordrecht* (1981).

[58] Joel Lobel. "Economists' Models of Learning". In: *Journal of Economic Theory* 94 (2000), pp. 241–261.

[59] Jan Lorenz. "A stabilization theorem for dynamics of continuous opinions". In: *Physica A: Statistical Mechanics and its Applications* 355.1 (2005), pp. 217–223. DOI: `10.1016/j.physa.2005.02.086`.

[60] Jan Lorenz. "Continuous opinion dynamics of multidimensional allocation problems under bounded confidence. More dimensions lead to better chances for consensus". In: *European Journal of Economic and Social Systems* 19 (2006), pp. 213–227.

[61] Jan Lorenz. "Continuous opinion dynamics under bounded confidence: A survey". In: *International Journal of Modern Physics C* (2007).

[62] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. "How social influence can undermine the wisdom of crowd effect". In: *Proceedings of the National Academy of Sciences* 108.22 (2011), pp. 9020–9025. DOI: 10.1073/pnas.1008636108.

[63] Charles Mackay. *The extraordinary and popular delusions and the madness of crowds*. 4th. Ware, UK: Wordsworth Editions Limited, 1841.

[64] Albert E. Mannes. "Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision". In: *Management Science* 55 (8 2009), pp. 1267–1279.

[65] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. "Birds of a feather: Homophily in Social Networks". In: *Annual Review of Sociology* 27 (2001), pp. 415–444.

[66] Carl D. Meyer. *Matrix analysis and applied linear algebra*. Philadelphia: SIAM, 2000.

[67] Zhengzheng Pan. "Trust, influence, and convergence of behavior in social networks". In: *Mathematical Social Sciences* 60 (1 2010), pp. 69–78.

[68] D. Rosenberg, E. Solan, and N. Vieille. "Informational Externalities and Convergence of Behavior". Preprint. 2006.

[69] James Surowiecki. *The wisdom of crowds. Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. London: Little, Brown, 2004.

[70] Amos Tversky and Daniel Kahnemann. "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185 (4157 1974), pp. 1124–1131.

[71] Carl Wagner. "Consensus through respect: A model of rational group decision-making". In: *Philosophical studies: An international Journal for Philosophy in the Analytic Tradition* 34 (4 1978), pp. 335–349.

[72] Gwen M. Wittenbaum and Garold Stasser. "Management of information in small groups". In: *What's Social about Social Cognition*. Ed. by J.L. Nye and A.M. Brower. Thousand Oaks, CA: Sage, 1996, pp. 967–978.

[73] Ilan Yaniv. "The benefit of additional opinions". In: *American Psychological Society* 13 (2 2004), pp. 75–78.

[74] Ilan Yaniv and Eli Kleinberger. "Advice taking in decision making: Egocentric discounting and reputation formation". In: *Organizational Behavior and Human Decsion Processes* 83 (2 2000), pp. 260–281.

[75] Ercan Yildiz, Daron Acemoglu, Asuman Ozdaglar, Amin Saberi, and Anna Scaglione. *Discrete Opinion Dynamics with Stubborn Agents*. LIDS report 2870. to appear in *ACM Transactions on Economics and Computation*. 2012. URL: http://web.mit.edu/asuman/www/documents/voter-submit.pdf.

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

[77] Bo-Yu Zhang, Zhi-Gang Cao, Cheng-Zhong Qin, and Xiao-Guang Yang. *Fashion and homophily*. Available at SSRN: http://ssrn.com/abstract=2250898 or http://dx.doi.org/10.2139/ssrn.2250898. 2013.

Chapter 3

An agent-based sorting model for city size and wealth distributions

Abstract

We propose a new model for city size and wealth distributions in an economy. Our model is, first, based upon agents with preferences over neighborhoods; rich/wealthy neigborhoods are more attractive than poorer ones and agents generally want to 'sort' into the neighborhood whose wealth level is largest. A second feature is that neighborhoods have an impact upon the members of their community, which we define in terms of the neighborhood's average wealth level. Finally, agents are inert in the sense that they are unwilling to leave their current neighborhood without reason and generally incur costs of relocation; and agents are boundedly rational in that they do not anticipate/predict other agents' behavior and in that they perform local instead of global relocation decisions. We derive a few analytical results for this setup which characterize our model as one where 'the poor are chasing the rich', amongst other things. Moreover, we show by simulation that, under reasonable parametrizations, our proposed model generates Zipfean city size distributions with coefficient alpha close to 1. It does not, however, by itself, generate Pareto wealth distributions. To this end, we add a stochastic component to individual agents' wealth levels, which we specify such that it either entails linear or exponential average growth. Nontrivially, this seems to lead to the 'correct shape' of the wealth distribution function for both the linear and exponential growth paradigms, but with 'suitable' coefficient beta only under the exponential growth implication.

3.1 Introduction

One version of *Zipf's law* for city sizes states that if one ranks cities by size and plots city size versus rank in log-log-scale, one obtains, roughly, a straight with slope $-\alpha = -1$, where we refer to α as the *Zipf coefficient*. In Figure 3.1, we illustrate the law, using city size data for the United States for the year 2009. One sees that the fit is not perfect, with some deviation particularly for the largest cities New York, Los Angeles, and Chicago, but overall seemingly pretty good and an excellent rule of thumb. Remarkably, a related law apparently holds for the distribution of wealth among subjects in economies. If one plots the probability that an individual has at least wealth level w against w in log-log-scale, one obtains, for the 'rich' tail of the distribution, again, a straight line, this time with slope between $-\beta = -1$ and $-\beta = -3$, where we call β the *Pareto coefficient*. This relationship, which we exemplify in Figure 3.2, is termed *Pareto's law* for wealth distributions. Both laws were discovered around the turn of the 19th century and, although *prima facie* having different interpretations, can both be phrased as 'rank size' rules; for Pareto's law, we have that if one ranks richest subjects by wealth and plots wealth level versus rank in log-log-scale, one obtains a straight line with slope between $-1/3$ and $-1/1 = -1$ (see below). The 'rank size' formulation allows a simple conceptualization; assuming a Zipf coefficient of 1 and a Pareto coefficient of 2, then, the largest city in an economy has about double the size of the

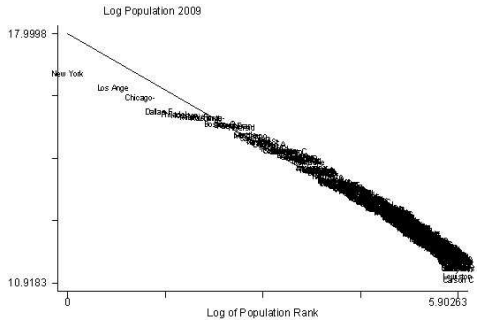


Figure 3.1: City size distribution of the United States for 2009, from `http://economix.blogs.nytimes.com/2010/04/20/a-tale-of-many-cities/`.

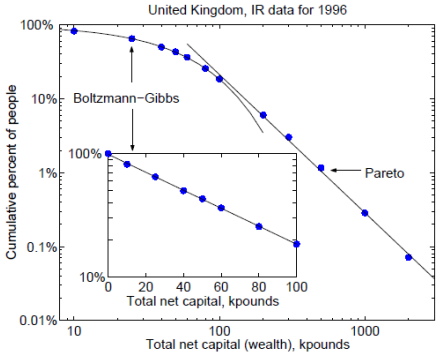


Figure 3.2: Wealth distribution of the United Kingdom for 1996, reprinted from Drăgulescu (2003).

second largest, three times the size of the third largest, etc. Accordingly, the wealthiest individual in an economy is about $\sqrt{2}$ times wealthier than the second wealthiest, $\sqrt{3}$ times wealthier than the third wealthiest, etc.

The two laws have received considerable attention ever since their discovery, particularly from physisicts, economists, and statisticians, where the leading question is about the mechanism(s) generating such outcomes. Two stochastic explanations for the Zipf and Pareto phenomena are as follows. For city sizes, Simon (1955) found that a 'preferential attachment' rule can imply the observed regularity; cities grow proportionally to their current size, that is, larger cities obtain more new inhabitants. For wealth, a distribution process in analogy to those observed in statistical physics has been suggested (cf. A. Chatterjee, Chakrabarti, and Manna, 2003) where at each time step two randomly selected individuals 'gamble' about an individually determined share of their current wealth, leaving total wealth unchanged. While these models may be at least partly convincing and undoubtedly appealing in their abstractness and simplicity, they do not tell us about instrinsic motivations and preferences of the agents involved in their setups. Some other models address this issue, for example, by defining *optimizing* economic agents that, in the case of cities, relocate, e.g., on the basis of preferences defined over population density; for example, agents may rejoice in companionship, but certainly want to avoid overcrowding.

In the current work, we undertake to propose a new model for explaining both city size and wealth distributions in an economy. Our model is, first, based upon agents with preferences over *neighborhoods*; rich/wealthy neigborhoods are more attractive than poorer ones and agents generally want to 'sort' into the neighborhood whose wealth level is largest.[1] A second, separate but related, feature is that

[1]The fact that human agents, or more generally, biological organisms, undertake to settle in rich ecological neighborhoods seems self-evident to us. We emphasize with a few examples. First, immigration from 'third-world' countries, such as Africa, to the industrialized nations, such as the United States, Europe, etc. Secondly, in the animal world, usually rivers, food-rich habitats, etc., exogenous sources of wealth (sometimes also called "locational fundamentals", cf. Krugman, 1996), attract multitudes of organisms. Thirdly, the 'poor chasing the rich' literature (cf. Tiebout, 1956; Strahilevitz, 2005; Bucovetsky

neighborhoods have an impact upon the members of their community, which we define in terms of the neighborhood's average wealth level; we take this circumstance as one justification of why agents would want to sort into richer neighborhoods. Moreover, such neighborhood or peer effects are well-known and well-documented in economics (cf. Dietz, 2002; Falk and Ichino, 2006, etc.) and, for example, in, e.g., businesses it is estimated that income and productivity of individual workers increase considerably as the average income and productivity of co-workers increase (cf. Ichino and Maggi, 2000; Shvydko, 2008). Finally, agents are *inert* in the sense that they are unwilling to leave their current neighborhood without reason and generally incur costs of relocation; and agents are *boundedly rational* in that they do not anticipate/predict other agents' behavior — they play best responses to the current 'state of affairs' without anticipating changes — and that they perform local instead of global relocation decisions.[2] We show by simulation that, under reasonable parametrizations, this model generates Zipfean city size distributions with coefficient α close to 1. It does not, however, by itself, generate Pareto wealth distributions (since, in the long run, our model tends to imply a too equal distribution of wealth among agents, as we also argue from analytical results we derive in Section 3.5, see below). To this end, we add a stochastic component to individual agents' wealth levels, which we specify in Section 3.6 such that it either entails linear or exponential average growth. Nontrivially, while not infringing upon the Zipfean city size distribution law, this seems to lead to the 'correct shape' of the wealth distribution function — straight line for the rich tail in log-log-scale, exponential regime for the poor tail, see below — for both the linear and exponential growth paradigms, but with 'suitable' coefficient β only under the exponential growth implication.

While we believe our model to be abstract and general enough to potentially apply to the migratory behavior of many living organisms, we note that it is, at the same time, rooted in economic theory. For example, the classical Tiebout sorting model (cf. Tiebout, 1956) holds that individuals sort into neighborhoods based upon the latters' attractiveness (in terms of taxes, public goods, etc.). Moreover, as indicated already, *neighborhood effects*, or, more specificially, *peer effects*, have been well-studied in economics and may include such phenomena as status formation motives, aversion to pay inequality,[3] learning, or exogenous effects due to environmental characteristics of an area (cf. Manski, 2000).

Our approach is, from an abstract perspective, closely related to a number of other modelings. Schelling (1978) posits, as do we, that micro motives — individuals' desire to live in wealthy neighborhoods, in our case, vs. individuals' desire to be close to agents of the same type, in Schelling's model — entail macro behavior — Zipfean city size distributions, in our case, vs. segregation, in Schelling's model. As in Page (1999) and Mansury and Gulyás (2007), two highly related models of 'city formation', we assume that attributes defined over neighborhoods, or, 'areas' of space, enter agents' utility functions and motives to relocate.[4] Novel about our approach is that we consider *wealth* — rather than a location's population density, as in the former two models — as an *endogenous* decision variable (e.g., rather than exogenous beauty of places, as in Rand et al., 2003). In considering wealth as a decision variable, our agents' utility functions — and hence, their motivations — may also be quite different from those in Page (1999) and Mansury and Gulyás (2007), where utility on a location's population density may be encoded in a quadratic function, implying, exogenously, both attractive and repulsive forces for agent relocation decisions. In constrast, we assume that utility on wealth is monotonically increasing in wealth, whence repulsion and attraction arise endogenously in our model, since the poor are attracted by the rich (as they have positive utility from more wealth), while the latter want to escape the former (for the same reasons regarding their utility functions). Contrary to the named three models, our model also incorporates (the economically well-founded concept of) *neighborhood effects*, so that agents' attributes — their wealth levels — are affected and modified over time. Our approach is, to the best of our knowledge, one of the few, general, agent-based approaches to modeling Zipfean city size distributions and the first model to discuss city size and wealth distributions in a unified framework.

This paper is structured as follows. In Section 3.2, we review the concept of power law distributions,

and Glazer, 2010) has as key insight that people care for the average income or wealth in the community in which they live for at least three reasons: status, peer groups, and taxes; see also our discussion below.

[2] Agents can only relocate within a pre-defined radius of their present location as in the model of Mansury and Gulyás (2007).

[3] Sometimes also referred to by the phrase 'Keeping up with the Joneses'.

[4] Our model, like those of Schelling (1978), Page (1999), and Mansury and Gulyás (2007), does not include economic variables — such as wages, land prices, etc. — except for wealth.

of which the Zipf city size and the Pareto wealth distributions are special cases, in more detail, and discuss empirical facts about city size and wealth distributions across countries. In Section 3.3, we more thoroughly outline related work, which includes economic models, in a narrow sense, as well as models based on stochastic random growth processes. In Section 3.4, we outline our model mathematically and discuss its individual components, whereafter we present a few intuitive analytical results about it in Section 3.5. Our first result is that our model implies a 'poor chasing the rich' scenario without (pure strategy Nash) equilibria, as we will render more precisely then, and our second result is a 'convergence in wealth' tendency inherent in our modeling due to the neighborhood effects. We also briefly consider the case when agents are slightly more rational (or, less 'myopic') than we have assumed in our modeling, and attempt to derive city size distributions implied by our framework for small values n of agents, which are the primary goal of this work. Because the mathematics to derive these for arbitrary values of n escapes us, we resort to simulation in Section 3.6, where we detail outcomes under a variety of different model calibrations. Finally, in Section 3.7, we conclude. In the apppendix, we derive one of our results, convergence of agents' wealth levels, under the assumption of agents optimizing in continuous time, showing that it agrees with the discrete time result.

3.2 Zipf's law, Power law distributions, and empirics

After briefly reviewing the concept of power law distributions, of which the Zipf and Pareto distributions are special cases, we address in more detail empirical distributions of *city sizes* and of *wealth*.

Zipf and power law distributions

Generally, a *power law distribution* is a probability distribution of the form

$$p(x) \sim x^{-\gamma}, \tag{3.2.1}$$

for a coefficient $\gamma \geq 1$. Power law distributions are ubiquitious as distribution functions of social, physical, biological, or technological systems. For example, according to Newman (2005), the cumulative density functions (cdfs) of the following quantities, amongst many others, are claimed to follow power law distributions: word frequencies, citations of scientific papers, web hits, copies of books sold in the United States, telephone calls received, magnitude of earthquakes, diameter of moon craters, intensity of solar flares, intensity of wars, wealth (of the richest individuals), frequencies of family names, and city sizes. Of course, each of these phenomena may potentially be described by different power law coefficients, and the power law distribution form may probably also only hold for some part of the range of variables. Moreover, some of the examples mentioned may be specific, for instance, to particular cultures (or simply, circumstances); e.g., power law distributions seem to hold for US American and Japanese family names (cf. Miyazima et al., 2000), but not so for Korean family names, which apparently follow an exponential distribution (cf. Kim and Park, 2005).

A *Zipf distribution* is a special kind of power law distribution with coefficient $\alpha := \gamma = 1$.[5] The Zipf distribution form was originally discovered by George Kingsley Zipf (1902-1950) as underlying word frequencies, but Felix Auerbach (1856-1933) is nowadays credited as the first to observe the (now-called) Zipf regularity in the distribution of city sizes. Both for word frequencies and city size distributions, a slightly different interpretation than the one given above, using cumulative distribution functions, prevails in the literature. Usually, *Zipf's law* is interpreted as a 'rank-size rule'; if words or cities are ranked in descending order according to their frequency or population size, respectively, one obtains the following relationship

$$\log \text{Rank} = \text{Constant} - \alpha \log \text{Size}, \tag{3.2.2}$$

where α is approximately 1. In log-log-space, thus, this is as a linear relationship. One remark is in place here; as in Figure 3.1, we find it more intuitive to regress a city's size on its rank, although (3.2.2) seems to be the prevailing regression equation in the economics literature and Zipf coefficients and their

[5] We will refer to the 'general' power law coefficient as γ, and call it α and β in the Zipf and Pareto case, respectively.

ranges are usually specified in terms of the above relationship. If one regresses size on rank, one obtains a coefficient that is, intuitively, the inverse of α above. We treat Pareto wealth distributions and their coefficients β, as special cases of power law distributions, below.

Empirical facts about city size and wealth distributions

Since the work of Auerbach (1913) and Zipf (1949), it has generally been accepted that *city size* distributions follow a power law with Zipf coefficient $\alpha = 1$. Many (simulational) studies to this date adhere to this 'Zipf benchmark' (cf. Mansury and Gulyás, 2007; Axtell and Florida, 2001, etc.). More recent research on the empirics of city sizes, however, indicates that the coefficient for city size distributions rather falls in an interval around $\alpha = 1$.[6] For example, Rosen and Resnick (1980) examine the distribution of city sizes for 44 countries in 1970; one of their main findings is that the average coefficient is 1.13, with all of the countries in their sample having an exponent between 0.8 and 1.5. Soo (2005) confirms most of these findings for his sample of 75 countries for the period 1970-2000, for which he estimates an average power law coefficient of 1.10; moreover, 71 out of his 75 countries have an exponent between 0.8 and 1.5. Two further remarks are in order here. First, as Krugman (1996) and Brakman, Garretsen, Merrewijk, et al. (1999) point out, Zipf's law for city sizes holds best when very small cities are excluded, and, in the United States, is most accurate for cities between 200,000 and 20,000,000 inhabitants.[7] Next, the result of a Zipf distribution for city sizes may to some degree also depend on the *definition* of a city; e.g., the postulation of what *is* a city may partly be arbitrary and/or underlie historical contingencies, cf. Gabaix and Ioannides (2004), Newman (2005).

Vilfredo Pareto (1848-1923) is credited as the first to quantitatively investigate the distribution of *wealth* in a society. In Pareto's original work (Pareto, 1896), he discovered that this distribution follows a power law for large values m of wealth, that is,

$$\Pi(m) \sim m^{-\beta}, \quad \text{for } m \text{ large enough,}$$

where $\Pi(m)$ is the distribution function giving the probability that an individual's wealth is at least m (cf. Richmond et al., 2006). Pareto (1896) finds power law exponents β between 1.24 (for Basle 1887) and 1.89 (for Prussia 1852), while recent published empirical data estimates coefficients β between just under 1 to almost 3, with an average exponent of around 2 (cf. Santos et al., 2007; Coelho et al., 2008). Next, as mentioned, the power law distribution of wealth seemingly holds only for the richest subjects of a society. Santos et al. (2007), Drăgulescu (2003), and others, note that the remaining range of wealth distributions may follow other distributional laws, such as the (Boltzmann-Gibbs type) exponential or log-normal distribution (cf. Figure 3.2). A problem arises here because the definition of 'the richest of a society' is not specified but is required to quantitatively determine β from data; it is, however, common practice to regard the top 10%, 5%, 3%, or 1% wealthiest of a society (cf. Cowell, 2011) as the 'richest'. Finally, it must be mentioned that, except for the Forbes magazine rankings, which are based on wealth (fortune), most econometric studies actually use *income* as a proxy for wealth, as most of the available data about personal richness comes from individual income tax declarations (cf. Santos et al., 2007).

3.3 Related literature

A multitude of different models have been proposed to explain Zipf's law for city sizes, on the one hand, and wealth distributions in an economy, on the other. These models often stem from very different scientific fields, whereby a large fraction of approaches is based on random stochastic processes. Newman (2005) summarizes some of the best known and widely applied stochastic mechanisms that have been proposed to generate power laws. Among them is the Yule process, which was originally introduced in

[6] To make things worse, Benguigui and Blumenfeld-Lieberthal (2007) observe that, for several countries, it seems that a non-linear function in log-log space is a better fit to city size distributions than a linear function.

[7] This point is also made by Rossi-Hansberg and Wright (2007), who observe that, in contemporary city size data, small cities are under-represented and big cities are too small, compared with the Zipf benchmark. See also Dittmar (2010), who in addition makes the case that Zipf's law in Europe has only established since the 1500s, because, before, "land and land-intensive intermediates entered city production as quasi-fixed factors", slowing down growth of big cities. In contrast, Davis and Weinstein (2002) find that Zipf-like laws hold in Japan for time periods stretching back thousands of years.

Yule (1925) to explain the distribution of biological taxa. It has later been generalized and adapted, e.g., by economist Herbert Simon to explain the distribution of city sizes; cf. Simon (1955) and, for a more recent contribution, Gabaix (1999). Its principle mechanism is that of *preferential attachment* — cf. Barabási and Albert (1999), also called *Gibrat's law* in Simon (1955). This principle means that objects, such as cities, receive additional entities, such as people, in proportion to the number of entities they already have, that is, 'the rich get richer'. A major flaw of the standard Yule process is that it ignores that objects (and entities) may become extinct but, among the stochastic explanations for, e.g., city size distributions, it has become the most widely accepted theoretical model. Problematic about the Yule process is moreover that objects may never relocate once they have attached to a particular location (cf. Mansury and Gulyás, 2007). Combination of exponentials, whereby a quantity of interest — such as city sizes, or more generally, sizes of biological populations — is exponentially related to an exponentially distributed random variable — such as time of death of the organisms constituting the populations — may also entail power law distributions, cf. Reed and Hughes (2002).

Concerning wealth distribution models, a few important models have been developed within the field of so-called *econophysics*, many of which are analogous to models of collisions between molecules as considered in statistical physics. Here, wealth is assumed to be exchanged between two randomly selected economic agents like the exchange of energy between two molecules, including the law of conservation of energy; what one agent wins, the other loses. In pure gambling games (cf. Gupta, 2006), the sum $w(t)$ of the wealths of two agents i and j at time t is at disposal and a random draw $\epsilon \in [0, 1]$ determines the share of $w(t)$ that both agents have in the next period,

$$w_i(t+1) = \epsilon[w_i(t) + w_j(t)],$$
$$w_j(t+1) = (1-\epsilon)[w_i(t) + w_j(t)].$$

Such forms of interactions between agents lead to Boltzmann-Gibbs type exponential distributions in individual wealth but variations thereof, incorporating savings of agents, may lead to other distributions, such as the power law distribution, cf. the models of A. Chatterjee, Chakrabarti, and Manna (2003), Slanina (2004), etc. Several network models of wealth distribution that successfully reproduce the wealth power law coefficients have also been proposed, i.e., models in which agents live on networks and exchange or distribute wealth, cf. Dorogovtsev and Mendes (2003), Coelho et al. (2008), Santos et al. (2007). A more complete overview is, for instance, provided in Chen (2011).

Of course, problematic about all the above models, from a microeconomic viewpoint, is that — while they are certainly elegant and appealing in their generality and abstractness — the models do not deduce their results from individual agents' preferences over outcomes, as is one of the basic tenets of modern microeconomic theory. Other models rest more thoroughly on economic principles. Axtell and Florida (2001), for example, hold that city and firm sizes are both Zipf distributed and inherently correlated. In fact, they claim that in pre-industrial times, city size distributions were less skew than they are now and that there "are deep and important connections between firm and city size distributions". Their model rests upon the interaction of workers and firms, where a city is in their model defined as an agglomeration of firms. Workers form firms, and firms select locations, but workers may also change between firms when they find it welfare-improving. Economically, workers are driven by motives of providing effort for team production and may join another firm if this increases their utility; decisions of individual workers to change firms may have influence on other team members. The authors show that the city size distributions obtained from their experiments are centered around 'true' Zipf distributions, defined by a coefficient of $\alpha = 1$. A few interesting macroeconomically founded models, aimed at generating Zipf city size distributions, have also been proposed. Duranton (2002)'s modeling approach to city size distributions rests on Grossman and Helpman (1991)'s quality-ladder model of growth, which assumes innovation driven technology shocks as sources of city and industry formation and development. Cities grow or decline as they win or lose industries following new innovations. So small innovation-driven technology shocks are the main engine behind the growth and decline of cities. The model matches both the empirical distributions of US and French city sizes, as opposed to an abstract 'Zipf benchmark' of 1, which themselves are considerably dissimilar, when key parameters are calibrated accordingly. The new economic geography model of Brakman, Garretsen, Merrewijk, et al. (1999), which extends Krugman (1992)'s general equilibrium location model by the introduction of external

diseconomies, congestion costs, etc., and rests upon Brakman, Garretsen, Gigengack, et al. (1996), requires an "industrialization" setting in order to generate Zipf coefficients near 1. Pre- and post-industrialization scenarios are associated with coefficients larger than 1; see also Gabaix and Ioannides (2004) and Mansury and Gulyás (2007) for more extensive overviews and additional references.

Equilibrium-based wealth distribution models are offered, among many others, by Wang (2007), who applies a stochastic consumption rule which captures precautionary savings motives to a self-insurance model with inter-temporally dependent preferences, Quadrini (2000), and S. Chatterjee (1994), who investigates wealth distribution in a neoclassical growth model, for example. All these papers do not try, however, to generate power law distributions for wealth distributions, but rather focus on 'secondary' characteristics of wealth distributions, such as skewness, wealth concentration, inequality, etc. Fiaschi and Marsili (2009) study the equilibrium distribution of wealth in a macroeconomic model with firms, households, and government taxes and find a Paretian law in the top tail of the wealth distribution function in case of incomplete markets. Benhabib, Bisin, and Zhu (2011) also find a Pareto distribution in the rich tail of the wealth distribution function in their overlapping generations model; see also Chen (2011), and others.

An issue with the afore-mentioned economic models, from our perspective, is that many of them rely on quite demanding assumptions regarding the structuring of society, such as the existence of firms, workers, industries, 'technology shocks', 'industrialization', etc. More parsimonious economic frameworks are discussed, for instance, by Krugman (1996) — who focuses on the tension between attraction and repulsion (in his terminology, centripetal and centrifugal forces) as sources of city formation — Schelling (1978) and Page (1999). In Schelling (1978)'s model — whose general motivation is to derive 'macro behavior' from 'micro motives', which Schelling (1978) considers a general structuring principle — two types of agents (black vs. white, male vs. female, etc.) live on a two-dimensional grid. Each type requires a minimum number of agents of the same type in his neighborhood[8] — and when this threshold is not reached, the agent randomly relocates to a new grid place. One of the model's surprising results is that segregation is likely to emerge even when agents are tolerant toward the other type. Page (1999) models the emergence of cities by assuming preferences of agents, distributed on a two-dimensional grid as in Schelling (1978)'s model, over a location's population and its separation (that is, its average distance to other agents) and shows that 'cities' form under such preferences; similarly as we do, he derives both analytical as well as simulational results for his modeling. Neither Krugman's, Schelling's, nor Page's approach compare resulting city size distributions, however, to a 'Zipf' benchmark.[9]

Mansury and Gulyás (2007)'s model, in contrast, aims at explicitly deriving Zipf city size distributions. In their modeling, population density is, as in Page (1999), the decision variable for agents' migratory behavior; more precisely, quadratic preferences of agents over the population size at any given grid point are assumed such that agents rejoice in companionship but want to avoid overcrowding. They derive a Zipf distribution for city sizes under certain restrictions on agents' spatial reach. In Rand et al. (2003), agents have preferences over a location's natural (and exogenous) beauty and its distance to 'service centers'; under their parametizations, the authors note a power relationship between frequency and cluster size, that is, a Zipf-like relationship.

A non-stochastic agent-based wealth distribution model is, for example, Wilensky (1998)'s netlogo model, adapted from Epstein and Axtell (1996)'s sugarscape model. In this model, agents accumulate grain, their wealth, on a two-dimensional grid, while attempting to maximize their wealth. Agents have life-expectancy and produce off-spring on death. This model typically produces power law wealth distributions.

3.4 Model

We first succintly present our model and its basic terminology in an abstract manner, whereafter we elaborate on the importance of each of its aspects, thereby illustrating more concretely their possible realizations. We deliberately keep the model general, first, before discussing a concrete version of it in the results section, Section 3.6.

[8] This model has some similarity with Conway's 'game of life' (cf. Gardner, 1970) model.

[9] Of course, Schelling (1978) does not even interpret his objects as 'cities'.

As mentioned, our model assumes that agents optimize their *wealth* structure, whereas other models attempting to reproduce the emergence of cities (or similar entities), or, more particularly, city size distributions, rely on other decision variables, such as *density* (cf. Mansury and Gulyás, 2007; Page, 1999) or *homophily* (cf. Schelling, 1978). In Section 3.4, we briefly discuss how these quantities could be incorporated in our model, although we do not include them in the current exposition, for reasons of parsimony.

Setup

A set of *agents* (or *players*) $[n] = \{1, 2, \dots, n\}$, for $n \geq 2$, inhabits a *world* — which we also refer to as *grid* or *lattice* — $X \subseteq \mathbb{N}^k$, for $k \geq 1$. Only a fraction $s \in [0, 1]$ of all places (or points) $p \in X$ are inhabited, the remaining are empty — we remark that each position $p \in X$ may either be empty or is otherwise occupied by a *single* agent. Each agent i has *payoff*, which we refer to as his *wealth*, $Y_{i,t}$ in *periods* $t = 0, 1, \dots, T$. Payoffs are determined by the agent's current payoff and his environment's payoffs according to

$$Y_{i,t+1} = Y_{i,t} + \delta(\bar{Y}_{p_i,t} - Y_{i,t}) + \epsilon_{i,t+1} = (1 - \delta)Y_{i,t} + \delta\bar{Y}_{p_i,t} + \epsilon_{i,t+1}, \tag{3.4.1}$$

where $\delta \in (0, 1)$ is the *adaption rate*, $\bar{Y}_{p_i,t}$ is some 'average' payoff at agent i's location $p_i \in X$ at time t, and $\epsilon_{i,t+1}$ is a random component. Generally, the *average payoff* $\bar{Y}_{p_i,t}$ at position p_i, which summarizes 'average' wealth 'in the neighborhood' of p_i, is determined according to

$$\bar{Y}_{p_i,t} = \sum_{j \in [n]} w_{p_i,p_j} Y_{j,t}, \tag{3.4.2}$$

where $w_{p_i,p_j} \geq 0$ and $\sum_{j \in [n]} w_{p_i,p_j} = 1$. We interpret the *weight* w_{p_i,p_j} as the influence of an agent j, at position p_j, on the average wealth of location p_i, at which agent i resides. We generally assume, as indicated in our notation, that these weights depend on the (relative) positions of agents. Figures 3.3 and 3.4 schematically illustrate our model's basic setup.

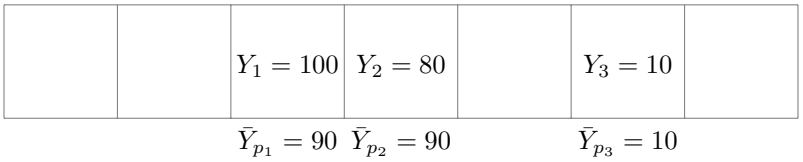


Figure 3.3: World X as a finite one-dimensional grid. For each position $p \in X$, we indicate whether it is occupied or not. We also indicate average wealth levels $\bar{Y}_{p,t}$ for the occupied places. Here, $\bar{Y}_{p,t}$ is determined by uniformly averaging neighbors' wealth levels within one unit of distance to the current place p.

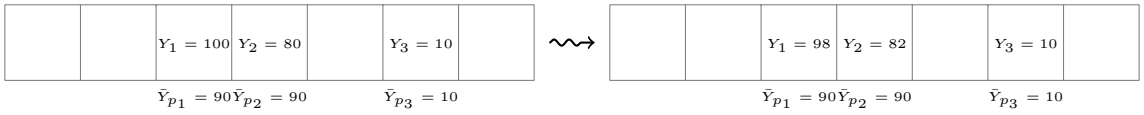


Figure 3.4: Same setup as in Figure 3.3. Once all agents have made their relocation decisions, their wealth levels are updated via Equation (3.4.1). Here, we let $\delta = 0.2$.

Relocation Dynamics

All agents receive random initial wealth $Y_{i,0}$, for $i = 1, \dots, n$. Then, for all periods $t = 1, \dots, T$, they solve the following maximization problems

$$\max_{p \in X} \mathbb{E}_t[u_Y(Y_{i,t+1})], \tag{3.4.3}$$

where $u_Y : \mathbb{R} \to \mathbb{R}$ is a *utility function* on wealth and $\mathbb{E}_t[\cdot]$ denotes the (conditional) expectation operator, conditional on information available at time t. In other words, at time t, agents search for the positions p on grid X that promises highest expected utility in period $t+1$. Crucially, we assume that agents have limited horizon and solve maximization problem (3.4.3) by playing a myopic best response to the world as it is at time t and, more precisely, at the time point of their relocation decision:[10] of all currently free positions in world X, agent i chooses the one with the *current* highest utility, which would, in fact, be a utility maximizer for period $(t+1)$ wealth of agent i provided that, subsequent to agent i's decision, all other agents remained at their position prior to i's relocation decision.[11] If u_Y is non-decreasing ("more wealth cannot be worse"), this strategy has an obvious solution since $Y_{i,t+1}$ has only one component which is not determined exogenously at time t, namely $\bar{Y}_{p_i,t}$.[12] Thus, in period t, once it is her turn to relocate, agent i chooses position $p \in X \backslash \mathcal{O}$ — where $\mathcal{O}$ is the set of occupied places in X — with the current highest *average wealth level* $\bar{Y}_{p,t}$. Since $\bar{Y}_{p,t}$ summarizes average wealth 'in the neighborhood of p', we interpret agent i's choice as a choice for the 'wealthiest' neighborhood. Importantly, the order in which agents are allowed to choose their positions in period t is determined randomly; agents who move later make more 'informed' decisions.[13] Wealth levels are adapted, via rule (3.4.1), only *after* all n agents have made their relocation decisions. Figure 3.5 schematically illustrates agent 3's situation when deciding to relocate.

		$Y_1 = 100$	$Y_2 = 80$		$Y_3 = 10$	
$\bar{Y}_p = 10$	$\bar{Y}_p = 55$	$\emptyset$	$\emptyset$	$\bar{Y}_p = 45$	$\bar{Y}_p = 10$	$\bar{Y}_p = 10$

Figure 3.5: It is agent 3's turn to make a relocation decision. The position p immediately to the left of agent 1 has currently highest average wealth (under the assumption that agent 3 would move to this position) — $\frac{100+10}{2} = 55$ — so agent 3 myopically maximizes her expected utility if she relocates from her current position to p. 'Unreachable' positions, i.e., because they are occupied, are indicated by $\emptyset$.

To avoid 'inflationary' movements and to account for individuals' 'inertia', we introduce positive moving costs

$$c : X^2 \to \mathbb{R}_{\geq 0}, \qquad c : (p, q) \mapsto c(p, q).$$

This changes the optimization conditions only slightly; agents now solve the utility maximization problems

$$\max_{p \in X} \mathbb{E}_t[u_Y(Y_{i,t+1}) - c(p_i, p)] \tag{3.4.4}$$

where p_i is agent i's position prior to her relocation decision. Again, agents solve this problem in a myopic best response manner. Moreover, as an additional aspect of 'bounded rationality', we generally restrict agents to conduct a *local* search for optimal grid positions instead of a global search; i.e., they may be restricted to choose positions in the vicinity of their current habitat,[14] which can be modeled by letting moving costs be infinite for distant places.

As we have mentioned, when all n agents have moved — some may have remained at the position they were occupying before — their wealth levels are updated via Equation (3.4.1). We summarize the relocation dynamics in Algorithm 1: first, agents receive random initial wealth levels (line 3), and then,

[10] Our model shares this assumption with that of Page (1999) or Mansury and Gulyás (2007).

[11] Agent i assumes that the world stays as it is except for her own hypothetical movement. Otherwise, if this was not considered, agents would avoid 'empty regions' as these have low average income values.

[12] The variable $\bar{Y}_{p_i,t}$ is not exogenously determined at time t since the weights w_{p_i,p_j} depend on the position of agent i relative to agent j.

[13] The assumption of random movements is the same as in Page (1999). He also, in a footnote, discusses incentive based asynchronous updating, whereby those that gain most from relocating are allowed to move first, which may significantly alter the relocation dynamics.

[14] This also reduces computational burden, in the simulations.

for a total of T periods, relocate by playing best responses as described (lines 5-7), whereafter their wealth levels are updated (lines 8-10).

Algorithm 1 Relocation Dynamics (RD)

1: **procedure** RD(T, n) ▷ T is the number of periods, n the number of players
2: $t \leftarrow 0$
3: $Y_{i,t} \leftarrow Z_{i,t}$ where $Z_{i,t}$ is an iid random draw from a distribution with cdf F_Z, for $i = 1, \ldots, n$
4: **while** $t < T$ **do**
5: **for** $i \in [n]$ in random order **do**
6: Myopically solve optimization problem (3.4.4) by choosing $p \in X \backslash \mathcal{O}$ that maximizes $\bar{Y}_{p,t} - \frac{c(p_i,p)}{\delta}$
7: **end for**
8: **for** $i \in [n]$ **do**
9: Update $Y_{i,t}$ via Equation (3.4.1) to obtain $Y_{i,t+1}$
10: **end for**
11: $t \leftarrow t + 1$
12: **end while**
13: **end procedure**

Discussion of the model

To specify the agents' world $X \subseteq \mathbb{N}^k$, in the agent-based models that we have reviewed, usually a one or two-dimensional grid is assumed. For example, we might specify that $k = 1$ and $X = \{1, \ldots, m\}$ or that $k = 2$ and, e.g., $X = \{1, \ldots, m\} \times \{1, \ldots, l\}$,[15] for positive integers m and l, both of which would result in a *finite* world, that is, with finitely many places. In our simulations in Section 3.6, we specify X as one-dimensional because this is the simplest specification and also reduces the difficulty of defining a city, which we generally identify as a subset C of X of *contiguous occupied grid points*, whereby the exact specification of 'contiguity' may be problematic, however, unless the grid is one-dimensional, in which case a contiguous set of occupied grid points is obviously any connected array of points, all of which are occupied. In Figures 3.3 and 3.4, hence, there would be two cities — one constisting of agents 1 and 2, on the one hand, and one consisting of agent 3, on the other — if we follow our just mentioned definition of a city.

As to the agents' interactions, Equation (3.4.1) models the neighborhood effects discussed above. Note that, disregarding the error, $\delta \to 1$ implies that $Y_{i,t+1} \to \bar{Y}_{p_i,t}$ and $\delta \to 0$ implies $Y_{i,t+1} \to Y_{p_i,t}$, so that the adaption rate determines how strongly an agent's wealth is affected by average wealth at i's current position p_i and his last period wealth, respectively. Note also that Equation (3.4.1) has appeared in several contexts in the economics and non-economics literature. For example, assuming the random components to be zero for the moment, the equation in the form $Y_{i,t+1} = (1 - \delta)Y_{i,t} + \delta\bar{Y}_{p_i,t}$ illustrates that an agent's next period wealth appears as a convex combination of current period (own) wealth and current average ('other' agents') wealth at the given location. Perceived thus, the model has, i.a., strong resemblance with Falk and Ichino (2006)'s model of worker productivity interdependence.[16] Intruigingly, the wealth update equation is also a special case of Friedkin and Johnsen (1990)'s social influence model.[17] The specification is, moreover, a variant of the updating rule in self-organzing maps (cf. Kohonen, 1984), neural networks (cf. Hopfield, 1982), etc., and thus related to the literature on

[15] In our model, as in Page (1999), we do not 'connect' the end points of X, which would result in the topology of a torus or a ring for X. As Page (1999) points out, although the earth is round, most countries are, topologically, much more similar to a rectangle, i.e., with bounds on each side.

[16] They consider just two agents, whereas we consider more generally $n \geq 2$ agents. They also attach potentially non-convex weights to own productivities and other agents' productivities.

[17] And, thus, the model is also related to the literature on *opinion dynamics*. In fact, wealth update equation (3.4.1) is a generalization of the belief updating equation specified in DeMarzo, Vayanos, and Zwiebel (2003). A distinction here, however, is that, in the opinion dynamics models, opinion updating via weighted averages is usually consider a (boundedly) rational behavior of agents, while Equation (3.4.1) does not describe a *choice* but non-deliberate influence. In fact, a rational agents' choice would be to increase her wealth indefinitely, provided that she has strictly increasing utility on wealth, rather than to mix wealth levels with peers'.

unsupervised self-organizing adaptive systems, etc. Its general role in this context is to increase the similarity between two objects, usually represented as vectors in Euclidean space.

The random components in Equation (3.4.1) need not necessarily be realized as white noise, i.e., as a sequence of zero-mean, independent random variables.[18] On the contrary, as discussed in Section 3.1, we might want to design our model in such a way that the *rich are getting richer*. To do so, the conditional expectation $\mathbb{E}[\epsilon_{i,t+1} \,|\, Y_{i,t}]$ could be specified as an increasing function of $Y_{i,t}$. A simple choice, leading to exponential growth and implementing 'preferential attachment', would be to let $\mathbb{E}[\epsilon_{i,t+1} \,|\, Y_{i,t}] = \mu Y_{i,t}$ for $\mu > 0$. Note that this would introduce the following interdependencies between the ϵ and Y variables.

$$\begin{array}{ccccccccc} \dots & & Y_{i,t-1} & \to & Y_{i,t} & \to & Y_{i,t+1} & \to & \dots \\ & & \uparrow & \searrow & \uparrow & \searrow & \uparrow & \searrow & \\ \dots & & \epsilon_{i,t-1} & & \epsilon_{i,t} & & \epsilon_{i,t+1} & & \dots \end{array}$$

In Section 3.6, we consider both a linear and an exponential growth paradigm of $Y_{i,t}$, as determined by the choice of $\epsilon_{i,t+1}$.

Next, concerning the determination of average wealth, Equation (3.4.2), a particular instantiation of the weighting scheme w_{p_i,p_j} would be to set the weights uniformly within a fixed radius $r > 0$ of agent i's position $p_i \in X$,

$$w_{p_i,p_j} = \begin{cases} \frac{1}{|B_r(p_i) \cap \mathcal{O}|} & \text{if } p_j \in B_r(p_i), \\ 0 & \text{else,} \end{cases} \tag{3.4.5}$$

where $B_r(p_i) = \{x \in X \,|\, d(x, p_i) < r\}$ is the open ball with radius r around p_i, and d is a metric on X, such as Euclidean distance; as before, $\mathcal{O}$ is the set of occupied places in X. In Section 3.5, we consider the uniform weighting scheme in the analytical results we derive, because it is a very convenient and simple choice of weighting scheme, but generally ignores the fact that influence may decrease in distance, even within a predefined neighborhood. Another possibility, accounting for the latter issue, is to use the density of the multivariate normal distribution centered at p_i as weighting factor,[19]

$$w_{p_i,p_j} = \frac{1}{C} \exp\left(-\frac{1}{2} (p_j - p_i)^\intercal \boldsymbol{\Sigma}^{-1} (p_j - p_i) \right), \tag{3.4.6}$$

where C is a normalization constant such that $\sum_{j \in [n]} w_{p_i,p_j} = 1$. In the experiments in Section 3.6, we truncate the normal distribution so that $w_{p_i,p_j} = 0$ outside a predefined interval.

For the moving costs function $c : X^2 \to \mathbb{R}_{\geq 0}$, we assume that c is symmetric, $c(p, q) = c(q, p)$, and that $c(p, p) = 0$. Moreover, we assume that c is non-decreasing in the distance between two points $p, q \in X$. A simple choice we make use of in the simulations is

$$c(p, q) = \chi \cdot \|p - q\|, \tag{3.4.7}$$

where $\|\cdot\|$ is the Euclidean distance (absolute distance, in the one-dimensional case) and $\chi \in (0, 1)$ is a moving cost parameter.

Additional variables

Of course, wealth need not be the only decision variable in our model. We generally assume that factors defined over neighborhoods of locations influence an agent's decision to relocate but these factors may also include, e.g., the *population density* D of a position p_i[20] — as has been outlined as a relevant criterion in related work — or *inequality* E, e.g., defined as the absolute difference between $Y_{i,t}$ and $\bar{Y}_{p_i,t}$, $\left\|Y_{i,t} - \bar{Y}_{p_i,t}\right\|$. Agents utilities u_D and u_E on these variables D and E could be such that u_D is quadratic, e.g., $u_D(x) = x - x^2$ (that is, low if there are too few or too many agents around), where

[18] As we show in Section 3.5, such a choice would not result in Pareto wealth distributions.

[19] The multivariate normal distribution becomes univariate when X is unidimensional, as we consider in Section 3.6.

[20] E.g., defined as the number of agents located around p_i or, better, as a normalized value in $[0, 1]$.

$x \in [0,1]$, and u_E is non-increasing. Then, total utility of agent i could be an additive linear utility function $u : \mathbb{R} \times [0,1] \times \mathbb{R} \to \mathbb{R}$, separable in Y, D, and E, e.g.,

$$u(Y, D, E) = \alpha u_Y(Y) + \beta u_D(D) + \gamma u_E(E), \tag{3.4.8}$$

with some coefficients α, β, γ. Investigating such an extended model, e.g., by simulation, would very likely be insightful, but we do not undertake it in the current work. Moreover, introducing additional variables renders our model more complex, and, by Occam's razor,[21] scientific models should try to reproduce their objectives — in our case, city size and wealth distributions — by the most parsimonious approach that can explain them.

3.5 Analytical results

All throughout this section, we focus on the following particular setup, unless explicitly stated otherwise. Random shocks to wealth $\epsilon_{i,t}$ are zero for all i and t. Moreover, assume that $c(p,q) = 0$ for all $(p,q) \in X^2$, i.e., moving costs are also zero such that relocating is for free. Assume, moreover, that the average income $\bar{Y}_{p_i,t}$ is determined by uniformly averaging all agents' wealth levels within a radius $r > 0$ of agent i's position p_i, for $i = 1, \ldots, n$, that is, weights are computed according to Equation (3.4.5). For our first result, below, recall that a *Nash equilibrium* is a state in a game where no player has a (unilateral) profitable deviation.

Instability and convergence of wealth levels

We now consider the following game. Assume that, rather than choosing grid places sequentially, agents choose them simulatenously. Fix some time period $t \geq 0$ and wealth levels $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})$. Then, once each agent i has chosen her position $p_i \in X$, agents receive payoffs $u_Y(Y_{i,t+1})$, where, as before, $Y_{i,t+1} = Y_{i,t} + \delta(\bar{Y}_{p_i,t} - Y_{i,t})$. We must resolve a technical issue here because, in the simultaneous move game, it might happen that two distinct agents i and j choose the same positions $p_i = p_j$, which is disallowed in our model. Accordingly, we simply call the corresponding strategy profile $\mathbf{p}$ *invalid*, and otherwise we call $\mathbf{p}$ *valid*. Thus, for a choice of valid positions $\mathbf{p} = (p_1, \ldots, p_n) \in X^n$, we let agent i's utility be

$$U_{i,Y}(\mathbf{p}; \mathbf{Y}_t) = u_Y(Y_{i,t+1})$$

and we let $U_{i,Y}(\mathbf{p}; \mathbf{Y}_t)$ be undefined if $\mathbf{p}$ is invalid. We now consider the valid Nash equilibria of the game $([n], X^n, U_Y(\cdot; \mathbf{Y}_t))$, where $U_Y(\cdot; \mathbf{Y}_t)$ is the vector $(U_{1,Y}(\cdot; \mathbf{Y}_t), \ldots, U_{n,Y}(\cdot; \mathbf{Y}_t))$, showing that this game has, in fact, no such pure strategy Nash equilibria.

Proposition 3.5.1 ("Instability")**.** Let $t \geq 0$ and wealth levels $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})$ be fixed. Consider the normal form game $([n], X^n, U_Y(\cdot; \mathbf{Y}_t))$. If u_Y is strictly increasing in Y for all agents and if X and $1 - s$ are 'sufficiently large', r is not 'too big' and weights are computed according to (3.4.5), then, unless $Y_{1,t} = \ldots = Y_{n,t}$, there are no pure strategy Nash equilibria in the normal form game as outlined.

Proof. If $Y_{1,t} = \ldots = Y_{n,t} \equiv Y$, then, for all valid $\mathbf{p}$ and individual choices p_i in $\mathbf{p}$, $\bar{Y}_{p_i,t} = Y$. Hence, $Y_{i,t+1} = Y_{i,t}$ for all agents $i \in [n]$ and by changing to a position p_i', agent i could not improve her payoff. Hence, no player has a unilateral profitable deviation. Consequently, all valid $\mathbf{p}$ are Nash equilibria in this situation.

Now, assume that it is not true that $Y_{1,t} = \cdots = Y_{n,t}$. Without loss of generality, assume that there is a unique richest agent with wealth Y_t and that there is a second richest agent with wealth y_t, $0 < y_t < Y_t$.[22] If the agent with wealth Y_t — call him Y_t, for short — is within another agent's radius — that is, agent Y_t's position is within distance r to another agent's position — Y_t has a profitable

[21] *Occam's razor* is a principle of parsimony, economy, or succinctness attributed to the English philosopher William of Occam (1287–1347).

[22] If there are multiple richest agents, the proof follows along the same lines as the one outlined. The difference is then that we have to distinguish whether the group of richest agents is isolated (as a group) or not.

deviation, namely, to move out of this radius to a place where he is not within any other agent's radius (by assumption such a place exists, since X is sufficiently large, r is not 'too big' and there are enough free places, i.e., $1 - s$ is sufficiently large). This is so because at Y_t's current position p, it holds that $\bar{Y}_{p,t} < Y_t$ since (in the equation, C is a normalizing constant)

$$\bar{Y}_{p,t} = \sum_{p_j \in B_r(p)} \frac{1}{C} Y_{j,t} < Y_t \sum_{p_j \in B_r(p)} \frac{1}{C} = Y_t,$$

and hence, by Equation (3.4.1), agent Y_t's wealth would be smaller in period $t+1$ than in period t if he stayed where is his. Only moving to an isolated place can prevent this.

Conversely, if Y_t is not within any other agent's radius, agent y_t has a profitable deviation. If she stayed where she currently is, her next period payoff would be at most y_t since she is the second richest agent and the richest is not within her radius. Moving within Y_t's radius would be a profitable deviation as then $\bar{y} = \frac{y_t + Y_t}{2} > y_t$ and hence, for all positive δ, we have, by Equation (3.4.1),

$$y_{t+1} = y_t + \delta(\underbrace{\frac{y_t + Y_t}{2} - y_t}_{>0}) > y_t.$$

□

Remark 3.5.1. It is important to note that in the second case discussed above — Y_t is not within any other agent's radius — not all 'poor' agents have a profitable deviation by moving within Y_t's radius. To illustrate, suppose that there are 5 agents with wealth levels $1, 4, 4, 4, 5$. If agents $1, 4, 4, 4$ form a group and 5 is isolated, 1 does not want to move within 5's radius, as $\frac{1+4+4+4}{4} = \frac{13}{4} > 3 = \frac{1+5}{2}$. In other words, the average wealth is higher at 1's current position, with agents whose wealth levels are $4, 4, 4$ close by, than it is at a position within 5's radius, for an agent with wealth level 1.

Remark 3.5.2. We note that Proposition 3.5.1 is equivalent to the following statement. Let π be any permutation on the set of agents $[n]$ that describes movement orders, such that $\pi(i) = \tau$ says that agent i is the τ-th agent, for $1 \leq \tau \leq n$, to make his relocation decision. When π is fixed, our relocating model is a deterministic process that can fully be determined via considering the myopic best responses played by agents to the current state of affairs. Let $\mathbf{p}_t \in X^n$ be a valid profile of positions at time t, that is, $\mathbf{p}_t$ describes agents' locations at time t, before relocation decisions are implemented. Then, let P_π be the deterministic operator that maps $\mathbf{p}_t$ to $\mathbf{p}_{t+1}$, the profile of positions *after* all n agents have conducted their relocation decisions, for a given fixed movement order, as encoded in π. Then, as indicated, Proposition 3.5.1 is equivalent to the statement that $P_\pi(\mathbf{p}_t) \neq \mathbf{p}_t$ for any valid profile $\mathbf{p}_t$ and any permutation π (unless all agents have the same wealth levels at time t). In other words, the condition of no pure strategy Nash equilibrium precisely means that the operator P_π has no fixed points, no matter the permutation π. This leads to our next remark.

Remark 3.5.3. By our previous remark, Proposition 3.5.1 does not merely describe a theoretical result — non-existence of pure one-shot valid Nash equilibria in the simultaneous move game — but the proposition implies actual consequences for the dynamics of the sequential move 'game'. Namely, at the beginning of each new time period t, if it is not the case that all agents have the same wealth levels, then, by the proposition, we know that at least one agent is 'unhappy' with the current placement of agents on the grid. If this agent were first to move, she would surely immediately switch to another place. If she is not the first to choose a new position, then, if no one before her relocates, she will do so once it is her turn. Hence, the proposition implies that, in each new time period, if not all agents have the same wealth levels, then *there will always be movement and relocations* — in other words, instability — among the agents; which is precisely what the condition $\mathbf{p}_{t+1} = P_\pi(\mathbf{p}_t) \neq \mathbf{p}_t$ in the last remark states. Moreover, Proposition 3.5.1 states that these movements are driven by wealth inequalities, and that, more precisely, the rich want to escape from the poor in our model, the latter which are chasing the former. Thus, the proposition furthers insight into the dynamics of actual agent relocations, in each time period, of the model outlined in Section 3.4.

We note, however, that the proposition might be false in case positive moving costs are introduced, for example, which might make certain relocations unprofitable.

Rather than focussing on relocation dynamics, we now investigate wealth evolution, at least in the restricted setting under scrutiny in this section, of the process described in Section 3.4 when $T \to \infty$. First, let us assume that, for some reason, all agents would remain within radius r (say, since r is sufficiently large) for all periods $t = 1, 2, \ldots$. Then our payoff process would evolve according to, under Equations (3.4.1) and (3.4.5),

$$\mathbf{Y}_{t+1} = f(\mathbf{Y}_t), \quad f(\mathbf{z}) = \mathbf{z} + \delta(\bar{\mathbf{z}}\mathbb{1} - \mathbf{z}),\ f : \mathbb{R}^n \to \mathbb{R}^n,\ t = 0, 1, 2, \ldots \tag{3.5.1}$$

where $\mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$, $\mathbf{Y}_t$=$(Y_{1,t}, \ldots, Y_{n,t})$, and for a vector $\mathbf{z} \in \mathbb{R}^n$, $\overline{\mathbf{z}}$ denotes the average $\frac{z_1+\ldots+z_n}{n}$.[23]
We first note that $\overline{\mathbf{Y}_s} = \overline{\mathbf{Y}_t}$ for all s, t, which we could interpret as a "zero sum condition".[24] This follows inductively since

$$\begin{aligned}\overline{\mathbf{Y}_{t+1}} = \overline{f(\mathbf{Y}_t)} = \overline{\mathbf{Y}_t + \delta(\overline{\mathbf{Y}_t}\mathbb{1} - \mathbf{Y}_t)} &= \frac{\sum_{i=1}^n Y_{i,t} + \delta\sum_{i=1}^n (\frac{Y_{1,t}+\ldots+Y_{n,t}}{n}) - \delta\sum_{i=1}^n Y_{i,t}}{n} \\ &= \frac{\sum_{i=1}^n Y_{i,t}}{n} + \frac{\delta\sum_{i=1}^n Y_{i,t} - \delta\sum_{i=1}^n Y_{i,t}}{n} = \overline{\mathbf{Y}_t}.\end{aligned}$$

This in turn means that $\overline{\mathbf{Y}_t} = \overline{\mathbf{Y}_0}$ for all periods t, where $\mathbf{Y}_0$ refers to the initial conditions $(Y_{1,0}, \ldots, Y_{n,0})$. Therefore

$$\mathbf{Y}_{t+1} - \overline{\mathbf{Y}_0}\mathbb{1} = f(\mathbf{Y}_t) - \overline{\mathbf{Y}_0}\mathbb{1} = \mathbf{Y}_t + \delta(\overline{\mathbf{Y}_0}\mathbb{1} - \mathbf{Y}_t) - \overline{\mathbf{Y}_0}\mathbb{1} = (1-\delta)(\mathbf{Y}_t - \overline{\mathbf{Y}_0}\mathbb{1}).$$

Hence

$$\frac{\left\|\mathbf{Y}_{t+1} - \overline{\mathbf{Y}_0}\mathbb{1}\right\|}{\left\|\mathbf{Y}_t - \overline{\mathbf{Y}_0}\mathbb{1}\right\|} = 1 - \delta,$$

which means that $\mathbf{Y}_t \to \overline{\mathbf{Y}_0}\mathbb{1}$ as $t \to \infty$, with a linear rate of convergence, since $1 - \delta \in (0, 1)$.[25] This immediately leads to our next result.

Proposition 3.5.2 ("Convergence of wealth levels")**.** Consider the model sketched in Section 3.4, under the specializations indicated in this section (random components are zero, determination of $\bar{Y}_{p,t}$ by uniform averaging of agents' wealth levels within radius r, zero moving costs). Assume that movement order — who makes the first relocation decision in each round? — is determined randomly (and independently) by a process that selects player 1 as the first player to move with probability p, for $0 < p < 1$. Then, if there are only $n = 2$ agents, their payoffs/wealths $Y_{1,t}$ and $Y_{2,t}$ converge to $\frac{Y_{1,0}+Y_{2,0}}{2}$ as $t \to \infty$ almost surely, for their initial endowments $Y_{1,0}$ and $Y_{2,0}$.

Proof. We assume that $Y_{1,0} \neq Y_{2,0}$, for otherwise the proposition is trivially true. Without loss of generality, let $Y_{2,0} > Y_{1,0}$. By Proposition 3.5.1 and its proof, the poor agent will 'chase' the rich, who in turn tries to escape. Whenever the poor is successful (when he is last to move), both players are within distance r, so $\mathbf{Y}_{t+1} = f(\mathbf{Y}_t)$. Otherwise (when the rich is last to move), $\mathbf{Y}_{t+1} = \mathbf{Y}_t$. So if there is a random process determining which agent will move first, then

$$\mathbf{Y}_{t+1} = \begin{cases} \mathbf{Y}_t & \text{with probability } p, \\ f(\mathbf{Y}_t) & \text{with probability } 1-p, \end{cases}$$

Hence, for all $0 \leq p < 1$, $\mathbf{Y}_t \to \overline{\mathbf{Y}_0}\mathbb{1}$ almost surely,[26] by our previous derivations. □

[23] This process can be analyzed by use of results from Markov chain theory — or, opinion dynamics, for that matter — but we sketch an elementary solution in the following.

[24] What one agent loses in wealth, the other gains. In sum, the average remains unchanged.

[25] In the appendix, we prove this result when time is continuous instead of discrete as considered here.

[26] The convergence is "almost surely" since the probability mass of the set $\{a \in \{0,1\}^{\mathbb{N}} \mid a \text{ has only finitely many } 1\}$ under any measure that assigns positive probability to 1 is zero. In other words, if you throw a coin infinitely often, the probability is zero that there will be only heads from some time onward.

Remark 3.5.4. For the case of $n > 2$ agents, the situation is (only) slightly different. In general, it is possible for some players to form a group first (say, in a three player setting, the poor will chase the rich), while others "stay out" (the middle agent sticking to his position). Once a group's wealth approaches an average value, inter-group chasing will take place. Finally, all agents will have the same payoff almost surely (although, in the general $n > 2$ agent case, this common payoff does not have to be $\overline{\mathbf{Y}_0}$).

Remark 3.5.5. Proposition 3.5.2 is, in a sense, a counterpart to Proposition 3.5.1. By the latter, if the process discussed in Section 3.4 is started with random initial endowments for all agents, there will be no equilibrium, in each round of relocation decisions, whence agents will be relocating and, more particularly, chasing each other. By the former proposition, this chasing will eventually terminate — at infinity, at the latest — when all agents have converged to the same wealth levels. At this point, again by the latter result, a Nash equilibrium is reached.

To summarize Propositions 3.5.1 and 3.5.2 and the corresponding remarks, in the current setup, agents will generally relocate on the basis of wealth inequalities; in fact, the poor will be chasing the rich. In the long-run, due to the neighborhood effects and the fact that agents cannot — e.g., exogenously — increase their wealth levels, this chasing will eventually entail assimilation of wealth levels. When all agents' wealth levels are finally equal, agents will reach an equilibrium where no one, trivially, has an incentive to deviate, that is, to relocate. We also note that our results imply that this standard model — without, e.g., random, exogenous, shocks to agents' wealth levels — cannot entail, at least in the long run, a Pareto distribution for agent wealth because this requires a sufficient degree of inequality among agents' wealth levels. Concerning Zipf distributions for city sizes, our analysis has provided no insights, except, maybe, that clustering tendencies can generally be expected, since most agents will chase someone richer than them.

Remark 3.5.6. As our final remark in this subsection, we note that we may provide a kind of 'master equation' for our city size and wealth dynamics process via the notation introduced in this section. Assuming the random components $\epsilon_{i,t}$ to be zero, wealth and cities evolve according to, for $\mathbf{p}_0$ and $\mathbf{Y}_0$ given,

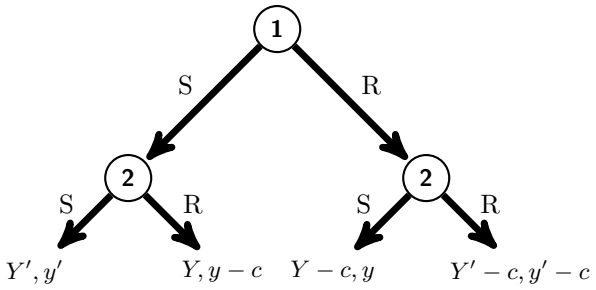
$$\begin{aligned} \mathbf{Y}_{t+1} &= (1-\delta)\mathbf{Y}_t + \delta \mathbf{W}_{\mathbf{p}_{t+1}} \mathbf{Y}_t, \\ \mathbf{p}_{t+1} &= \begin{cases} P_{\pi_1}(\mathbf{p}_t; \mathbf{Y}_t), & \text{with probab. } \Pi(\pi_1), \\ P_{\pi_2}(\mathbf{p}_t; \mathbf{Y}_t), & \text{with probab. } \Pi(\pi_2), \\ \vdots \end{cases} \end{aligned} \tag{3.5.2}$$

where $\mathbf{W}_{\mathbf{p}_t}$ is the matrix with entries $[\mathbf{W}_{\mathbf{p}_t}]_{ij} = w_{\mathbf{p}_{it},\mathbf{p}_{jt}}$ and $\Pi(\pi)$ denotes the probability of drawing permutation $\pi : [n] \to [n]$ from the set of all permutations on $[n]$; in the random uniform case, $\Pi(\pi) = \frac{1}{n!}$. Here, in the notation of $P_\pi(\mathbf{p}_t; \mathbf{Y}_t)$, we have also indicated the dependence of relocation decisions on the wealth structure.

Strategic agents

As we have mentioned before, the agents we have introduced in our model play myopic best responses to the 'current state of the world'. In this sense, these agents are in fact 'short-sighted': their best responses may be pretty bad choices once other agents have relocated, in the current period. Moreover, even if their choice is good for period t, another may have been preferable had periods $t+1, t+2, \ldots$ also been taken into account. Of course, a *strategic* player, knowing of the short-sightedness of the other players, could change his behavior accordingly, by prediction of the myopic players' decisions.[27] To illustrate, consider a two-player setup where player 1 has, in some fixed period, wealth Y and player 2 has wealth y. Assume that $Y > y$, so that agent 1 is the richer. Assume a situation as in the game tree shown. Agent 1 is first to make her relocation decision, and assume that agent 1 currently resides at a position *inside* a radius r of agent 2's location. To simplify, we think of the basic choices of both agents as either to stay at their current positions (S) or to relocate to a position inside/outside the influence range of the

[27] For an in-depth discussion of this general topic, see Schipper (2011).



other agent (R) (the relocation choice of the wealthier agent can only be to escape, while that of the less wealthy agent can only be to chase the wealthier). If player 1 were myopic, as discussed hitherto, her decision to relocate would soley be based on the moving costs. Assume, for simplicity, that these costs are lump-sum of size $c > 0$ whenever an agent decides to relocate and 0 whenever an agent decides to stay. Hence, if player 1 were myopic, she would relocate whenever $\frac{Y+y}{2} - 0 < Y - \frac{c}{\delta}$, or, equivalently, $c < \delta\frac{Y-y}{2}$. In contrast, if 1 were strategic, then 1 would base her decision to relocate on whether agent 2 would 'fight' her relocation (and also relocate, within radius r of agent 1) or not and on whether relocating is generally more profitable than staying. Hence, she would relocate whenever agent 2 would not relocate and when $Y - c \geq Y' = (1-\delta)Y + \delta\frac{Y+y}{2}$ — the left-hand side of the last equation is the maximum she could earn when she chose R in the above game tree and the right-hand side is what she would earn if she chose S, since player 2 would, in fact, never play R in the latter case. But, since the specification is symmetric, agent 1 would thus relocate whenever $c \geq \delta\frac{Y-y}{2}$ and $c < \delta\frac{Y-y}{2}$, since myopic agent 2 would relocate precisely when a myopic agent 1 would. Hence, since this requirement on c is contradictory, a strategic agent 1 would, in fact, never relocate, if she was first to move (trivially, she would not relocate if she was outside a radius r of agent 1's location). The reason is, to summarize, that moving costs are either so high that it does not pay to relocate, or else, if they are low enough, then agent 1 would also relocate, resulting in the lowest possible payoff of $Y' - c$ for agent 1.

Thus, the behaviors of a strategic agent and a myopic agent can be quite different. However, the inference problems that a strategic agent faces may become quite challenging, as agent number size increases and as the strategic agent may also want to include future time points $t, t+1, \ldots$, appropriately discounted, into consideration. Moreover, the problem becomes even more complicated when moving costs are not lump-sum but, e.g., linear in distance, and when w_{p_i,p_j} assumes more complex forms (than a uniform weighting when the distance between p_i and p_j is less than r, and zero otherwise). Thus, we find the inclusion of more rational players unrealistic, from a practical viewpoint concerning the inference problems such agents would have to solve,[28] while it is, of course, unclear how strategic agents would change resulting distributions of city sizes, as is our work's main goal.

City size distributions

We analytically derive city size distributions for our simple setup as discussed in Section 3.4 for small numbers n of agents under the assumptions stated at the beginning of the current section; in particular, we assume that moving costs are always zero. For simplicity, assume that the agents' world X is a one-dimensional grid of size N, with N points, that is, $X = \{1, \ldots, N\}$. Define, as indicated above, a city as a 'continguous' group of agents on the grid with no 'empty' spaces in between. Moreover, assume that r in Equation (3.4.5) is 1, so that agents have to 'live' directly next to each other in order to be in the same neighborhood (i.e., be influenced by each other's wealth levels) and assume that N is sufficiently large.[29] Also assume that all agents have different initial incomes Y and that they are initially placed uniformly randomly on the grid (with the restriction that players cannot occupy the same grid positions). We remark that our analysis here is no more than a toy analysis because what really interests us are city size distributions in the case when n becomes large. Nonetheless, even the present analysis will be

[28] Full rationality would also require knowledge of whether other agents are myopic or strategic, knowledge of *their* moving costs in case of moving cost heterogeneity, etc.

[29] Such that, e.g., agents can always relocate to empty regions.

insightful, in particular, since it outlines agents' clustering tendencies in our model.

The case $n = 2$. In the case of two agents, there can only be one or two cities; under random uniform initialization, the probability of one city is $\frac{2}{N}$, and the probability of two cities is $1 - \frac{2}{N}$.[30] Naturally, if there are two cities, both will consist of one agent, and if there is one city, it will consist of two agents. In the first round $t = 0$, if the poor player is first to move, there will subsequently be two cities since the rich player, being the last to move, will move away from the poor. If the rich player is first to move, there will be one city. So, if both events are equally likely, there will be two cities with probability $1/2$ and one city with probability $1/2$. This probability distribution is stable over time.

The case $n = 3$. Assume there are three players A, B, C with initial incomes $A < B < C$. In the case of three players, there can be one, two, or three cities. Initially, the probability of one city with three agents is $\frac{6}{N(N-1)} = O(1/N^2)$, the probability of two cities with one and two agents is $\frac{6(N-3)}{N(N-1)} = O(1/N)$, and the probability of three cities with one agent each is the remainder, $1 - \frac{6(N-3)}{N(N-1)} - \frac{6}{N(N-1)} = 1 - \frac{6(N-4)}{N(N-1)}$. After one round of relocations, city size distributions are as shown in Table 3.1. Note that, although the case $n = 3$ is really just a toy case, the 'Zipf outcome' — largest city has double the size of the second largest, which obtains in the case of two cities — is much more likely now than under random placement of the agents. Also note that clustering — either one or two cities — now has probability $2/3$ (independent of the grid size N), whereas it has probability $O(1/N)$ for random allocations, which quickly converges toward zero as N becomes large. This probability distribution is stable over time.

Movement order	1 city	2 cities	3 cities
A,B,C			1
A,C,B		1	
B,A,C			1
B,C,A		1	
C,A,B	1		
C,B,A	1		
Total Probability	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{2}{6}$

Table 3.1: Table entries are probabilities of the respective outcome (1, 2, or 3 cities) under the given movement order. Missing probabilities are taken as zero.

3.6 Simulation results

We consider the following parametrization. The world is a one-dimensional grid of size N, with places $p \in X = \{1, \ldots, N\}$, where we choose $N = 1000$. Moreover, we set the number of agents, n, to 300. Agents conduct a local search for optimal grid positions by considering the ρ, $\rho \in \mathbb{N}$, places 'around' their current position for potential relocation. Agents $i \in [n]$ have moving costs of

$$c(p_i, q) = \begin{cases} \chi \left\| p_i - q \right\| & \text{if } \left\| p_i - q \right\| \leq \rho \\ \infty & \text{else} \end{cases},$$

where $p_i, q \in X$ (p_i being agent's i current position), and where we set χ to the 'low' value of 0.001. Each agent i is initially assigned a random place $p_i \in X$ and a random wealth level $Y_{i,0} \sim U(x_0, x_1)$, where $U(x_0, x_1)$ is the continuous uniform density function on the interval $[x_0, x_1]$. As to determining the average wealth at any given grid point $p \in X$, we use the (normalized and truncated) normal density function centered at p with the 'low' variance of 1; we set the density weights to zero for agents at positions $q \in X$, with $\|p - q\| > 5$. Finally, as before, we define a city as a continguous set of occupied grid points with no empty spaces in between. We summarize the model calibration in Table 3.2.

[30] If n players are placed in N bins, the probability that they are all next to each other — i.e., their positions are $(i, i+1, \ldots, i+n)$ for some $i = 1, \ldots, (N-n)$ — is $\frac{N-(n-1)}{\binom{N}{n}} = \frac{n!}{N(N-1)\cdot\ldots\cdot(N-n+2)}$.

Parameter	Value	Meaning
N	1000	grid size
n	300	number of agents
χ	0.001	inertia/moving cost parameter
x_0	50	$Y_{i,0} \sim U(x_0, x_1)$
x_1	60	$Y_{i,0} \sim U(x_0, x_1)$
δ	$\in [0, 1]$	adaption rate
ρ	$\in \{5, 10, 20, 30, 100, 400, N\}$	spatial reach
T	$\in \{5000, 10000\}$	number of periods
w_{p_i,p_j}	$\tilde{f}(p_j; p_i, 1, p_i - 5, p_i + 5)$	weight for agent j, at p_j, with respect to agent i, at p_i

Table 3.2: Model calibration. By $\tilde{f}(x; \mu, \sigma^2, a, b)$ we denote the (adequately normalized and truncated) normal density function with mean μ, variance σ^2, and truncation parameters a and b.

In the subsequent analyses, all values discussed are averages over 10 runs, and numerical results are taken after $T = 5000$ iterations, unless stated otherwise. Also note that we estimate β on the basis of the 10% richest, mainly due to our small size of n (e.g., it would not be expedient to estimate a regression on the basis of $1\% \cdot n = 3$ data points).

3.6.1 Linear growth

First, consider the case where the random component in (3.4.1) is close to zero and iid across agents. We set

$$\epsilon_{i,t+1} \sim \mathcal{N}(0.2, 1),$$

where the mean of 0.2 is chosen so as to avoid that agents get poorer, on average, from period to period due to positive moving costs. Note also that if δ was zero (and ignoring moving costs) this would entail that $Y_{i,t}$, defined in (3.4.1), follows, in expectation, an affine-linear growth process,

$$\mathbb{E}[Y_{i,t}] = \mathbb{E}[Y_{i,t-1} + \mu] = \mu \cdot t + Y_{i,0},$$

where, in our case, $\mu = 0.2$. By simulation, we find that such a process, by itself, leads to a large Pareto coefficient β for the top 10% wealthiest of around 36.69 after $T = 5000$ periods, where the fit is very good, with R^2 value of about 98%.[31]

We outline results in Figures 3.6, 3.7, 3.8, 3.9 and Table 3.3. Figures 3.6 and 3.7 show a distribution of wealth and agents across time and grid positions for different parameter values of ρ and δ. In the figures, the y-axis has periods $t = 0, \ldots, T$ (from top to bottom) (with $T = 5000$), and the x-axis has grid positions $1, \ldots, N = 1000$; the blue color marks empty grid points and a stronger red signals higher wealth of the agent occupying the respective position. We see how different parameter values affect the distributional pattern of cities in the world.

An increase of the adaption rate δ (shown in Figure 3.6) leads to a larger (and faster) agglomeration of wealth and agents in a particular area of the world. Note that, when δ becomes larger, agents do not become richer, on average, but wealth becomes shared by a larger group of agents, which entails increased clustering since the group of agents that attracts others increases. For sufficiently large δ and ρ, we find a typical 'poor chasing the rich' outcome; rich agents cluster in one 'end' of the world, with the poorer attaching close-by. That the rich concentrate in an 'end' of the world is not surprising in this case, since in our model, the best way for the rich to escape the poor is to locate in an area with as little potential influence of the poor upon them as possible. It is intuitive that this leads to a large (rich) city in one 'end' of the world and several smaller cities in the direction of the opposite end, because all agents attempt to secure a place within the world's (single) rich habitat, and because rich agents have no incentive — due to the size of ρ, which is not large enough to reach isolated grid points, and the lack

[31] We simulated with different levels of x_0 and x_1 and always found that β is significantly larger than 10 after 5000 periods.

of attractiveness of reachable points — to leave their neighborhood. Moreover, the fact that places in the rich neighborhood are thus scarcer than other places may be considered an endogenous analogue of an increased 'rental price' in rich environments that would typically be observed in real economies.

Figure 3.7 illustrates how ρ affects distributional patterns of cities in the world. The smaller ρ, the more does the world decompose into smaller local 'settlements' distributed more evenly. In the bottom right plot of Figure 3.7, we display a situation where ρ is a random variable; with probability π_i (proportional to the size of agent i's wealth), we set $\rho = N$ and with $1 - \pi_i$, we set $\rho = 10$. This leads to a slightly larger agglomeration of wealth and agents again, but also to individual rich agents moving to isolated places from time to time, and thus to a more even spread of agents across the world than in Figure 3.6. It may additionally lead to slightly 'better' Zipfean city size distributions, as discussed in Figure 3.8 and Table 3.3.

From Table 3.3, we deduce that for many different calibrations of ρ and δ, our model entails city size distributions with Zipf parameter α in the 'right' range between 0.8 and 1.5. As a reference, note that we found by simulation that distributing $n = 300$ agents randomly among $N = 1000$ grid points implies a coefficient α of around 2.128 with R^2 value of around 0.839. Our model performs much 'better', even with a value of δ equal to 0; for example, for $\delta = 0$ and $\rho \in \{10, 20\}$, α is on average smaller than 1.5 after $T = 5000$ periods and the R^2 value is close to 90%. We obtain 'best' results for δ above/around 1% and ρ relatively small, in the range between $\rho = 20$ and $\rho = 100$; the R^2 fit is then usually larger than 90% and α is close to unity. Moreover, Figure 3.9 shows that, at least for specific parameter settings of ρ and δ, α is quite stable over time and it usually takes fewer than 1000 periods for it to settle within a narrowly defined band around its asymptotic value.

Concerning the Pareto wealth coefficient β, its size is typically magnitudes too large under the given calibration and the 'linear' wealth growth process. While the fit is usually good (R^2 above 90% for small δ) — note also that the 'correct' *form* of the wealth distribution function is usually apparently reproduced by our model (cf. Figure 3.8), i.e., a Boltzmann-Gibbs type distribution for 'small' wealth levels w and a Pareto distribution for 'large' w — the lowest value recorded in the simulations summarized in Table 3.3 is more than five times larger than observed in real economies. This is a rather unsurprising result, given the high level of β under $\delta = 0$ and ignoring moving costs (see above), the system's inherent tendency toward convergence (cf. Proposition 3.5.2), and the thus implied assimilation of agents' wealth levels. Accordingly, Gini coefficients, which measure inequality, are also quite small, usually only marginally exceeding 0.2, which is a lower value than observed for most real economies world-wide, where wealth Gini coefficients are usually found to lie in the range 0.6-0.8 and income Gini coefficients in the range 0.3-0.5 (cf. Davies et al., 2009). Moreover, in Figure 3.9, we see that β is much less stable than α under the given calibration, displaying considerable fluctuation over time.

		δ					
		0.001	0.005	0.015	0.05	0.1	0.5
α	size	1.662	1.198	0.994	0.961	0.942	0.938
	R^2	0.883	0.898	0.920	0.941	0.920	0.905
β	size	142.85	90.90	45.45	15.62	14.92	15.62
	R^2	0.976	0.956	0.938	0.876	0.834	0.767

		ρ				
		5	10	20	400	10, π_i
α	size	0.651	0.847	0.892	1.563	0.888
	R^2	0.770	0.909	0.918	0.893	0.830
β	size	142.86	125.0	90.90	50.0	166.67
	R^2	0.972	0.939	0.921	0.978	0.965

		ρ			
		5	10	20	10, π_i
α	size	0.725	0.758	0.868	0.823
	R^2	0.892	0.849	0.911	0.943
β	size	200.0	76.92	76.92	55.55
	R^2	0.972	0.865	0.818	0.908

		ρ		
		5	10	20
α	size	0.705	1.448	1.483
	R^2	0.570	0.893	0.890
β	size	38.46	33.33	25.64
	R^2	0.963	0.968	0.977

Table 3.3: Calibration as in Table 3.2. From left to right and top to bottom: $\rho = 100$, $\delta = 0.05$, $\delta = 0.015$, $\delta = 0$. Linear growth.

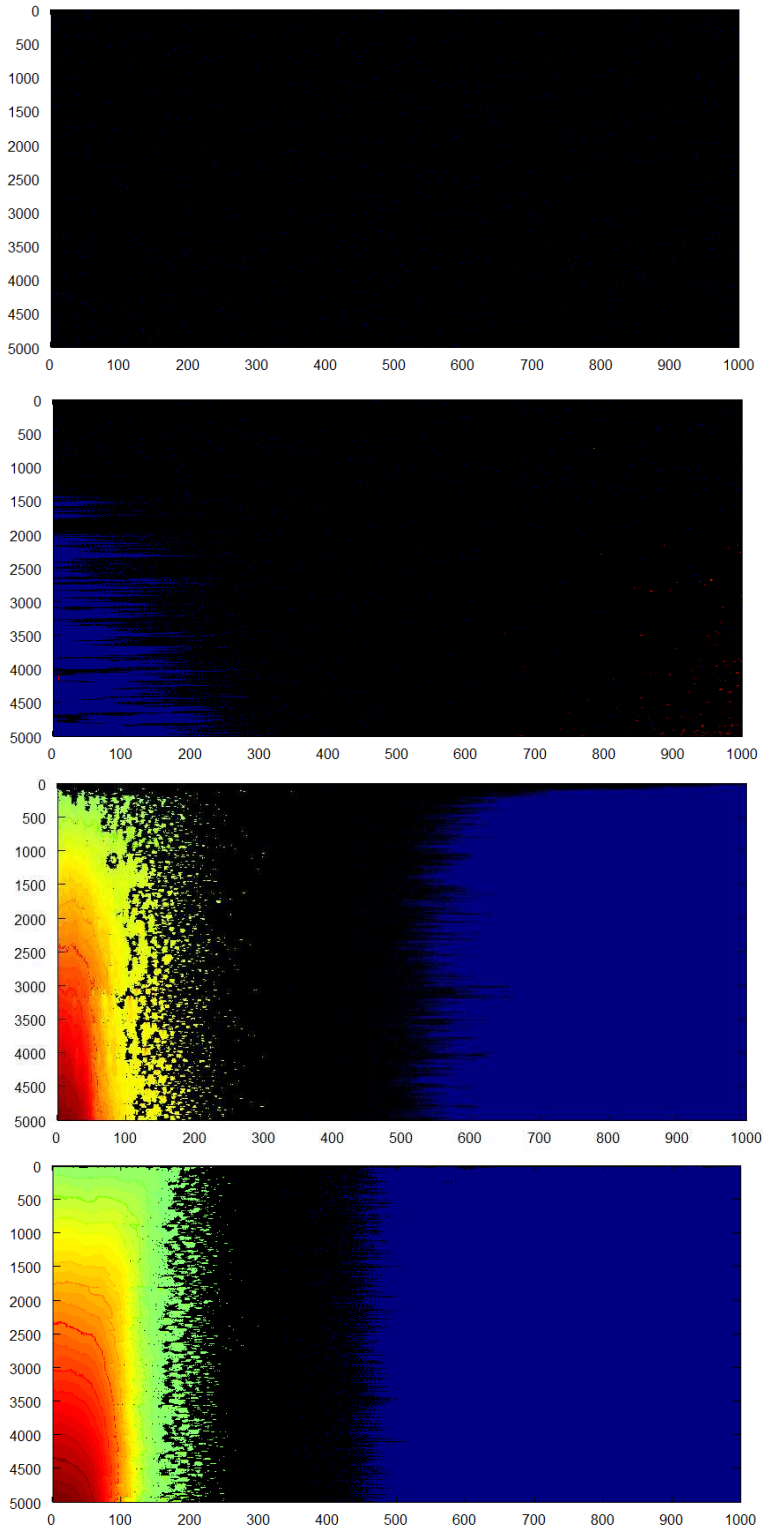


Figure 3.6: Calibration as in Table 3.2, and $\rho = 100$ throughout. From top to bottom: $\delta = 0.001$, $\delta = 0.005$, $\delta = 0.05$, $\delta = 0.5$. Linear growth.

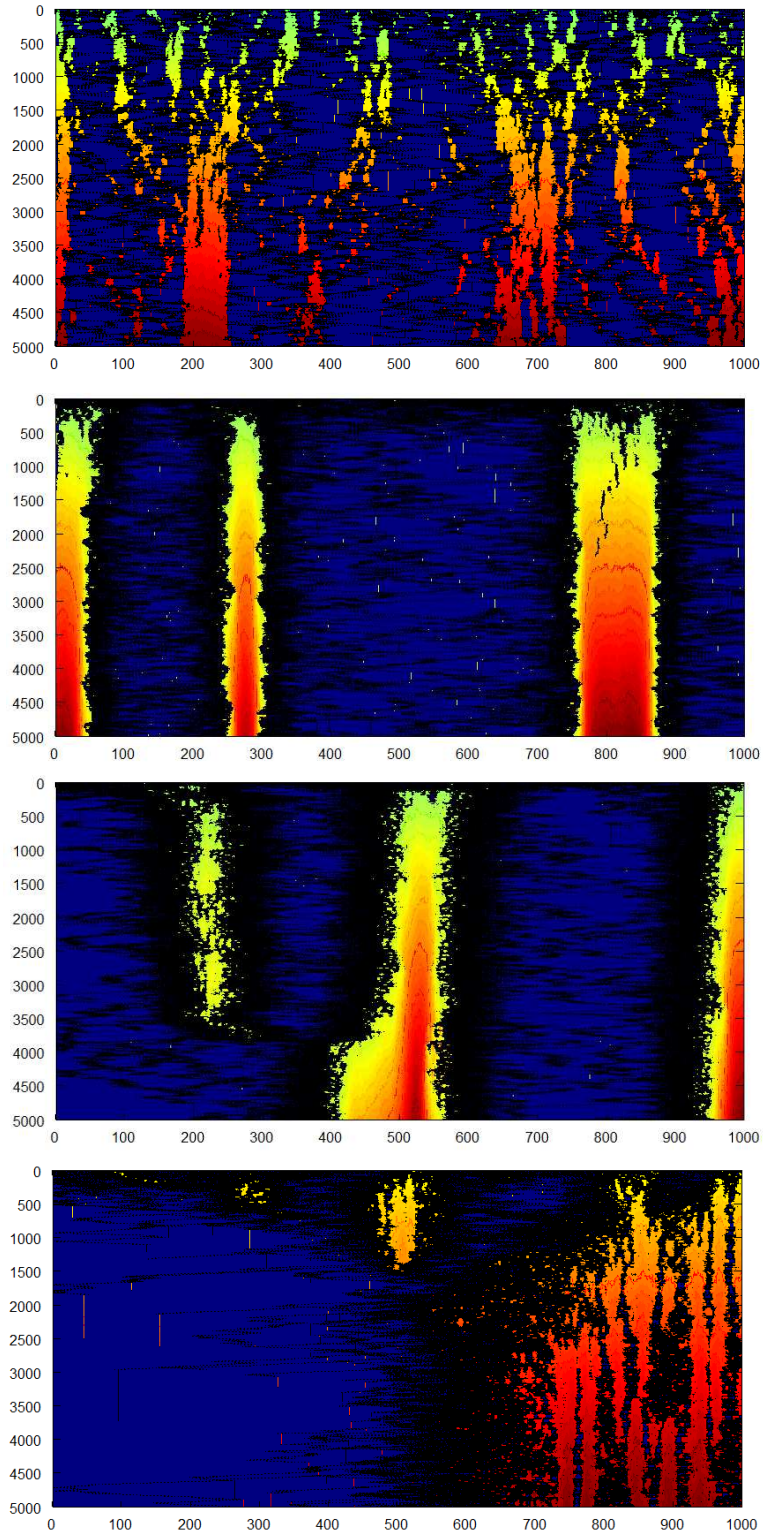


Figure 3.7: Calibration as in Table 3.2, and $\delta = 0.05$ throughout. From top to bottom: $\rho = 5$, $\rho = 10$, $\rho = 20$, $\rho = N$ with prob. π_i and $\rho = 10$ with prob. $1 - \pi_i$. Linear growth.

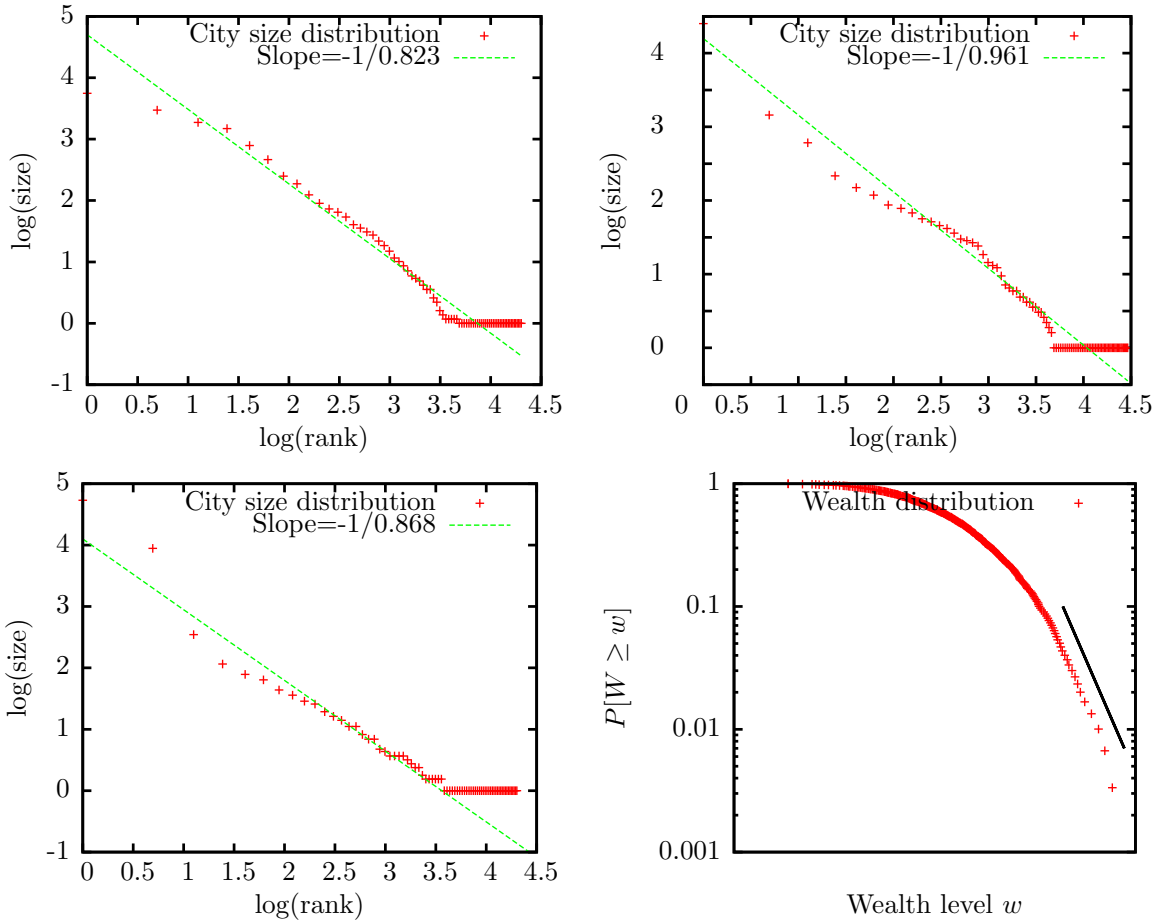


Figure 3.8: City size and wealth distributions after $T = 5000$ periods. Calibration as in Table 3.2, and $\delta = 0.05$ throughout. From left to right and top to bottom: $\rho = 10$ with prob. $1 - \pi_i$, $\rho = 100$, $\rho = 20$, $\rho = 5$. The marked slope in the wealth distribution curve has value -200. Linear growth.

3.6.2 Exponential/proportional growth

Next, consider the following specification of $\epsilon_{i,t+1}$,

$$\epsilon_{i,t+1} \sim \mathcal{N}(\mu Y_{i,t}, \kappa \|Y_{i,t}\|), \tag{3.6.1}$$

where $\mu \in (0,1)$ and $\kappa \in (0,1)$. Note that, if δ, the adaption rate toward a neighborhood's average wealth level, were zero and ignoring moving costs, this would entail the following evolution of $Y_{i,t+1}$, as defined in (3.4.1):

$$Y_{i,t+1} = (1+\mu)Y_{i,t} + \nu_{i,t+1}, \tag{3.6.2}$$

where $\nu_{i,t+1} \sim \mathcal{N}(0, \kappa \|Y_{i,t}\|)$, which implies, in expectation, an exponential growth process,

$$\mathbb{E}[Y_{i,t}] = (1+\mu)^t Y_{i,0}.$$

Specifications (3.6.1) and (3.6.2) can be interpreted as an instantiation of 'proportional attachment'; the increase in an agent's wealth level between two periods is proportional, in expectation, to the size of the agent's current wealth level — the richer experience a larger increment. That we choose the variance of $\epsilon_{i,t+1}$ to be a function of $\|Y_{i,t}\|$ seems plausible and entails a constant relative standard deviation of $\epsilon_{i,t+1}$ across agents. In the following, we set $\mu = 0.0005$ and $\kappa = 0.1$. By simulation, we find that this

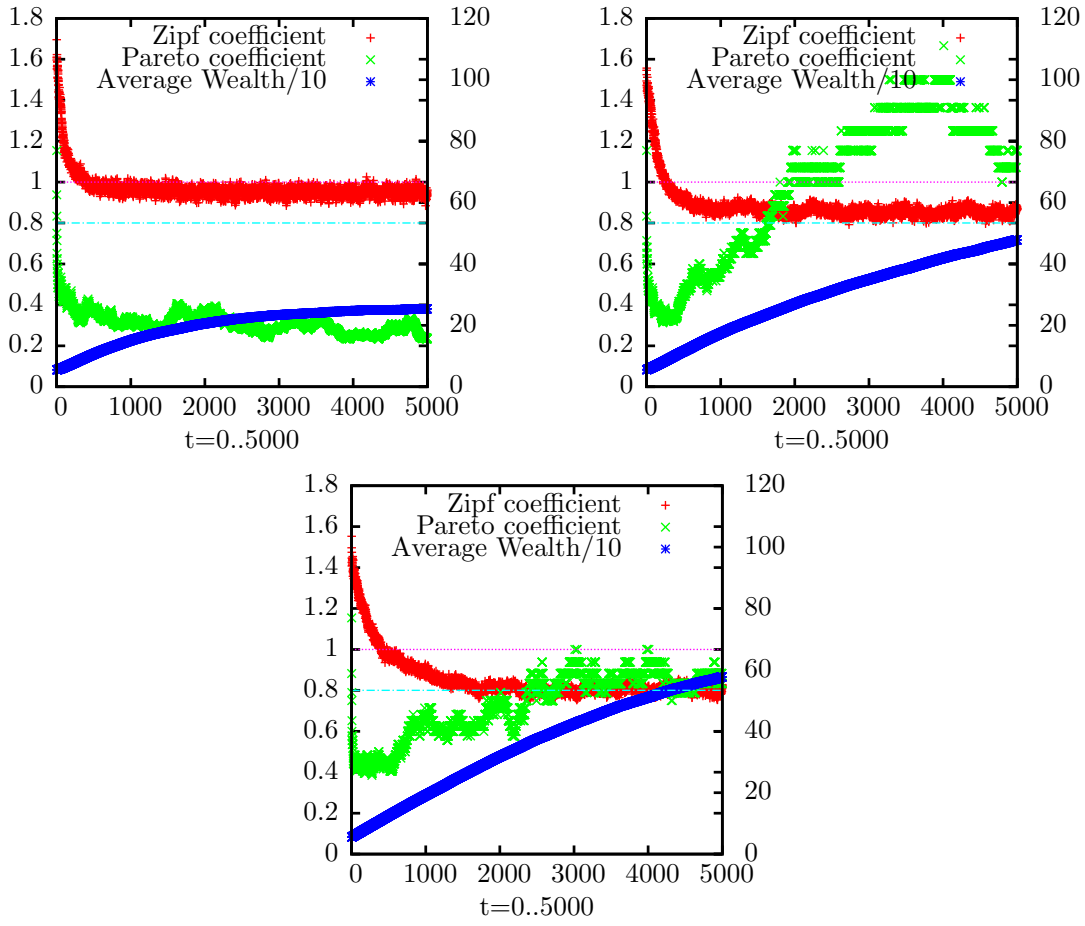


Figure 3.9: Parameter evolution of α, β, and average wealth over time. Average wealth and β are drawn with respect to the y_2-axis indicated on the right side of each plot. Calibration as in Table 3.2, and $\delta = 0.05$. From left to right: $\rho = 100$, $\rho = 20$, $\rho = 10$ and π_i. Linear growth.

parametrization leads to a coefficient β of around 3.02, with R^2 value of around 0.950, under $\delta = 0$ and no moving costs.

We summarize results in Table 3.4 and Figure 3.10. When contrasting with the results under linear growth, we find that the growth process may also affect city size distributions. For example, under $\delta = 0.015$, α is closer to unity under exponential growth than under linear growth (for $\rho = 5, 10, 20$), with no worse R^2 values. The best result is, again, obtained for δ around 1% and ρ relatively small, this time around 30.[32] In that case, α is close to unity and β is smaller than 3 after 5000 periods, with R^2 values larger than 90%. Figure 3.10 shows that β is in the range between 2 and 3 after about 4500 periods and seems to remain stable, although still fluctuating more heavily than α.

3.7 Conclusion

In the current work, we have investigated city size and wealth distributions in a unified framework. Our primary goal was the study of city size distributions in an economy, for which we have proposed a Tiebout-like sorting model in which boundedly rational agents have utilities on wealth and relocate, by playing myopic best responses to the current 'state of affairs', on the basis of the attractiveness of

[32] For larger ρ, we frequently noticed a 'poverty trap', in which agents become poorer initially — due to higher average moving costs — and can then not recover due to the small size of μ.

		δ			
		0.001	0.005	0.015	0.02
α	size	1.490	1.500	1.157	1.147
	R^2	0.890	0.891	0.927	0.938
β	size	3.401	6.667	2.873	2.994
	R^2	0.960	0.947	0.921	0.813

		ρ			
		5	10	20	10, π_i
α	size	0.664	0.969	0.988	1.281
	R^2	0.852	0.928	0.938	0.891
β	size	12.821	11.765	6.172	15.385
	R^2	0.930	0.930	0.939	0.955

		ρ		
		5	10	20
α	size	0.850	1.295	1.321
	R^2	0.566	0.855	0.875
β	size	2.81	2.73	2.403
	R^2	0.948	0.947	0.947

Table 3.4: Calibration as in Table 3.2. From left to right and top to bottom: $\rho = 30$, $\delta = 0.015$, $\delta = 0$. Exponential growth.

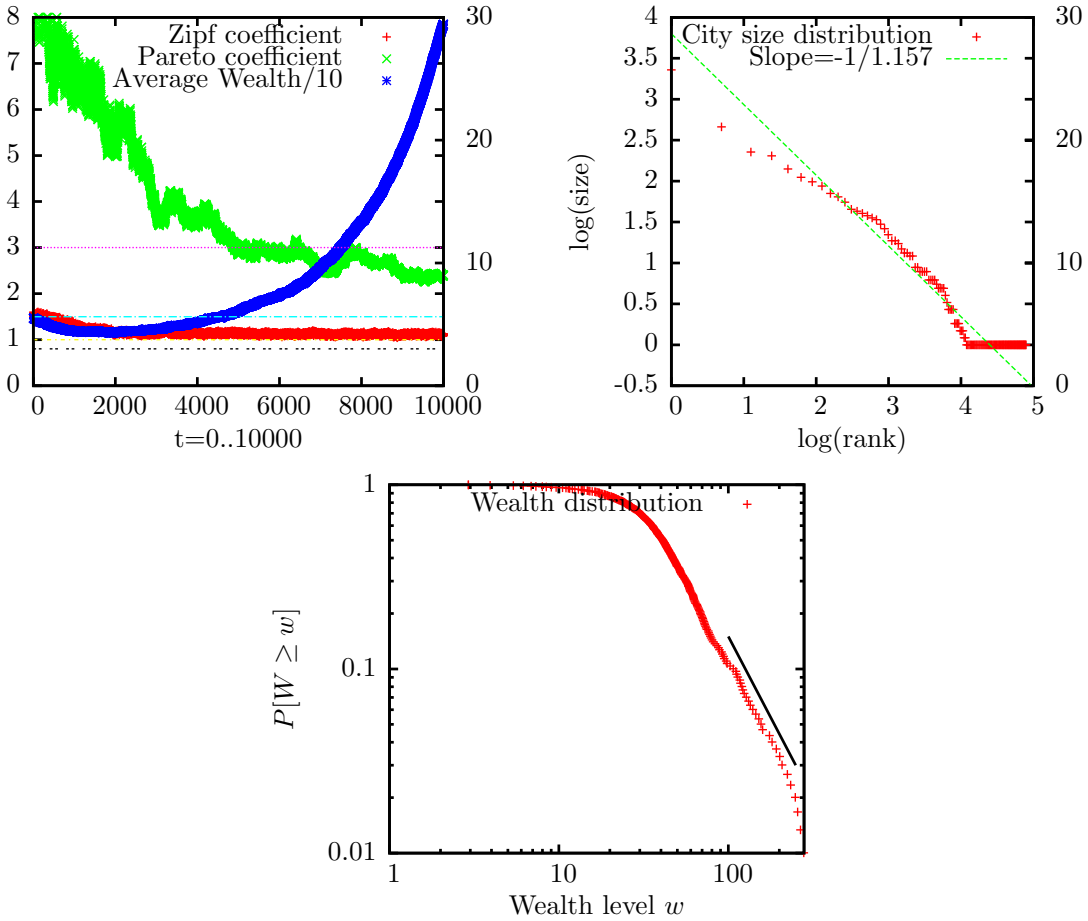


Figure 3.10: Calibration as in Table 3.2, and $\delta = 0.015$ and $\rho = 30$. Parameter evolution of α, β, and average wealth over time, and city size and wealth distributions after $T = 5000$ periods. Average wealth is drawn with respect to the y_2-axis indicated on the right side of the respective plot. The marked slope in the wealth distribution curve has value -2.873. Exponential growth.

neighborhoods, which, in turn, have (positive or negative) externalities upon their community members. We have examined analytical results for this setup — such as the non-existence of pure strategy Nash

equilibria in the simultaneous move game underlying our agents' relocation decisions such that the model we have specified is one of the type where the poor are chasing the rich who, in turn, flee from the former — and we have also challenged the myopia condition on our agents, showing that behavior of more rational agents would plausibly be different, although also considerably more costly in terms of processing costs for inference. Moreover, we have shown, by simulation, that our model seems to be very well capable of reproducing the Zipf phenonemon whereby a plot of city size versus rank yields (approximately) a straight line in log-log-space with slope closely centered around -1. Since our agents' decision variables are wealth, we are naturally led, within our framework, to the investigation of wealth distributions in our economy, which, 'in reality', follow a Pareto distribution form. We have argued, in the analytical results in Section 3.5, that our approach entails an assimilation of wealth levels, over time, thus generating a too 'flat' or 'equal' distribution among the richest of a society. To this end, we have added a stochastic component to individual agents' wealth levels, which may be specified such that it is in accordance with the rule of proportional attachment and implies an exponential growth process. Nontrivially, this apparently leads to the desired result of an economy in which, concurrently, both the Zipf and Pareto law are obeyed.

A few remarks on our results must be made. First, the fact that exponential growth is required by our model seems to be too specific an assumption, on the one hand, since exponential economic growth is presumably a feature of the last few hundred years exclusively, in effect only since the industrial revolution, whereas most of human history was, by all appearances, characterized by virtually no growth at all, at least on average. On the other hand, it evidently opposes the wealth distribution models proposed in econophysics, which assume zero-sum wealth processes. Two things might be answered to this; namely, that a) the universal applicability of Pareto's law is apparently still not fully ascertained to date, and some authors call it a property of capitalist societies (cf. Düring, Matthes, and Toscani, 2008), for which exponential growth is valid, allegedly. Moreover, b) it must be said that our model does presumably not necessarily require exponential growth but could probably do with a proportional attachment rule without such implications (e.g., due to high enough moving costs, some agents having large negative net wealth and thus clearing the balance, etc.); also note that in Figure 3.10, there is in fact practically no growth at all for the first 5000 periods and still the Pareto coefficient approaches 3.

Next, the result that Zipf's law is apparently 'best' reproduced for small values of δ around 1% appears to be a plausible and consistent outcome. It is known that, for example, peer effects at the work-place are usually between 5 and 15% (cf. Ichino and Maggi, 2000; Shvydko, 2008), and neighborhood effects in a community should certainly be a bit lower, as the degree of interaction between individuals, there, is plausibly lower. Overall, in the simulations, we find δ on the order of 1% and a 'moderate' spatial reach ρ of, about, 20 to 100 — these numbers plausibly depend on the grid size N — to lead to city size distributions that match well the Zipf paradigm.[33] Thus, in our model, as in Mansury and Gulyás (2007), 'bounded rationality' in the form of restricted spatial reach seems to be a contributing factor to generating Zipfean city size distributions.

For future work, it might be of interest to find an 'agent-based' — instead of a stochastic — solution for generating Pareto wealth distributions within our Tiebout-like sorting model. Potentially, the inclusion and adequate weighting of further variables such as those discussed in Section 3.4 might be helpful here, but we think this unlikely. Presumably, unless a process is defined whereby agents veritably lose and gain wealth beyond the rather small neighborhood effects implemented — be it through 'gambling' or other mechanisms — a wealth distribution that is unequal 'enough' in the rich tail will not be achieved. Another aspect that might be worth investigating is to individualize the spatial reach parameter ρ and/or moving costs, e.g., by time-dependent adaptation of χ (increases might be thought of as representing the concept of 'familiarity'/stronger social ties over time) or wealth-dependent adaptation (it might be argued that wealthier agents should incur smaller costs, possibly, due to superior technology, information asymmetries, etc.). Although we are uncertain whether this can qualitatively affect outcomes, it must be noted that individualizing saving propensities in the wealth distribution model of A. Chatterjee, Chakrabarti, and Manna (2003) has turned an exponential regime into a Pareto regime. To check for the robustness of our results, implementing our model in a two-dimensional scenario might be a further

[33] It is worthwhile pointing out that $\delta = 0$, that is, absence of neighborhood effects, typically does not result in Zipf coefficients close to 1, as indicated in the respective tables given above.

aspect of concern,[34] as well as running more extensive simulations, including larger grid sizes N and agent set sizes n.

Appendix 3.A Continuous time

It is insightful to solve the problem discussed in Section 3.5, the convergence of $\mathbf{Y}_t$ defined in Equation (3.5.1) as $t \to \infty$, in continuous time. To this end, first rewrite (3.5.1) as

$$\mathbf{Y}_{t+1} - \mathbf{Y}_t = \delta(\overline{\mathbf{Y}_t}\mathbb{1} - \mathbf{Y}_t).$$

In continuous time ($t \in [0, \infty)$), this equation becomes

$$\dot{\mathbf{Y}}(t) = \mathbf{A}\mathbf{Y}(t), \qquad \mathbf{A} = \frac{\delta}{n} \underbrace{\begin{pmatrix} 1-n & 1 & \dots & 1 \\ 1 & 1-n & \dots & 1 \\ \vdots & \dots & \ddots & \vdots \\ 1 & 1 & \dots & 1-n \end{pmatrix}}_{=:\mathbf{B}}. \tag{3.A.1}$$

Matrix $\mathbf{B}$ has characteristic polynomial $p_{\mathbf{B}}(\lambda) = \det(\mathbf{B} - \lambda \mathbf{I}_n) = \big(\lambda - (-n)\big)^{n-1}\lambda$. To see this, consider more generally the matrix $\mathbf{C}_n \in \mathbb{R}^{n \times n}$,

$$\mathbf{C}_n := \begin{pmatrix} \alpha & 1 & \dots & 1 \\ 1 & \alpha & \dots & 1 \\ \vdots & \dots & \ddots & \vdots \\ 1 & 1 & \dots & \alpha \end{pmatrix},$$

for $\alpha \in \mathbb{R}$, whose characteristic polynomial is found by considering the following determinant,

$$\begin{aligned} \det(\mathbf{C}_n - \lambda \mathbf{I}_n) &= \det \begin{pmatrix} \alpha-\lambda & 1 & \dots & 1 \\ 1 & \alpha-\lambda & \dots & 1 \\ \vdots & \dots & \ddots & \vdots \\ 1 & 1 & \dots & \alpha-\lambda \end{pmatrix} = \det \begin{pmatrix} \alpha-\lambda-1 & -(\alpha-\lambda-1) & \dots & 0 \\ 1 & \alpha-\lambda & \dots & 1 \\ \vdots & \dots & \ddots & \vdots \\ 1 & 1 & \dots & \alpha-\lambda \end{pmatrix} \\ &= (\alpha-\lambda-1)\det(\mathbf{C}_{n-1} - \lambda \mathbf{I}_{n-1}) + (\alpha-\lambda-1)^{n-1}. \end{aligned}$$

We claim that $\mathbf{C}_n$ has characteristic polynomial $\big(\lambda - (\alpha-1)\big)^{n-1}(\lambda - (\alpha+n-1))$, and, via the given representation of $\det(\mathbf{C}_n - \lambda \mathbf{I}_n)$, this easily follows inductively. Then, substituting $\alpha = 1-n$ yields our above claim for the matrix $\mathbf{B}$. Consequently, matrix $\mathbf{A}$ above has eigenvalues $(\lambda_1, \dots, \lambda_{n-1}, \lambda_n) = (-\delta, \dots, -\delta, 0)$. Since $\mathbf{A}$ is symmetric, we can therefore diagonalize $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{V} \begin{pmatrix} -\delta & 0 & \dots & \dots & 0 \\ 0 & -\delta & 0 & \dots & 0 \\ \vdots & \dots & \ddots & \dots & \vdots \\ 0 & 0 & \dots & -\delta & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} \mathbf{V}^\intercal,$$

where $\mathbf{V}$ is an orthonormal matrix, i.e., $\mathbf{V}^\intercal\mathbf{V} = \mathbf{I}_n$ (denoting by $\mathbf{V}^\intercal$ the transpose of $\mathbf{V}$). Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ denote the columns of $\mathbf{V}$, i.e., the eigenvectors corresponding to the eigenvalues $-\delta, \dots, -\delta, 0$. Since (3.A.1) represents a system of linear differential equations, it is well known (cf. Hirsch and Smale, 1995) that its general solution is given by

$$\mathbf{Y}(t) = \sum_{i=1}^{n} c_i \exp(\lambda_i t)\mathbf{v}_i,$$

[34] In a preliminary two-dimensional study, we could replicate Zipf city size distributions, under plausible parameter calibrations.

where $c_i \in \mathbb{R}$ are constants. In our case, therefore

$$\mathbf{Y}(t) = \sum_{i=1}^{n-1} c_i \exp(-\delta t)\mathbf{v}_i + c_n \exp(0 \cdot t)\mathbf{v}_n = \sum_{i=1}^{n-1} c_i \exp(-\delta t)\mathbf{v}_i + c_n \mathbf{v}_n.$$

Hence, since $\delta > 0$, $\mathbf{Y}(t) \to c_n \mathbf{v}_n$ as $t \to \infty$. Let us determine the limit vector $c_n \mathbf{v}_n$. Note that the eigenvalue zero (to which $\mathbf{v}_n$ corresponds) entails

$$\mathbf{A}\mathbf{v}_n = \mathbf{0} \cdot \mathbf{v}_n = \mathbf{0}.$$

By the structure of $\mathbf{A}$, this then implies that $\mathbf{v}_n = (\beta, \ldots, \beta)^\intercal$, for some constant $\beta \in \mathbb{R}$. Moreover, we obtain the coefficients vector $\mathbf{c} = (c_1, \ldots, c_n)^\intercal$ by evaluating $\mathbf{Y}(t)$ at time 0,

$$\mathbf{Y}_0 = \mathbf{Y}(0) = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \cdots + c_n \mathbf{v}_n = \mathbf{V}\mathbf{c}.$$

Since $\mathbf{V}$ is orthonormal, then,

$$\mathbf{c} = \mathbf{V}^\intercal \mathbf{Y}_0.$$

By inspection of both sides of this equality, c_n is the sum $\mathbf{v}_n^\intercal \mathbf{Y}_0 = v_{n,1} Y_{0,1} + \cdots + v_{n,n} Y_{0,n} = \beta(Y_{0,1} + \cdots + Y_{0,n}) = \beta \overline{\mathbf{Y}_0} n$. Thus,

$$c_n \mathbf{v}_n = \overline{\mathbf{Y}_0} \cdot n \left(\beta^2 \ldots \beta^2\right)^\intercal = \overline{\mathbf{Y}_0} \cdot n \left(1/n \ldots 1/n\right)^\intercal = \overline{\mathbf{Y}_0} \mathbb{1},$$

since $\mathbf{v}_n$ is orthonormal, i.e., $n\beta^2 = \mathbf{v}_n^\intercal \mathbf{v}_n = 1$. In other words, $\mathbf{Y}(t) \to \overline{\mathbf{Y}_0} \mathbb{1}$ as $t \to \infty$, which is the same result we have derived in Section 3.5 for the discrete analogue of Equation (3.A.1). In Figure 3.11 below, we show a sample evolution path of $\mathbf{Y}(t)$ in $(t, Y_1(t))/(t, Y_2(t))$ space for $n = 2$ agents and initial endowments $Y_1(0) = 0.9$, $Y_2(0) = 0.2$, and $\delta = 0.1$ and $\delta = 0.2$.

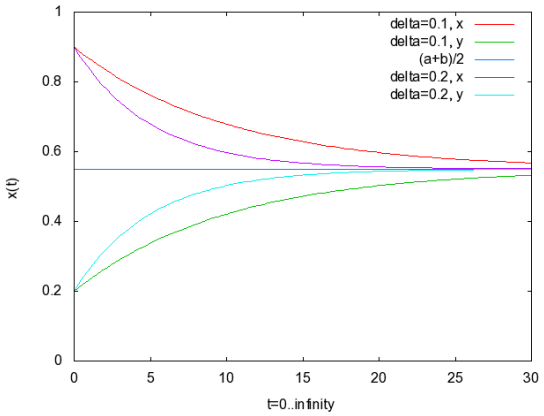


Figure 3.11: Sample evolution path of $\mathbf{Y}(t) = (Y_1(t), Y_2(t))$ in $(t, Y_1(t))/(t, Y_2(t))$ space for $n = 2$ agents and initial endowments $a = Y_1(0) = 0.9$, $b = Y_2(0) = 0.2$, and $\delta = 0.1$ and $\delta = 0.2$. In the plot, we denote Y_1 as x and Y_2 as y.

Bibliography

[1] Felix Auerbach. "Das Gesetz der Bevölkerungskonzentration". In: *Petermann's Geographische Mitteilungen* 59 (1913), pp. 74–76.

[2] Robert L. Axtell and Richard Florida. *Emergent cities: a microeconomic explanation of Zipf's Law.* Paper presented at the Society of Computational Economics, Yale University. 2001.

[3] Albert-László Barabási and Réko Albert. "Emergence of scaling in random networks". In: *Science* 286 (1999), pp. 509–512.

[4] Lucien Benguigui and Efrat Blumenfeld-Lieberthal. "A dynamic model for city size distribution beyond Zipf 's law". In: *Physica A: Statistical Mechanics and its Applications* 384 (2 2007), pp. 613–627.

[5] Jess Benhabib, Alberto Bisin, and Shenghao Zhu. "The distribution of wealth and fiscal policy in economies with finitely lived agents". In: *Econometrica* 79 (1 2011), pp. 123–157.

[6] Steven Brakman, Harry Garretsen, Richard Gigengack, Charles van Marrewijk, and Rien Wagenvoort. "Negative feedbacks in the economy and industrial location". In: *Journal of Regional Science* 36 (4 1996), pp. 631–652.

[7] Steven Brakman, Harry Garretsen, Charles van Merrewijk, and Marianne Van Den Berg. "The return of Zipf: a further understanding of the rank-size distribution". In: *Journal of Regional Science* 39 (1 1999), pp. 183–213.

[8] Sam Bucovetsky and Amihai Glazer. *Peer Group Effects, Sorting, and Fiscal Federalism.* Working Papers 091006. University of California-Irvine, Department of Economics, May 2010. URL: `http://ideas.repec.org/p/irv/wpaper/091006.html`.

[9] Arnab Chatterjee, Bikas K. Chakrabarti, and Smarajit S. Manna. "Money in Gas-Like Markets: Gibbs and Pareto Laws". In: *Physica Scripta T* 2003.T106 (2003), p. 36. URL: `http://stacks.iop.org/1402-4896/2003/i=T106/a=008`.

[10] Satyajit Chatterjee. "Transitional dynamics and the distribution of wealth in a neoclassical growth model". In: *Journal of Public Economics* 54.1 (May 1994), pp. 97–119. URL: `http://ideas.repec.org/a/eee/pubeco/v54y1994i1p97-119.html`.

[11] Siyan Chen. "Agent-based modeling for wealth and income distributions". PhD thesis. Universita Politecnica Delle Marche, Nov. 2011.

[12] Ricardo Coelho, Peter Richmond, Joseph Barry, and Stefan Hutzler. "Double power laws in income and wealth distributions". In: *Physica A* 387 (2008), pp. 3847–3851. DOI: `10.1016/j.physa.2008.01.047`. URL: `http://dx.doi.org/10.1016/j.physa.2008.01.047`.

[13] Frank A Cowell. *Inequality among the Wealthy.* CASE Papers /150. Centre for Analysis of Social Exclusion, LSE, June 2011. URL: `http://ideas.repec.org/p/cep/sticas/-150.html`.

[14] James B. Davies, Susanna Sandström, Anthony B. Shorrocks, and Edward N. Wolff. *The Level and Distribution of Global Household Wealth.* Working Paper 15508. National Bureau of Economic Research, Nov. 2009. URL: `http://www.nber.org/papers/w15508`.

[15] Donald Davis and David Weinstein. "Bones, Bombs, and Breakpoints: the Geography of Economic Activity". In: *American Economic Review* 92 (5 2002), pp. 1269–1289.

[16] Peter M. DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. "Persuasion Bias, Social Influence, And Unidimensional Opinions". In: *The Quarterly Journal of Economics* 118.3 (Aug. 2003), pp. 909–968. URL: http://ideas.repec.org/a/tpr/qjecon/v118y2003i3p909-968.html.

[17] Robert D. Dietz. "The estimation of neighborhood effects in the social sciences: An interdisciplinary approach". In: *Social Science Research* 31.4 (Dec. 2002), pp. 539–575. DOI: 10.1016/S0049-089X(02)00005-4. URL: http://www.sciencedirect.com/science/article/B6WX8-474GH72-3/2/405267dcf89813e5b06c79b624e79b35.

[18] Jeremiah Dittmar. *Cities, Institutions, and Growth: The Emergence of Zipf's law*. Working Paper. 2010.

[19] Serguei N. Dorogovtsev and Jose F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[20] Adrian A. Drăgulescu. *Applications of physics to economics and finance: Money, income, wealth, and the stock market*. Papers. arXiv.org, 2003. URL: http://EconPapers.repec.org/RePEc:arx:papers:cond-mat/0307341.

[21] Gilles Duranton. *City Size Distributions As A Consequence of the Growth Process*. CEP Discussion Papers dp0550. Centre for Economic Performance, LSE, Oct. 2002. URL: http://ideas.repec.org/p/cep/cepdps/dp0550.html.

[22] Bertram Düring, Daniel Matthes, and Giuseppe Toscani. *Kinetic equations modelling wealth redistribution: a comparison of approaches*. eng. Discussion paper series // Zentrum für Finanzen und Ökonometrie, Universität Konstanz 2008,03. Konstanz: Universität Konstanz, 2008. URL: http://hdl.handle.net/10419/32177.

[23] Joshua M. Epstein and Robert L. Axtell. *Growing artificial societies: social science from the bottom up*. Washington, DC, USA: The Brookings Institution, 1996. ISBN: 0-262-55025-3.

[24] Armin Falk and Andrea Ichino. "Clean Evidence on Peer Effects". In: *Journal of Labor Economics* 24.1 (Jan. 2006), pp. 39–58. URL: http://ideas.repec.org/a/ucp/jlabec/v24y2006i1p39-58.html.

[25] Davide Fiaschi and Matteo Marsili. *Distribution of Wealth and Incomplete Markets: Theory and Empirical Evidence*. Discussion Papers 2009/83. Dipartimento di Economia e Management (DEM), University of Pisa, Pisa, Italy, Apr. 2009. URL: http://ideas.repec.org/p/pie/dsedps/2009-83.html.

[26] Noah E. Friedkin and Eugene C. Johnsen. "Social influence and opinions". In: *Journal of Mathematical Sociology* 15 (3-4 1990), pp. 193–205.

[27] Xavier Gabaix. "Zipf's Law For Cities: An Explanation". In: *Quarterly Journal of Economics* 114 (1999), pp. 114–739.

[28] Xavier Gabaix and Yannis M. Ioannides. "The evolution of city size distributions". In: *Handbook of Regional and Urban Economics*. Ed. by J. V. Henderson and J. F. Thisse. Vol. 4. Handbook of Regional and Urban Economics. Elsevier, 2004. Chap. 53, pp. 2341–2378. URL: http://ideas.repec.org/h/eee/regchp/4-53.html.

[29] M. Gardner. "Mathematical Games: The Fantastic Combinations of John Conway's New Solitaire Game Life". In: *Scientific American* 223 (1970), pp. 120–123.

[30] Gene M. Grossman and Elhanan Helpman. "Quality Ladders in the Theory of Growth". In: *Review of Economic Studies* 58.1 (Jan. 1991), pp. 43–61. URL: http://ideas.repec.org/a/bla/restud/v58y1991i1p43-61.html.

[31] Abhijit K. Gupta. "Models of wealth distributions: a perspective". In: *ArXiv Physics e-prints* (Apr. 2006). arXiv: physics/0604161. URL: http://arxiv.org/abs/physics/0604161.

[32] Morris W. Hirsch and Stephen Smale. *Differential Equations, Dynamical Systems and Linear Algebra*. Pure and Applied Mathematics. Index. London: Academic Press, 1995. ISBN: 0-12-349550-4. URL: http://opac.inria.fr/record=b1099208.

[33] John J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the National Academy of Sciences* 79.8 (Apr. 1, 1982), pp. 2554–2558. ISSN: 1091-6490. URL: http://www.pnas.org/content/79/8/2554.abstract.

[34] Andrea Ichino and Giovanni Maggi. "Work Environment And Individual Background: Explaining Regional Shirking Differentials In A Large Italian Firm". In: *The Quarterly Journal of Economics* 115.3 (Aug. 2000), pp. 1057–1090. URL: http://ideas.repec.org/a/tpr/qjecon/v115y2000i3p1057-1090.html.

[35] Beom Jun Kim and Sung Min Park. "Distribution of Korean family names". In: *Physica A: Statistical Mechanics and its Applications* 347 (Mar. 1, 2005), pp. 683–694. URL: http://www.sciencedirect.com/science/article/B6TVG-4D98BVV-7/1/3a44d5adb331dfee8f11e36d19c5eee7.

[36] Teuvo Kohonen. *Self-Organization and Associative Memory*. Berlin: Springer Verlag, 1984.

[37] Paul R. Krugman. *Geography and Trade*. 1st ed. Vol. 1. The MIT Press, 1992. URL: http://EconPapers.repec.org/RePEc:mtp:titles:0262610868.

[38] Paul R. Krugman. *The Self-Organizing Economy*. 1. publ. Cambridge, Mass. [u.a.]: Blackwell, 1996. VI, 122. ISBN: 1557866996. URL: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+186794517&sourceid=fbw_bibsonomy.

[39] Charles F. Manski. "Economic Analysis of Social Interactions". In: *Journal of Economic Perspectives* 14 (3 2000), pp. 115–136.

[40] Yuri Mansury and László Gulyás. "The emergence of Zipf's Law in a system of cities: An agent-based simulation approach". In: *Journal of Economic Dynamics and Control* 31.7 (July 2007), pp. 2438–2460. URL: http://www.sciencedirect.com/science/article/B6V85-4M5WHR4-1/1/7f1c1260f7c57c58a4dd58c86eca6053.

[41] Sasuke Miyazima, Youngki Lee, Tomomasa Nagamine, and Hiroaki Miyajima. "Power-law distribution of family names in Japanese societies". In: *Physica A: Statistical Mechanics and its Applications* 278.1-2 (Apr. 1, 2000), pp. 282–288. URL: http://www.sciencedirect.com/science/article/B6TVG-3YYTPKF-R/1/cdec3c89c55b5f2f191bcbe19512d192.

[42] Mark E. J. Newman. "Power laws, Pareto distributions and Zipf's law". In: *Contemporary Physics* 46 (Dec. 2005), pp. 323–351. arXiv: cond-mat/0412004. URL: http://arxiv.org/abs/cond-mat/0412004.

[43] Scott E. Page. "On the Emergence of Cities". In: *Journal of Urban Economics* 45.1 (1999), pp. 184–208. URL: http://EconPapers.repec.org/RePEc:eee:juecon:v:45:y:1999:i:1:p:184-208.

[44] Vilfredo Pareto. *Cours d'Economie Politique*. Genève: Droz, 1896.

[45] Vincenzo Quadrini. "Entrepreneurship, Saving and Social Mobility". In: *Review of Economic Dynamics* 3.1 (Jan. 2000), pp. 1–40. URL: http://ideas.repec.org/a/red/issued/v3y2000i1p1-40.html.

[46] William Rand, Daniel G. Brown, Scott E. Page, Rick Riolo, Luis E. Fernandez, and Moira Zellner. "Statistical validation of spatial patterns in agent-based models". In: *Proceedings of Agent Based Simulation* 4 (2003).

[47] William J. Reed and Barry D. Hughes. "From Genes families and genera to incomes and Internet file sizes: why power-laws are so common in nature". In: *Physical Review E* 66 (2002), pp. 067103+. DOI: 10.1103/PhysRevE.66.067103. URL: http://link.aps.org/abstract/PRE/v66/e067103.

[48] Peter Richmond, Stefan Hutzler, Ricardo Coelho, and Przemek Repetowicz. "A Review of Empirical Studies and Models of Income Distributions in Society". In: *Econophysics & Sociophysics: Trends & Perspectives*. Ed. by B. K. Chakrabarti, A. Chakraborti, and A. Chatterjee. WileyVCH, 2006. Chap. 5, p. 121.

[49] Kenneth T. Rosen and Mitchel Resnick. "The size distribution of cities: An examination of the Pareto law and primacy". In: *Journal of Urban Economics* 8.2 (1980), pp. 165–186. URL: http://EconPapers.repec.org/RePEc:eee:juecon:v:8:y:1980:i:2:p:165-186.

[50] Esteban Rossi-Hansberg and Mark L.J. Wright. "Establishment Size Dynamics in the Aggregate Economy". In: *American Economic Review* 97 (5 2007), pp. 1639–1666.

[51] Maria A. Santos, Ricardo Coelho, Géza Hegyi, Zoltán Néda, and José J. Ramasco. "Wealth distribution in modern and medieval societies". In: *Eur. Phys. J. Special Topics* 143 (2007), pp. 81–85.

[52] Thomas C. Schelling. *Micromotives and Macrobehavior*. W. W. Norton & Company Incorporated, 1978. URL: http://www.amazon.com/exec/obidos/tg/detail/-/0393090094/104-8667083-6377538?v=glance.

[53] Burkhard Schipper. *Strategic control of myopic best reply in repeated games*. eng. Working Papers, University of California, Department of Economics 11,5. Davis, Calif.: University of California, Department of Economics, 2011. URL: http://hdl.handle.net/10419/58405.

[54] Tetyana Shvydko. "Essays in Labor Economics: Peer Effects and Labor Market Rigidities". PhD thesis. The University of North Carolina at Chapel Hill, 2008.

[55] Herbert A. Simon. "On a class of skew distribution functions". In: *Biometrika* 42.3–4 (1955), pp. 425–440. DOI: 10.1093/biomet/42.3-4.425. eprint: http://biomet.oxfordjournals.org/content/42/3-4/425.full.pdf+html.

[56] Franti šek Slanina. "Inelastically scattering particles and wealth distribution in an open economy". In: *Phys. Rev. E* 69.4 (Apr. 2004), p. 046102. DOI: 10.1103/PhysRevE.69.046102.

[57] Kwok Tong Soo. "Zipf's Law for cities: a cross-country investigation". In: *Regional Science and Urban Economics* 35.3 (May 2005), pp. 239–263. URL: http://ideas.repec.org/a/eee/regeco/v35y2005i3p239-263.html.

[58] Lior Strahilevitz. *Exclusionary Amenities in Residential Communities*. University of Chicago, Law & Economics Working Paper No. 250; U of Chicago, Public Law Working Paper No. 98. Available at SSRN: http://ssrn.com/abstract=757388 or http://dx.doi.org/10.2139/ssrn.757388. 2005.

[59] Charles M. Tiebout. "A Pure Theory of Local Expenditures". In: *Journal of Political Economy* 64 (1956), p. 416. URL: http://ideas.repec.org/a/ucp/jpolec/v64y1956p416.html.

[60] Neng Wang. "An equilibrium model of wealth distribution". In: *Journal of Monetary Economics* 54.7 (Oct. 2007), pp. 1882–1904. URL: http://www.sciencedirect.com/science/article/B6VBW-4N1SSH5-1/1/d6c964fde60147aee4bfaf4e82c42686.

[61] Uri Wilensky. *NetLogo Wealth Distribution model*. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. 1998. URL: http://ccl.northwestern.edu/netlogo/models/WealthDistribution.

[62] Udny G. Yule. "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S." In: *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character:* 213 (1925). Ed. by R. Jsto, pp. 21–87. URL: http://links.jstor.org/sici?sici=0264-3960(1925)213%5C<21:AMTOEB%5C>2.0.CO;2-5%5C&size=LARGE%5C&origin=ads.

[63] George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading MA (USA), 1949.

STEFFEN P. EGER

PRESENT ADDRESS
Albblick 1
72160 Horb-Betra
(+49) 7482 9137370

PERMANENT ADDRESS
Neuwiesenäcker 1
72160 Horb-Betra
(+49) 7482 913636

EDUCATION

Eduard-Spranger-Gymnasium Freudenstadt, 2000

Ruprecht-Karls-Universität Heidelberg
Magister Artium Computerlinguistik, VWL, Anglistik, 2007
Diplom Mathematik, 2008

Goethe-Universität Frankfurt
Dr. cand., seit 2009

INDUSTRY AND RESEARCH EXPERIENCE

Programmer	SAP AG Walldorf	August 2007-August 2008
Wissenschaftlicher Mitarbeiter	Institut für deutsche Sprache Mannheim	November 2008-Octob. 2010

COMPUTER SKILLS

Python, C (very good), Matlab, Java, Lisp, Prolog (good), Perl, R (basic).

LANGUAGES

German (native), English (very good), French (good), Latin (reading), Spanish, Navajo (elementary)

PEER-REVIEWED PUBLICATIONS

Steffen Eger. Sequence Segmentation by Enumeration: An Exploration. The Prague Bulletin of Mathematical Linguistics, No. 100, 2013, pp. 113131. doi: 10.2478/pralin-2013-0017.

Steffen Eger (2013): A contribution to the theory of word length distribution based on a stochastic word length distribution model. Journal of Quantitative Linguistics, 20:3, 252-265.

Eger, S. (2013): Sequence alignment with arbitrary steps and further generalizations, with applications to alignments in linguistics. Information Sciences, 237: 287-304.

S. Eger, Restricted weighted integer compositions and extended binomial coefficients. Journal of Integer Sequences 16 (2013), #13.1.3.

Steffen Eger, An Agent-Based Sorting Model for City Size and Wealth Distributions, Proceedings of the European Conference on Complex Systems 2012, Springer Proceedings in Complexity 2014, pp. 955-967

Eger, S. (2012): S-Restricted Monotone Alignments: Algorithm, Search Space, and Applications. COLING 2012: 781-798

Eger, S. (2012): Lexical semantic typologies from bilingual corpora - A framework. Proceedings of *SEM 2012 (First Joint Conference on Lexical and Computational Semantics)

Steffen Eger (2012): The Combinatorics of String Alignments: Reconsidering the Problem, Journal of Quantitative Linguistics, 19:1, 32-53.

Ineta Sejane and Steffen Eger. Semantic typologies by means of network analysis of bilingual dictionaries. Proceedings of the workshop on comparing approaches to measuring linguistic differences, Gothenburg, Oct. 2011

Eger, S., and Sejane, I. (2010): Computing semantic similarity from bilingual dictionaries. JADT 2010: 10th International Conference on Statistical Analysis of Textual Data, 9-11 June 2010, Rome, Italy; pp.1217-1225

ACCEPTED PAPERS

Eger, S. (2012): Stirling's approximation for central polynomial coefficients. Preprint available at arxiv.org

WORKING PAPERS

Eger, S. (2013): (Failure of the) Wisdom of the crowds in an endogenous opinion dynamics model with multiply biased agents. Preprint available at arxiv.org

Eger, S. (2013): Opinion dynamics under opposition. Preprint available at arxiv.org

POSTERS

(Failure of the) Wisdom of the crowds in an endogenous opinion dynamics model with multiply biased agents, at ECCS 2013, Barcelona

AWARDS

Karl-Steinbuch Stipendium, *Adiuvaris*, 2008

Ehrenwörtliche Erklärung

Ich habe die vorgelegte Dissertation selbst verfaßt und dabei nur die von mir angegebenen Quellen und Hilfsmittel benutzt. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind sowie alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.