The Dynamics of Group Polarization

Carlo Proietti

Lund University

Abstract. Exchange of arguments in a discussion often makes individuals more radical about their initial opinion. This phenomenon is known as Group-induced Attitude Polarization. A byproduct of it are bipolarization effects, where the distance between the attitudes of two groups of individuals increases after the discussion. This paper is a first attempt to analyse the building blocks of information exchange and information update that induce polarization. I use Argumentation Frameworks as a tool for encoding the information of agents in a debate relative to a given issue a. I then adapt a specific measure of the degree of acceptability of an opinion (Matt and Toni 2008). Changes in the degree of acceptability of a, prior and posterior to information exchange, serve here as an indicator of polarization. I finally show that the way agents transmit and update information has a decisive impact on polarization and bipolarization.

1 Introduction

Almost sixty years ago MIT student J. A. Stoner observed and studied a strange group phenomenon that he classified as "risky shift". This term categorizes the tendency of a group to make decisions that are riskier than the average of the individual decisions of members before the group met [27]. Subsequent research in social psychology showed that a similar pattern applies more generally to change of attitude and opinion after debate. This phenomenon is nowadays famous as *Group-induced attitude polarization*. Understanding the dynamics that lead to polarization is particularly relevant in the era of social networks, because of the dramatic global effects they may cause. Indeed, virtual forums and political debate seem to accrue so-called *bipolarization effects*, i.e. the tendency of different subgroups to radicalize their opinions towards opposite directions [28].

A long tradition in social psychology has regarded polarization and bipolarization as a byproduct of social influence in groups [26], where the main explanatory mechanism is *social comparison* [11].[1] An alternative explanation is

[1] According to social comparison explanations, such as [26], polarization may arise in a group because individuals are motivated to perceive and present themselves in a favorable light in their social environment. To this end, they take a position which is similar to everyone else but a bit more extreme. This kind of explanation assumes a lot. Indeed, models that explain bipolarization effects by social comparison mechanisms usually postulate both positive influence by ingroup members and negative influence by outgroup members [16, 12]. However, a number of criticisms have been addressed towards the accuracy of empirical research showing the presence of negative influence in social interaction [19].

provided by *persuasive arguments theory*, which was developed and tested in a number of lab experiments in the 1970s [30].[2]

Both social comparison and persuasive arguments theory provide interesting clues for explaining polarization phenomena. However, much more is hidden behind the mechanisms of *information* transmission and update among agents. The present work is a first attempt towards the formal description of such mechanisms. The aim is to unravel all the possible building blocks of polarization. In this context we need to understand the notion of information in a general sense, wider than, e.g., knowledge or rational belief. Polarization and bipolarization are in fact distinctive features of real-life dynamics, where people form their views (say, decide how to vote or what to buy) by exchanging information with others, or by trusting more or less authoritative channels. Such informational items typically need not to be consistent, nor are acquired via a careful process of individual inquiry and strict rules of belief revision and belief update. Indeed in many such situations individuals deviate from Bayesianism, insofar as they update their beliefs by discarding some available evidence. This happens, e.g., when they display a dogmatic or selective attitude towards the information received.

In the present paper I adopt Argumentation Frameworks [9] as a formalizing tool, which are the most versatile tool to encode the type of informational items we are interested in, as well as the argumentative process of information exchange and update. Indeed, the theoretical tools provided by abstract argumentation will serve to the purpose of

1. Describing both the *total information available* relative to a debated issue and one agent's *partial information*.
2. Provide a measure of the *degree of acceptability* of the debated issue given the available information.
3. Encode the most important policies of *information transmission* between agents and of *information update*.
4. Assess how such policies can impact polarization and bipolarization about the given issue

Schematically, the argumentative process generating polarization works as in the workflow of Figure 1. Agent i possesses some information about a given issue a, represented by $I_i(\mathcal{F}, a)$, she transmits some of her information to agent, say, j ($T_{i,j}(\mathcal{F}, a)$), and the latter updates her previous information $I_j(\mathcal{F}, a)$ by combining it with $T_{i,j}(\mathcal{F}, a)$ via some operation $\star$ to be specified.

[2] This explanation assumes that individuals become more convinced of their view when they hear novel and persuasive arguments in favor of their position, and therefore "Group discussion will cause an individual to shift in a given direction to the extent that the discussion exposes that individual to persuasive arguments favoring that direction" [15]. Typically, models inspired by persuasive arguments theory do not assume negative influence of any kind, but presuppose homophily, i.e. stronger interaction with like-minded individuals [23], or biased assimilation of arguments [21].

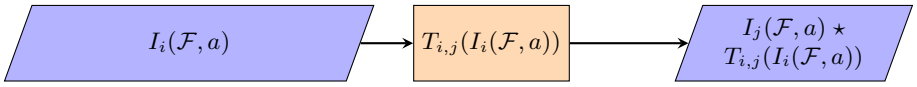


Fig. 1. Schematic flow of information transmission and update between agent i and j

I proceed as follows. Section 2 provides a short introduction to Argumentation Frameworks and shows how to define the scenario of a multi-agent debate. I also show how to apply the acceptability measure defined by [24] to encode the degree of acceptability of a given issue a. Section 3 introduces some relevant policies of information disclosure and update. Section 4 presents two main results that show the impact of these policies on polarization and bipolarization.

2 Argumentation Frameworks and multi-agent scenarios

An Argumentation Framework [9], AF for short, consists of a graph where nodes are arguments and a directed edge between a and b is to be read as "argument a attacks argument b". The formal definition is the following

Definition 1 (Pointed Argumentation Framework). *An Argumentation Framework is a 2-ple $\mathcal{F} = (A, R)$ where A is a finite and non-empty set of arguments and $R \subseteq A \times A$. A Pointed Argumentation Framework $\mathcal{F}, a$ consists of an Argumentation Framework $\mathcal{F}$ together with a specified $a \in A$.*

Example 1. Figure 2 provides the graphical representation of a pointed AF, where $A = \{a, b, c, d, e, f\}$, $R = \{(b, a), (c, a), (d, b), (e, c), (f, c)\}$ and the specified argument is a. This will serve us as a running example.

An AF is usually intended to represent a completed debate process. In our specific setting a pointed AF $\mathcal{F}, a$ is meant to encode what is sometimes called a "culturally given pool of arguments" [30] about one issue a, i.e. the full set of arguments and attacks between them that are available to a group of individuals debating over a. *Opinions* held by the participants are represented as sets of arguments they embrace. Conflicts between opinions can be formalised as attacks between sets of arguments. We say that an opinion X attacks an opinion Y if there is an attack $R(x, y) \in X \times Y$. For example, in Figure 2 it holds that $\{d, e, a\}$ attacks $\{b, c\}$ and viceversa.

One main purpose of argumentation theory is to identify which opinions are intuitively "acceptable". Such opinions are usually called *solutions* (or *extensions*). Typically, one solution should have at least two basic properties, i.e. *conflict-freeness* and *defense* of its own arguments. A set which combines these two properties is said to be *admissible*.

Definition 2 (Admissibility).

- *A set X is conflict-free if there is no $a, b \in X$ such that $R(a, b)$.*

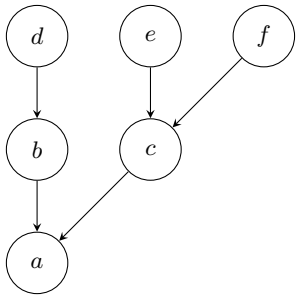


Fig. 2. An example of AF. Labelled nodes represent arguments. Relations of attack between arguments are indicated with an edge.

- *A set X defends an argument a if for all b such that $R(b, a)$ there is a $c \in X$ such that $R(c, b)$.*
- *A set X is admissible iff X is conflict-free and defends all its elements.*

Intuitively, conflict-freeness encodes the internal coherence of an opinion, in the sense that no argument attacks another. The largest conflict-free sets in Figure 2 are $\{a, d, e, f\}$ and $\{b, c\}$. The second condition (defending all its elements) encodes the fact that for an opinion to be *fully* acceptable it should be able to rebut all its counterarguments.[3] The AF in Figure 2 has ten admissible sets where the smallest is $\emptyset$ and the largest is $\{a, d, e, f\}$.

The opinions we are interested in for our case are the *pro* and the *contra* opinion about a given issue a. It is straightforward to identify the opinion contra a ($C(a)$) as the set of arguments that attack a, while the opinion pro a ($P(a)$) is the set of arguments that defend a, including a itself. What is left out is the *neutral* opinion $N(a)$, i.e. the set of arguments that neither attack nor defend a. These three opinions are then defined as follows:

Definition 3 (Pro, contra and neutral opinions).

- *$P(a)$ is the set of arguments b such that there is an R-path of even length, including length 0, from b to a*
- *$C(a)$ is the set of arguments b such that there is an R-path of odd length from b to a.*
- *$N(a)$ is the set of arguments b such that $b \notin P(a)$ and $b \notin C(a)$*

It is easy to ascertain that, in the pointed AF of Figure 2, $P(a) = \{a, d, e, f\}$, $C(a) = \{b, c\}$ and $N(a) = \emptyset$. Furthermore, the following holds:

Fact 4. *$P(a) \cap C(a) = \emptyset$ iff both $P(a)$ and $C(a)$ are conflict-free.*

In the following we assume that $P(a)$ and $C(a)$ are conflict-free for all the information frameworks that we consider.

[3] Admissibility is the basis of most of the solution concepts in the standard Dung's framework such as *preferredness*, *stability* and *groundedness*. For our present purposes we don't need to introduce them.

2.1 A multi-agent debate

If we regard a specific pointed AF $\mathcal{F}, a$ as the total information available about the issue a, then it is natural to encode the *partial information* available to a participant to the debate as a subgraph of $\mathcal{F}, a$, i.e. a partial representation of the argumentative pool.[4] This captures the fact that an individual may not be aware of some available arguments, or may even not be aware that some argument attacks another.[5] By consequence, the setup of a debate over a can be seen as a multiagent scenario where the information available to each agent i is defined as follows.

Definition 5 (Agent's information). *Given the total information* $\mathcal{F}, a = (A, R), a$*, the information available to agent* i *is* $I_i(\mathcal{F}, a) = (A_i, R_i), a$ *where* $A_i \subseteq A$*,* $a \in A_i$*,* $R_i \subseteq A_i \times A_i$ *and* $R_i \subseteq R$*.*

Hereafter I denote as $P_i(a)$, $C_i(a)$ and $N_i(a)$ the sets of pro, contra and neutral arguments of agent i about the issue a.

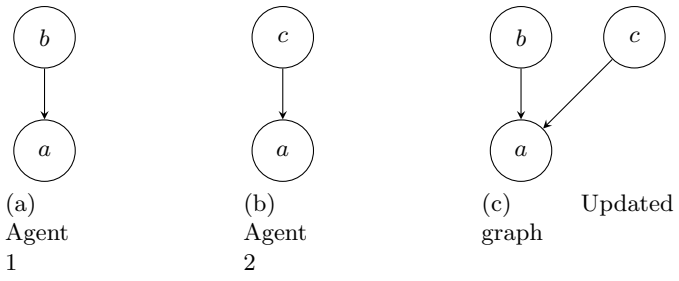


Fig. 3. Information of agents 1 and 2 and their merging

Example 2. Figure 3(a) (resp. 3(b)) represents the information $I_1(\mathcal{F}, a)$ available to Agent 1 (resp. $I_2(\mathcal{F}, a)$ available to Agent 2) at the initial state of a debate over a. Both are subgraphs of $\mathcal{F}, a$ in Figure 2. Here $C_1(a) = \{b\}$ and $C_2(a) = \{c\}$. Therefore both agents have distinct informational items contra a. Intuitively, when they merge such information together, as it happens in discussion, they should both get new information to the effect that $C_1(a) = C_2(a) = \{b, c\}$.

[4] A similar approach is taken by [25], [1] and [8]. There too the information base of an agent is encoded by a subset of a larger *universe* [8] or *universal argumentation framework* [25, 1].

[5] Being unaware that b attacks c is the case when one lacks the *warrant* for b to undermine c (see also [29]). To give an example, let c be the argument "Phosphorus is not visible in the sky tonight" and b be the argument "Look, Hesperus is there!". Clearly b constitutes an attack to c only if one is aware that Hesperus and Phosphorus are the same planet.

This illustrates polarization on the intuitive level, but does not yet provide a measure of it, which we shall present in the next section.

2.2 Degree of acceptability

We introduced admissibility as a criterion of full acceptability for an opinion in a debate. However, in many real-life scenarios involving a large number of arguments and rebuttals, full acceptability is a too strict requirement. In most cases the best thing one can do is weighting arguments *pro* and *contra*. This is also what seems to happen in the lab experiments on polarization and risky shift, where people is asked to provide *odds* for a certain decision or opinion [30] and polarization is measured as the shift between the initial and the final odds provided by the participants.

A large literature in abstract argumentation has recently developed to provide measures and weights for arguments and opinions in a debate (see among others [24, 20, 14]). For our present purposes we shall adopt a measure of acceptability provided by [24], which fulfills a series of useful properties for our case. We first define the set of attacks from set Y to set X in a framework $\mathcal{F}, a$, as $X^{\leftarrow Y}_{\mathcal{F},a} = \{(y,x) \in Y \times X \mid R(y,x)\}$. Then we can define the *degree of acceptability of X w.r.t. Y* as follows:

Definition 6 (degree of acceptability, Matt and Toni 2008).

$$d(X,Y) = \frac{1}{2}(1 + f(|Y^{\leftarrow X}_{\mathcal{F},a}|) - f(|X^{\leftarrow Y}_{\mathcal{F},a}|))$$

where $f : \mathbb{N} \to \mathbb{N}$ is defined as $f(n) = \frac{n}{n+1}$

We are specially interested in the measures $d(P(a), C(a))$ and $d(C(a), P(a))$. The nice properties of the measure d hang mostly on the fact that f is a monotonic increasing mapping s.t. $f(0) = 0$ and $lim_{n\to\infty} f(n) = 1$. Such properties are the following:

Fact 7. *The following properties hold for d.*

(a) $0 \leq d(X,Y) \leq 1$
(b) $d(P(a), C(a)) = 1 - d(C(a), P(a))$
(c) $d(P(a), C(a)) < \frac{1}{2}$ *iff* $|P(a)^{\leftarrow C(a)}_{\mathcal{F},a}| > |C(a)^{\leftarrow P(a)}_{\mathcal{F},a}|$
(d) $d(P(a), C(a)) > \frac{1}{2}$ *iff* $|P(a)^{\leftarrow C(a)}_{\mathcal{F},a}| < |C(a)^{\leftarrow P(a)}_{\mathcal{F},a}|$
(e) *If* $\mathcal{F}, a = (\{a\}, \emptyset), a$ *then* $d(P(a), C(a)) = d(C(a), P(a)) = \frac{1}{2}$

According to property (a) degrees of acceptability are scaled between 0 and 1. Properties (b), (c) and (d) mean that d measures the relative weight of arguments pro and contra a. (e) reflects the fact that a given issue with no arguments pro or contra is to be labeled as undecided.

Example 3. In our argumentative pool of Figure 2 the degree of acceptability $d(P(a), C(a))$ is $\frac{13}{24}$ and is therefore slightly favorable to the opinion supporting a.

Such measure d can work as the *actual degree of acceptability of* a. In a context where $\mathcal{F}, a$ represents the overall information available and there is no conclusive way of settling the truth, having a measure of this sort is the best we can hope for. In the following we indicate with $d(P_i(a), C_i(a))$ the *degree of acceptability of* a *for agent* i.

Example 4. It is straightforward to ascertain that $d(P_i(a), C_i(a))$ is $\frac{1}{4}$ for both agents in Figure 3(a) and Figure 3(b). If we merge the agent's information, as in Figure 3(c), then $d(P_i(a), C_i(a))$ becomes $\frac{1}{6}$ for each agent, i.e. it lowers. This is particularly interesting when we interpret Figure 3(c) as the new information available to Agent 1 and 2, as it would result from an information exchange and consequent information update. This is exactly what happens with attitude polarization: both agents radicalize their opinion contra a. Indeed, their degree of acceptability of a ends up being more extreme ($\frac{1}{6}$) than its average before entering the debate ($\frac{1}{4}$).

3 Information transmission and information update

Example 4 shows how polarization may arise in a group when agents share partial information about a given issue. This happened by agents merging their information together as the result of information exchange. However, the mechanisms of information exchange in a debate are way more complex than this. Such mechanisms need to be cut down into their basic components if one wants to capture all the possible ways in which polarization may arise. The basic components are clearly *information transmission* from a sender to a recipient and *information update* by the recipient.

3.1 Information transmission

Information transmission is the way agents disclose the information they possess. This is encoded as an operation $T^p_{i,j}$ where i is the sender, j the recipient and p is the disclosure policy adopted by the sender. The input of such operation is $I_i(\mathcal{F}, a)$, i.e. the information available to the sender. The disclosure policy p determines the output, which in principle can be any piece of information possessed by i. The latter is then a 2-ple (A', R') with the constraints $A' \subseteq A_i$ and $R' \subseteq R_i$.

There are many possible ways for a sender to transmit information to a recipient. One way is full disclosure o, which consists in one agent disclosing all the available information.

Definition 8 (Open disclosure). *The output of an open disclosure policy for agent* i *is*

$$T^o_{i,j}(I_i(\mathcal{F}, a)) = A_i, R_i$$

Policy o is typical of a debate where the agents' common interest is to share the best possible information or to settle an issue in the optimal way. But this is not what usually happens in more strategic situations where agents have different goals. People often discloses only the information that is useful for her to win a debate or to fortify her opinion among the audience, e.g. in judicial proceedings or panel discussions. A large set of strategies is available to agents in such contexts. Indeed, most of the situations that can be modeled by our framework are open to *cheap talk* [10], where agents are allowed to lie. In this context, the foundational game-theoretic analysis by [5] shows that, when the interests of the agents are not perfectly aligned, the equilibrium solution requires players not to be fully informative.[6]

Here we define a radical policy s, which consists in delivering only information that speaks in favor of one's opinion.

Definition 9 (Strategic disclosure). *The output of a strategic disclosure policy for agent i is defined by cases as $T^s_{i,j}(I_i(\mathcal{F}, a)) = A', R'$ where:*

- *if $d(P_i(a), C_i(a)) \geq \frac{1}{2}$*

$$A' = A_i \setminus \{b \in C_i(a)\}$$
$$R' = R_i \setminus \{(b, c) \in C_i(a) \times P_i(a)\}$$

- *if $d(P_i(a), C_i(a)) < \frac{1}{2}$*

$$A' = A_i \setminus \{b \in P_i(a)\}$$
$$R' = R_i \setminus \{(b, c) \in P_i(a) \times C_i(a)\}$$

3.2 Information update

When agent j receives new information from i she has to update her informational state in the light of it and of her prior information. This corresponds to an operation $\star^p$ which should output a new information state $I^p_j(\mathcal{F}, a)$ on the basis of her previous information I_j and the information $T_{i,j}$ received by i. Here again, many policies p are available to agents for such an update. For the present purposes I restrict my attention to two of them. The first option is again an open policy $\star^{ou}$, which consists in fully accepting the information received by the sender. This gives rise to the following definition

Definition 10 (Open update). *Let $X = (A, R), a$ and $Y = A', R'$ then the output for agent i of $\star^{ou}(X, Y)$ is a pointed AF $(A^{ou}, R^{ou}), a$ where*

$$A^{ou} = A \cup A'$$
$$R^{ou} = (R \cup R') \cap (A^{ou} \times A^{ou})$$

[6] As pointed out by Reviewer 1, modelling information transmission and update in cheap talk situations is a highly interesting venue, which we must leave for future research.

Here the receiver accepts all the new arguments communicated by the sender as well as all the new attacks, provided that both ends are in her updated argument set. In what follows the set of pro (resp. contra and neutral) arguments inherits the apex of the parent argument space, e.g. P^{ou} (resp. C^{ou} and N^{ou}) is the set of the pro (resp. contra and neutral) arguments in A^{ou}.

In most cases, however, information update is more critical than this. Agents often discard evidence that speaks against their prior beliefs, or else devote more scrutiny to it [13]. The latter option is what typically happens when agents try to reduce so-called *cognitive dissonance* [11]. The former is instead a form of what has been called "kripkean dogmatism" [18]. Such update procedures are more articulated. With the next definition we provide an example of one dogmatic policy $\star^d$ of such kind, which is built upon the previously defined $\star^{ou}$.

Definition 11 (Dogmatic update). *Let $X = (A, R), a$, $Y = A', R'$ and let $\star^{ou}(X, Y) = (A^{ou}, R^{ou}), a$ be as in Definition 10. Then the output for agent i of $\star^d(X, Y)$ is a pointed AF $(A^d, R^d), a$ where*

- *If $d(P_i(a), C_i(a)) \geq \frac{1}{2}$*

$$A^d = A^{ou} \setminus (C_i^{ou}(a) \cap (A' \setminus A))$$
$$R^d = R^{ou} \setminus ((C_i^{ou}(a) \times P_i^{ou}(a)) \cap (R' \setminus R)))$$

- *If $d(P_i(a), C_i(a)) < \frac{1}{2}$*

$$A^d = A^{ou} \setminus (P_i^{ou}(a) \cap (A' \setminus A))$$
$$R^d = R^{ou} \setminus ((P_i^{ou}(a) \times C_i^{ou}(a)) \cap (R' \setminus R)))$$

Under the policy $\star^d$ the agent i updates her framework on the basis of her degree of acceptability of a prior to the exchange of information. If she had a positive degree of accepability about the issue a, she discards all new arguments provided by the sender $(A' \setminus A)$ against a, as well as the new attacks against it $((C_i^{ou}(a) \times P_i^{ou}(a)) \cap (R' \setminus R))$. Otherwise she discards the pro arguments and the new attacks from pro to contra.

4 Results

I show two results holding for a two-agent debate between Agent 1 and 2. Agents follow some combinations of the previously defined policies of information transmission and update. All the following results show what happens after one round of mutual information transmission and update.

Theorem 1. *Let $\mathcal{F}, a = (A, R), a$, $I_1(\mathcal{F}, a) = (A_1, R_1), a$ and $I_2(\mathcal{F}, a) = (A_2, R_2), a$. Let $I_1^{ou}(\mathcal{F}, a) = (A_1^{ou}, R_1^{ou}), a = \star^{ou}(I_1(\mathcal{F}, a), T_{2,1}^{o}(I_2(\mathcal{F}, a)))$ and $I_2^{ou}(\mathcal{F}, a) = (A_2^{ou}, R_2^{ou}), a = \star^{ou}(I_2(\mathcal{F}, a), T_{1,2}^{o}(I_1(\mathcal{F}, a)))$. Suppose further that the distributed information of the agents covers the total information available, i.e. (tot) $A_1 \cup A_2 = A$ and $R_1 \cup R_2 = R$. Then*

(a) $I_1^{ou}(\mathcal{F}, a) = I_2^{ou}(\mathcal{F}, a) = \mathcal{F}, a$
(b) $d(P_1^{ou}(a), C_1^{ou}(a)) = d(P_2^{ou}(a), C_2^{ou}(a)) = d(P(a), C(a))$

Proof. (b) is an immediate consequence of (a). (a) is established as follows. By Definition 8 and 10 it follows that $A_1^{ou} = A_1 \cup A_2 = A_2^{ou}$ and $R_1^{ou} = R_1 \cup R_2 = R_2^{ou}$. Then, by the condition (*tot*) it follows that $A_1^{ou} = A_2^{ou} = A$ and $R_1^{ou} = R_2^{ou} = R$ and the result is established.

This result shows that by following open policies of information transmission and update all agents can align their opinion to the most reasonable one. However, we must notice that, for this to happen, (*tot*) is a necessary condition. Indeed, Example 4 shows that if such condition fails, the same policies may lead all agents far away from the most reasonable opinion.

Theorem 2. *Let* $\mathcal{F}, a = (A, R), a$ *with* $C(a) \cap P(a) = \emptyset$. *Let* $I_1(\mathcal{F}, a) = (A_1, R_1), a$. *Let* $I_1^d(\mathcal{F}, a) = (A_1^d, R_1^d), a = \star^d(I_1(\mathcal{F}, a), T_{2,1}^p(I_2(\mathcal{F}, a)))$ *and* $T_{2,1}^p(I_2(\mathcal{F}, a)) = A', R'$. *Furthermore suppose that* $N_1(a) = \emptyset$.

Then

(a) *if* $d(P_1(a), C_1(a)) \geq \frac{1}{2}$

$$d(P_1(a), C_1(a)) \leq d(P_1^d(a), C_1^d(a))$$

(b) *if* $d(P_1(a), C_1(a)) < \frac{1}{2}$

$$d(P_1(a), C_1(a)) \geq d(P_1^d(a), C_1^d(a))$$

Proof. We only prove (a), since (b) follows by the same reasoning. In order to establish (a) it is sufficient to prove that $|C_1(a)^{\leftarrow P_1(a)}| \leq |C_1^d(a)^{\leftarrow P_1^d(a)}|$ and that $|P_1(a)^{\leftarrow C_1(a)}| \geq |P_1^d(a)^{\leftarrow C_1^d(a)}|$ and the result will follow as a consequence of Definition 6. The former is established by ascertaining that $C_1(a)^{\leftarrow P_1(a)} \subseteq C_1^d(a)^{\leftarrow P_1^d(a)}$. The second inequality is established by showing that there are no elements b and c such that $(b, c) \in P_1^d(a)^{\leftarrow C_1^d(a)}$ and $(b, c) \notin P_1(a)^{\leftarrow C_1(a)}$. For reductio we suppose the contrary and then reason by cases.

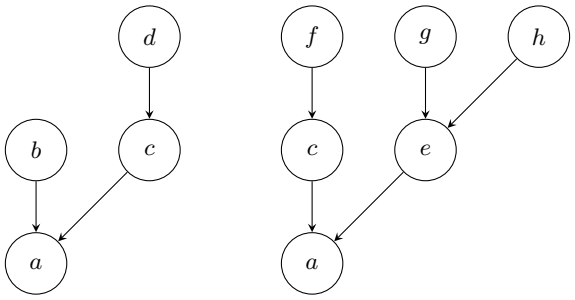
Case 1. $b \in A_1$, $c \in A_1$ and $(b, c) \in R_1$.
If $c \in P_1(a)$ then $b \in C_1(a)$ and $(b, c) \in P_1(a)^{\leftarrow C_1(a)}$ against the assumption. If $c \in C_1(a)$ then a fortiori $c \in C(a)$. However we also assumed that $c \in P_1^d(a)$ and therefore $c \in P(a)$. This however goes against the assumption that $C(a) \cap P(a) = \emptyset$ and we get a contradiction. The last possibility is $c \in N_1(a)$, but this again is excluded by the assumption that $N_1(a) = \emptyset$. Therefore Case 1 leads to a contradiction.

Case 2. Either $b \notin A_1$ or $c \notin A_1$ or $(b, c) \notin R_1$.
Case 2.1. If $b \notin A_1$ then it must be that $b \in (A' \setminus A_1)$. This excludes the fact that $b \in C_1^d(a)$ since this would imply also that $b \in C_1^{ou}(a)$ and b would therefore have been eliminated by the policy $\star^d$.
Case 2.2 If $c \notin A_1$ then $c \in (A' \setminus A_1)$. This entails that $(b, c) \in (R' \setminus R_1)$. The latter however contradicts the fact that $(b, c) \in P_1^d(a)^{\leftarrow C_1^d(a)}$, since a fortiori

$(b,c) \in P_1^{ou}(a)^{\leftarrow C_1^{ou}(a)}$ and therefore (b,c) would have been eliminated by the policy $\star^d$.
Case 2.3 If $(b,c) \notin R_1$ then $(b,c) \in (R' \setminus R_1)$ and a contradiction follows in the same way as in Case 2.2
Therefore Case 2 also leads to a contradiction and the proof is completed.

Under these conditions participants to a debate can only radicalize their prior opinions. This leaves the possibility open for bipolarization to happen. This is what the following example shows.

Example 5. Let the information states of Agent 1 and 2 be as in Figure 4. Then $d(P_1(a), C_1(a)) = \frac{5}{12}$ and $d(P_2(a), C_2(a)) = \frac{13}{24}$. Here $N_1(a) = N_2(a) = \emptyset$ since both graphs are connected. Suppose that both agents adopt the policies $T^o_{i,j}$ and $\star^d$. Then Agent 1 will update his information state with argument e and attack (e,a). Agent 2 will instead update her state by argument d and the attack (d,c). Therefore $d(P_1^d(a), C_1^d(a)) = \frac{3}{8}$ and $d(P_2^d(a), C_2^d(a)) = \frac{19}{30}$ The group has therefore bipolarized.



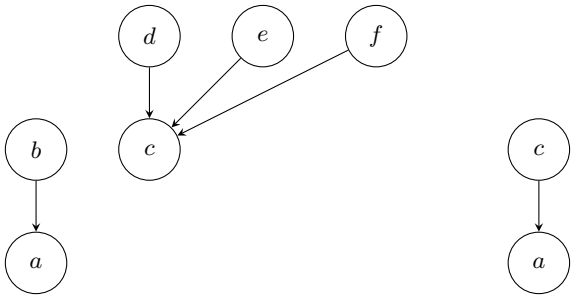
(a) Agent 1 (b) Agent 2

Fig. 4. Bipolarization

Theorem 2 may however not hold when $N_1(a)$ is not empty as the next example shows

Example 6. Let the information states of Agent 1 and 2 be as in Figure 5. Then $d(P_1(a), C_1(a)) = \frac{1}{4}$. Suppose that Agent 2 adopts the policy $T^o_{2,1}$ and Agent 1 adopts $\star^d$. Then Agent 1 will update his information state with the attack (c,a). As a result $d(P_1^d(a), C_1^d(a)) = \frac{13}{24}$. So Agent 1 shifts her opinion towards the opposite direction

This example illustrates how subtly information dynamics can influence opinion change. Indeed, under certain conditions, even providing information contra a may determine a significant positive shift in the degree of acceptability of a. In Example 6 this happens even though the recipient holds a dogmatic attitude in discarding information pro a.



(a) Agent 1 (b) Agent 2

Fig. 5. Opinion shift for Agent 1

5 Conclusions

One of the main aims of this work is to show how Argumentation Frameworks provide an efficient tool for a formal understanding of polarization and bipolarization dynamics in group discussion. Indeed, they serve to encode both the total information available in a debate about a given issue a, as well as the agents' partial information about it. A large number of policies of information exchange among agents can be defined as specific operations on Argumentation Frameworks. Moreover, the measure provided by [24] provides a way to quantify the degree of acceptability of a, both for the single discussants and on the absolute level. Therefore, it also serves to quantify the distance of the agents' opinion from the actual degree of acceptability of a. More importantly, it quantifies polarization and bipolarization prior and posterior to information exchange.

Theorem 1 of Section 4 shows how an open policy of information transmission and information update can help agents to align their views together with the most reasonable opinion about the debated issue. This is often cited as one of the virtues of an open discussion, held by agents without prejudices. Theorem 1 confirms such intuition, but only to a very specific extent. Indeed, the conclusion holds only under the condition (tot) that discussants have distributed knowledge of the total information available. Otherwise the possibility is open for polarization to happen and for the agents to be led far away from the most reasonable opinion, as shown in Example 4.

Theorem 2 shows instead that agents who follow a dogmatic policy of information update are bound to radicalize their opinion. This leaves the way open to polarization and bipolarization, as shown in Example 5. Therefore, this result provides an alternative explanation of bipolarization, one which fully lies in the agent's policy of information update and does not recur to standard explanations such as negative influence, as in [16] and [12], or homophily, as in [23]. Nonetheless, in this case too the conclusion holds only under a specific condition.

Example 6 illustrates how, when such condition fails, opinion shifts may happen even with a dogmatic policy of information update. This example put the finger on the complexity and the unpredictability of the opinion dynamics generated by information exchange. The study of such dynamics therefore discloses an interesting field of investigation for future research.

As a final consideration, Example 4 illustrates how polarization can happen under conditions of "full rationality", i.e. when agents openly disclose their information and update it on the basis of all the available evidence. On the other hand, Example 5 may seem to suggest that bipolarization requires instead that agents deviate from Bayesian standards, e.g. by being dogmatic and discarding some available evidence. However this conclusion jumps too far. Indeed, Theorem 2 only provides sufficient conditions for (bi)polarization, and not necessary ones. The general question remains open as to whether bipolarization entails such a deviation from Bayesianism. Answering this question goes beyond the scope of this paper, although it deserves the highest priority in this research agenda.[7]

Acknowledgements

This research is financed by the *Riksbankens Jubileumsfond. The Swedish Foundation for Humanities and Social Sciences* (RJ) through the project *Rationality and Group Behavior* (P16-0596:1). Special thanks go to Ilaria Jemos and Roberto Ciuni for their insightful comments and suggestions. I am also indebted to Davide Grossi and Paolo Di Paola for discussion and inspiration.

References

1. M. Caminada and C. Sakama. On the Issue of Argumentation and Informedness. *2nd International Workshop on Argument for Agreement and Assurance (AAA 2015)*, 2015.
2. C. Cayrol and M.C. Lagasquie-Schiex. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. *Lecture Notes in Computer Science*, 3571: 378–389, 2005.
3. C. Cayrol and M.C. Lagasquie-Schiex. Bipolarity in Argumentation Graphs: Towards a better Understanding. *International Journal of Approximate Reasoning*, 54(7): 876–899, 2013.
4. S. Coste-Marquis, C. Devred, S. Konieczny, M.C. Lagasquie-Schiex and P. Marquis. On the merging of Dung's argumentation systems. *Artificial Intelligence*, 171: 730–753, 2007.
5. V. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–1451, 1982.
6. J. Delobelle, S. Konieczny and S. Vesic. On the Aggregation of Argumentation Frameworks. *IJCAI 2015*: 2911–2917, 2015.
7. J. Delobelle, A. Haret, S. Konieczny, J. Mailly, J. Rossit. and S. Woltran. Merging of Abstract Argumentation Frameworks. *KR 2016*: 33–42, 2016.

[7] Thanks to Reviewer 1 for raising this issue.

8. F. Dupin de Saint-Cyr, P. Bisquert, C. Cayrol and M.C. Lagasquie-Schiex. Argumentation update in YALLA (Yet Another Logic Language for Argumentation). *International Journal of Approximate Reasoning*, 75: 57–92, 2016.
9. P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77 (2): 321–357, 1995.
10. J. Farrell Cheap talk, coordination, and entry. *The RAND journal of economics* 18(1): 34–39, 1987.
11. L. Festinger. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press, 1957.
12. A. Flache and M.W. Macy Small world and cultural polarization. *Journal of Mathematical Sociology* 35: 146–176, 2011.
13. T. Gilovich. *How we know what isnt so*. The Free Press, New York, 1991.
14. D. Grossi and S. Modgil On the Graded Acceptability of Arguments. *Proceedings of the IJCAI 2015*: 868–874, 2015.
15. D.J. Isenberg. Group Polarization: A critical review and a Meta-Analysis. *Journal of Personality and Social Psychology* 50 (6): 1141–1151, 1986.
16. W. Jager and F. Amblard Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory* 10: 295–303, 2004.
17. A. Jern, K.K. Chang and C. Kemp Belief Polarization is not always irrational. *Psychological Review* 121(2): 206–224, 2014.
18. T. Kelly Disagreement, Dogmatism, and Belief Polarization. *Journal of Philosophy* 105(10): 611–633, 2008.
19. Z. Krizan and R.S. Baron Group polarization and choice-dilemmas: How important is self-categorization? *European Journal of Social Psychology* 37: 191–201, 2007.
20. H. Li, N. Oren and T.J. Norman Probabilistic Argumentation Frameworks *Lecture Notes in Computer Science* 7132: 1–16, 2011.
21. Q. Liu, J. Zhao and X. Wang Multi-agent model of group polarisation with biased assimilation of arguments *Control Theory & Applications, IET 9.3*: 485–492, 2014.
22. C. Lord, L. Ross and M. Lepper Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology* 37 (11): 2098–2109, 1979.
23. M. Mäs and A. Flache Differentiation without Distancing. Explaining Bi-Polarization of Opinions without Negative Influence. *PLoS ONE* 8 (11): e74516, 2013.
24. P. Matt and F. Toni A Game-Theoretic Measure of Argument Strength for Abstract Argumentation. *Lecture Notes in Computer Science* 5293: 285–297, 2008.
25. C. Sakama. Dishonest Arguments in Debate Games. *COMMA 2012*, 75: 177–184, 2012.
26. G.S. Sanders and R.S. Baron Is social comparison irrelevant for producing choice shifts? *Journal of Experimental Social Psychology* 13: 303–314, 1977.
27. J.A. Stoner A comparison of individual and group decision involving risk MA thesis, Massachusetts Institute of Technology, 1961.
28. C. Sunstein. *Why societies need Dissent*. Cambridge, Harvard University Press, 2003.
29. S. Toulmin. *The Uses of Argument*. Cambridge, Harvard University Press, 1958.
30. A. Vinokur and E. Burnstein Effects of partially shared persuasive arguments on group-induced shifts, *Journal of Personality and Social Psychology* 29 (3): 305–15, 1974.