Computational Models of Argument
P. Baroni et al. (Eds.)

doi:10.3233/978-1-61499-686-6-41

Understanding Group Polarization with Bipolar Argumentation Frameworks

Carlo PROIETTI

Lund University

Abstract. Group polarization occurs when an initial attitude or belief of individuals becomes more radical after group discussion. Polarization often leads subgroups towards opposite directions. Since the 1960s this effect has been observed and repeatedly confirmed in lab experiments by social psychologists. Persuasive Arguments Theory (PAT) emerged as the most convincing explanation for this phenomenon. This paper is a first attempt to frame the PAT explanation more formally by means of Bipolar Argumentation Frameworks (BAFs). In particular, I show that polarization may emerge in a BAF by simple and rational belief updates by participants.

Keywords. Group Polarization, Persuasive Arguments Theory, Bipolar Argumentation Frameworks, Value-Based Argumentation Frameworks.

Introduction

Group-induced attitude polarization, also known as *risky shift* ([27]) occurs "when an initial tendency of individual group members toward a given direction is enhanced following group discussion. For example, a group of moderately profeminist women will be more strongly profeminist following group discussion" ([15]). This phenomenon occurs very often in real-life scenarios such as political debate ([28]) or discussion on virtual forums ([29]). Polarization often leads subgroups towards opposite directions, a phenomenon called *bipolarization*. Therefore, it speaks against the assumption that debate among informed individuals should lead to consensus and be truth-conducive. The fundamental question to ask is whether polarization is intrinsically irrational or not. A second question is whether it may happen in situations of perfect communication within a group. Both questions are very complex to disentangle insofar as rationality is a vaguely defined concept. However, formal approaches, as the one I adopt here, provide enough tools to capture the notion of rational update of information and therefore allow asking the question as to whether polarization may happen in situations of rational update by individuals.

Group polarization needs not to be confused with a similar phenomenon, called *belief polarization*.[1] The essential difference lies in the fact that debate and argumentation are essential ingredients of the former but not of the latter.

Large field experiments, mostly conducted in the 1970s, isolated two main concurrent explanations for this phenomenon. The first one builds upon *Social Comparison Theory* and the second upon *Persuasive Arguments Theory* (PAT). 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, people tend to take a position which is similar to everyone else but a bit more extreme. The PAT explanation ([30]) assumes instead 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]).

Both Social Comparison Theory and PAT have inspired multi-agent simulation models of opinion formation meant to explain bipolarization effects. Models inspired by Social Comparison explanations typically assume that agents are positively influenced by their ingroup members and negatively influenced by outgroup members ([12]). Alternatively, some models presuppose that the agents' opinions come closer to opinions of similar degree and instead shift away from opinions of a too different degree ([16]). Models inspired by PAT do not assume negative influence of any kind, but presuppose homophily, i.e. stronger interaction with like-minded individuals ([24]), or biased assimilation of arguments ([22]). Both kind of models can explain bipolarization effects. However, models based on social comparison fall back on a much stronger assumption. Furthermore, empirical research showing the presence of negative influence in social interaction is not immune from criticisms ([19]).

Other than being the most recognized by psychologists nowadays, the PAT explanation is also of main interest for answering our questions. Indeed it posits that polarization may arise by a rational process due to individuals refining their argumentative skills. However, the exact mechanisms of how this process may unfold are still unclear. To understand polarization we need to decompose it into its basic ingredients, i.e. (a) a plurality of agents, (b) a debated issue, (c) possibly different prior opinions held by the agents about the debated issue, (d) *pro* and *contra* arguments – possibly related with each other by relations of refutation, support, counterattack etc. – and (e) update, by the agents, of their argumentative basis.

All such ingredients can be formally framed by the help of Argumentation Frameworks ([10]), more specifically via *Bipolar Argumentation Frameworks* (BAFs) introduced by [3]. A BAF consists of a graph where nodes are arguments and directed links represent either *supports* or *attacks* among them. A specific BAF is originally meant to represent a completed process of argumentation, i.e. the situation where "everything is on the table". Here we give BAFs a dynamic

[1] Belief polarization ([23]) happens when two parties are lead to more extreme disagreement after considering the same evidence. Formal approaches based on *Bayesian networks* have already shown that this phenomenon needs not to be irrational ([17]).

turn in order to understand the steps of an argumentative debate among n agents. Indeed, given a BAF A, the information available to the participants to a debate can be represented as a subgraph of A. The result of a debate/exchange of arguments between two agents j and k can be framed as an operation on their respective subgraphs. It is very easy to show, even in this purely qualitative framework, that polarization may easily emerge throughout a debate.

I proceed as follows. Section 1 reviews the structure of some lab experiments meant to show the emergence of group polarization and to test the PAT explanation. Section 2 introduces BAFs and shows how to frame a debate and argumentative update in a group of n agents. It is shown how polarization towards opposite directions can arise due to incomplete communication in a group. Section 3 shows that polarization can also emerge in situations of full communication due to individual biases. Section 4 concludes by presenting some further research questions that can be answered by appeal to Argumentation Frameworks.

1. Group polarization in the lab

Many experimental studies have been conducted to show that persuasive and novel arguments can induce polarization ([30]). Such experiments have a more or less standard structure. Test subjects are presented with a binary choice between two options A and B, where A is a low-risk low-gain option and B is high-risk high-gain. Test subjects should provide their initial odds for switching from A to B.[2] Subjects are also asked to write down arguments *pro* and *contra* the decision of switching from A to B. Arguments are then circulated among the participants who should rank them on the basis of their *persuasiveness* and *novelty*. Participants are then asked again to give their odds for switching from A to B. The difference between the (average value of the) prior odds and the (average value of the) posterior odds gives the measure of polarization towards A or B. The same test is repeated over different pairs A and B: some pairs typically show polarization toward A while others toward B.

Experimental results established some important correlations:

(a) Prior to group discussion there exists a *culturally given pool of arguments* that determines the initial propensity of individuals towards A or towards B.

(b) The number and persuasiveness of the arguments pro (contra) are strongly correlated with the initial choice of odds in one direction or the other.

(c) Sharing of arguments is a necessary condition for polarization.

(d) Persuasiveness and novelty of the shared arguments pro or contra are strongly correlated with polarization in one direction or the other.

(e) Actual face to face debate among subjects does not increase polarization

Points (a) to (d) provide evidence in favor of PAT, while point (e) speaks against the social comparison explanation.

[2]Typically, test subjects should rate in a 1 to 10 scale how inclined they are to switch from A to B.

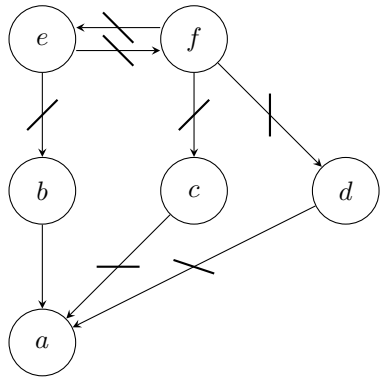


Figure 1. An example of $\mathcal{BAF}$. Labelled nodes represent arguments. Relations of support between arguments are indicated with a plain edge, while relations of attack are indicated with a barred one.

Pro and contra arguments play an essential role in this picture[3] and the experiments thus far presented are quite convincing. However, to fully understand the impact and the role of arguments in polarization we need a more fine-grained picture. As a first important point, it is simplistic to categorize argumentative moves in a debate simply as pro or contra. Arguments in a debate usually form a complex network, e.g. some argument x undermines y which in turn supports z (and therefore x also undermines z). To better estimate the impact of an argument in a debate we should then assess its impact on the overall network, and this is something that Argumentation Frameworks allow us to do. Secondly, we need to represent the network dynamics as generated by debate. We shall deal with both these issues in the next section.

2. Bipolar Argumentatiion Frameworks

Bipolar Argumentation Frameworks [3] are defined as follows.

Definition 1 (BAF) *A Bipolar Argumentation Framework $\mathcal{BAF}$ is a triple $(\mathcal{A}, \mathcal{R}^a, \mathcal{R}^s)$ where $\mathcal{A}$ is a finite and non-empty set of arguments and $\mathcal{R}^a, \mathcal{R}^s \subseteq \mathcal{A} \times \mathcal{A}$*

Here $\mathcal{R}^a$ and $\mathcal{R}^s$ are binary relations over $\mathcal{A}$, called the *attack* and the *support* relation. $a\mathcal{R}^a b$ means argument a attacks argument b, while $a\mathcal{R}^s b$ means a supports b. An example of a BAF is provided by Figure 1. Relations of support between arguments are represented by a plain directed edge, while relations of attack by a barred one. Here, for example, argument a receives support from b which, in its turn is attacked by e. In an intuitive sense, a is therefore indirectly attacked by e, which undermines one of its supports. Therefore, with respect to Dung's original

[3]It is important to stress that in such context, as in everyday discussions, 'pro' and 'contra' are quite independent notions. No specific constraint is given such as ,e.g., an argument pro A is an argument contra not-A. Therefore, in a formal context, we need to represent pro and contra as two independent binary relations among arguments.

framework we have more complex types of attack than the simple $\mathcal{R}^a$. They fall under two general categorizations, provided by the following definition (see [4]).

Definition 2 (Complex attacks) **(i)** *There is a supported attack from a to b if there is a sequence* $a_1\mathcal{R}_1 \ldots \mathcal{R}_{n-1}a_n$, $n \geq 3$, *with* $a_1 = a$, $a_n = b$, $\forall i = 1, \ldots, n-2$ $\mathcal{R}_i = \mathcal{R}^s$ *and* $\mathcal{R}_{n-1} = \mathcal{R}^a$.

(ii) *There is a secondary attack from a to b if there is a sequence* $a_1\mathcal{R}_1 \ldots \mathcal{R}_{n-1}a_n$, $n \geq 3$, *with* $a_1 = a$, $a_n = b$, $\forall i = 2, \ldots, n-1$ $\mathcal{R}_i = \mathcal{R}^s$ *and* $\mathcal{R}_1 = \mathcal{R}^a$.

In other words, a supported attack consists of an attack preceded by a chain of supports, while a secondary attack is a simple attack followed by a chain of supports. We shall use the term 'attack' to indicate both simple and complex attacks.

Given a particular $\mathcal{BAF} = (\mathcal{A}, \mathcal{R}^a, \mathcal{R}^s)$, its generating set $\mathcal{A}$ is meant to represent an argumentative pool (the "culturally given pool of arguments" from Section 2). A debated *issue* can therefore be regarded as a specific subset of $\mathcal{A}$; in our examples we shall use the singleton set $\{a\}$ as our debated issue.

In this framework, the acceptability of an argument depends on its membership of some sets, usually called *solutions* (or *extensions*). Solutions should have some specific properties. The basic ones among them are *conflict-freeness* and *collective defense* of their own arguments. Intuitively, conflict-freeness means that a set of arguments is coherent, in the sense that no argument attacks another in the same set.[4]

Definition 3 (Conflict-freeness) *A set S is conflict-free if there is no* $a, b \in S$ *s.t. a attacks b.*

The largest conflict-free sets in $\mathcal{BAF}$ of Figure 1 are $\{a, b, f\}$ and $\{c, d, e\}$. A solution should also be able to defend its arguments against external attacks. Such feature is provided by the definition of collective defense.

Definition 4 (Collective defense) *A set S defends collectively an argument a if for all b such that b attacks a there is a* $c \in S$ *s.t. c attacks b.*

These two notions are the basis of most of the solution concepts in the standard Dung's framework (admissibility, preferredness, stability and groundedness). Related solution concepts for BAF have been worked out by [3] and [4]. For our present purposes we need only to introduce the basic notion of d-admissibility (see [3]).[5]

Definition 5 (d-admissibility) *Let* $S \subseteq A$. *S is d-admissible iff S is conflict-free and defends all its elements*

We can see from our example of Figure 1 that two maximal different solutions are admissible: the sets $\{a, b, f\}$ and $\{c, d, e\}$. Argument a belongs to the first but

[4] A stronger notion of coherence is also provided in [3] under the name of 'safety'. However, we only need to introduce conflict-freeness for our present purposes.

[5] The letter 'd' stands for Dung. Indeed, two other notions of admissibility, c-admissibility and s-admissibility, are introduced in [3].

not to the second. Indeed the two sets represent quite opposite positions. If we see our example as the final stage of a debate, participants are in a difficult stand: they have to decide which solution to accept, and such solutions are opposite. However, there are many preliminary steps in a debate where polarization may emerge and participants can be pushed in one direction or the other. Our task for the next Section is precisely to clarify this process.

2.1. The dynamics of a debate

If we regard our $\mathcal{BAF}$ of Figure 1 as the final stage of a debate, then the cognitive state of someone entering the debate should be seen as a partial representation of such BAF: an individual may not be aware of some arguments on the table. She may also not be aware that some argument attacks or supports another. She may even have different opinions and think that some argument attacks another while this is not the case. If we rule out the latter option – which is reasonable to do in our context – then the state of an individual entering a debate is best represented as a *subgraph* of the larger $\mathcal{BAF}$.[6] By consequence, the initial setup of a debate among n agents can be encoded as a multiagent scenario where agents' states are represented by a subgraph of a given BAF. This gives rise to the following definition.

Definition 6 (Multiagent scenario) *Given $\mathcal{BAF}$, a multiagent scenario is a vector $(\mathcal{BAF}_1, \ldots, \mathcal{BAF}_n)$ of BAFs where each $\mathcal{BAF}_i$ (for $1 \leq i \leq n$) is a subgraph of $\mathcal{BAF}$*

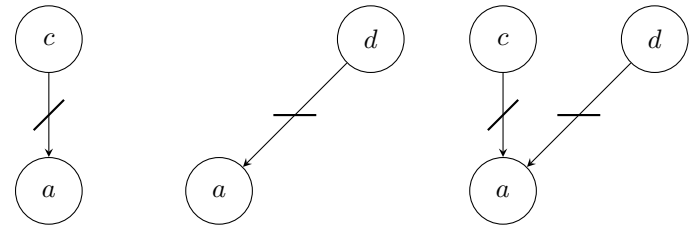
Once a multiagent scenario is set we need to model the successive steps of information exchange in a debate. There are many ways agents could merge new information when such information disagrees with the information they have (see [6]). All of the known merging procedures have some problematic aspect and none of them satisfies all the intuitive properties of an aggregation process (see [7] and [8]). However, in our scenario there is no disagreement possible on whether an arguments attacks or supports another argument. When the situation is such, an argumentative update after an exchange among n agents is modelled simply as the *union* of the participants' respective graphs.

Definition 7 (Argumentative update) *Given a vector $(\mathcal{BAF}_1, \ldots, \mathcal{BAF}_n)$ of BAFs we define, for each i, the update after information exchange as $\mathcal{BAF}_i^* = (\bigcup_{i=1}^n \mathcal{A}, \bigcup_{i=1}^n \mathcal{R}_i^a, \bigcup_{i=1}^n \mathcal{R}_i^s)$*

It is very easy to see, even in this purely qualitative framework, that polarization may easily emerge through debate. Consider a simple example of an exchange on a specific issue a with two agents 1 and 2 where both have arguments against a. Suppose that their respective initial states are represented by $\mathcal{BAF}_1 = (\{a, c\}, \{(c, a)\}, \emptyset)$ (Figure 2a) and $\mathcal{BAF}_2 = (\{a, d\}, \{(d, a)\}, \emptyset)$ (Figure

[6] Analogous approaches have been extensively developed by [25], [2] and [9] to encode multiagent debate dynamics with argumentations systems. Here too the knowledge base of an agent is encoded by a BAF. The agent's knowledge base is a subset of a larger *universe* ([9]) or *universal argumentation framework* ([25] and [2]) whose role is analogous to our argumentative pool.

2b). $\mathcal{BAF}_1 \cup \mathcal{BAF}_2$ is clearly $(\{a, c, d\}, \{(c, a), (d, a)\}, \emptyset)$. Both 1 and 2 have a new arguments against a. In other words, both get more radical and, therefore, the group "shifts" in the direction against a. This dynamic is typically called an *echo chamber*: people become more radical than their original position because they share information with other people who have similar views. Needless, to say, an echo chamber may lead the group towards the opposite direction as well. This happens when people with arguments in favor of a discuss together.



(a) agent 1 (b) agent 2 (c) Updated graph

Figure 2. Argumentative update for agents 1 and 2

A typically suggested policy to prevent echo chambers is to diversify opinions by favoring the interaction of people with different priors.[7] Back to our example, it is easy to see the effect of such mixing if we add a third agent with an argument in support of a to our debate at its initial state. Suppose indeed that agent 3's initial state is $\mathcal{BAF}_3 = (\{a, b\}, \emptyset, \{(b, a)\})$. Then the argumentative update will be as in Figure 3. Here the echo chamber effect is prevented. Indeed, both arguments *pro*

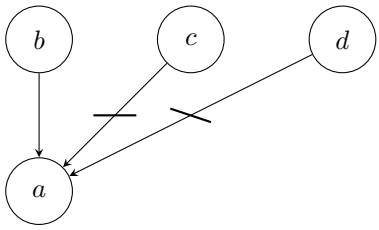


Figure 3. Argumentative update in a mixed group

and *contra* a are available to everybody, the debate is closer to a "tie" and there is no straightforward solution to choose at this stage. However this is not the end of the story. There is much psychological evidence to the fact that polarization can happen at this point too. Indeed people polarize towards opposite directions also in situations of high connectivity, e.g. online political debates ([28]). To see how this is possible we shall incorporate in our framework two explanatory clues provided by social psychology and legal reasoning.

[7] For example, larger representation of minorities in panels and decision committees goes in this direction.

3. Psychological processes and values

For purposes of decision making, agents 1, 2 and 3 in our example often need to break the tie and decide which arguments to save as more relevant when contrasting information is available. There are at least two ways this is done in real life scenarios as we shall see in this section.

3.1. Cognitive dissonance

The presence of inconsistent information usually makes individuals unconfortable and motivates them to reduce so-called *cognitive dissonance* ([11]). This can happen in different manners. People may avoid information which would likely increase the dissonance. They may also discard evidence against their prior beliefs. Or else, they may devote more scrutiny to hypotheses and explanations that speak against their prior beliefs [13].

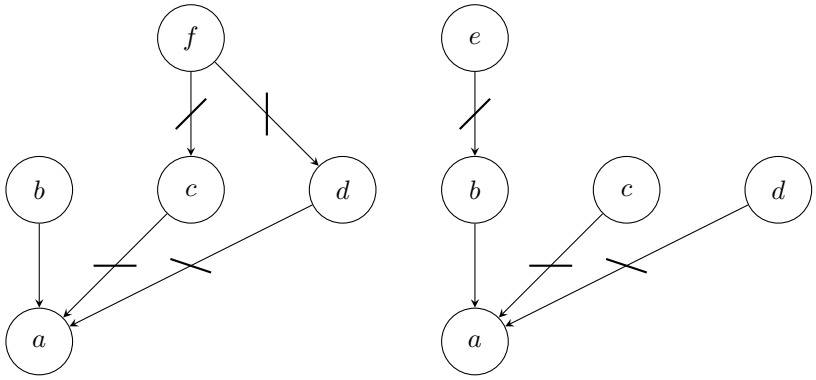
The third possibility seems to explain belief polarization without necessarily assuming that individuals are irrational ([18]). We can easily explain how this works in our framework by reference to our example. Agents 1, 2 and 3 are all in the same state after their first argumentative update (Figure 3). However, their initial state was quite different: agents 1 and 2 had evidence against a, while agent 3 had evidence in favor of a. In addition to that, more arguments are potentially available in their pool, such as e and f in Figure 1. Agent 1 reached her present state by receiving arguments b and d as new information. In an intuitive sense b speaks against her prior beliefs, while d does not. What should then happen when agent 1 scrutinizes b more closely? Intuitively, she should be more likely to find out arguments that undermine b if any. But argument e attacks b in our pool $\mathcal{A}$ of arguments. It may therefore be likely that agent 1 ends up as in Figure 4(b). Here admissible sets are $\{c, d, e\}$ and its subsets and all of them (directly or indirectly) attack a.

On the other hand agent 3 reached her present state by incorporating arguments c and d, which both go against her prior belief. Therefore, she is likely to find out arguments that undermine c and d, if any. Such an argument is f. It is therefore likely that agent 3 ends up as in Figure 5(a),where admissible sets are $\{a, b, f\}$, $\{b, f\}$, $\{b\}$ and $\{f\}$. Such sets contain only arguments supporting a. Agent 1 and 3 will therefore disagree and polarization is back again.

Such a way of updating takes into account not only the agent's present state but also the previous ones. This, of course, is not the full story of how agents may scrutinize new evidence that contrasts with their prior beliefs, but it tells us that polarization may be very resilient and difficult to contrast even when people with different priors interact in an open and large debate.

3.2. Values

In cases like the one represented in Figure 1 a dispute cannot be settled. Indeed, there are two maximal disjoint admissible solutions for the graph:$\{c, d, e\}$ and $\{a, b, f\}$. As stressed by [1], this is often the case in contexts of practical reasoning, law or ethical debate, which are also contexts where polarization often arises. In many cases the dispute is solved by appeal to the arguments' intrinsic value.



(a) Solving cognitive dissonance one way

(b) or another

Figure 4. How subjects may solve cognitive dissonance

> Often no conclusive demonstration of the rightness of one side is possible: both sides will plead their case, presenting arguments for their view as to what is correct. Their arguments may all be sound. But their arguments will not have equal value for the judge charged with deciding the case: the case will be decided by the judge preferring one argument over the other. And when the judge decides the case, the verdict must be supplemented by an argument, intended to convince the parties to the case, fellow judges and the public at large, that the favoured argument is the one that should be favoured. ([1], p.429-430)

Arguments are very often attached with *values* in public debate too. As an example, an argument against gun-control may be often associated with *individual freedom*, while arguments for gun-control have a special inclination towards *non-violence*.[8] Indeed, different groups of people may hold different value rankings. This is an explanatory clue for polarization in many contexts. To make this point we need to define Value-based Bipolar Argumentation Frameworks (VBAF). This is done by expanding Bench-Capon's definition in [1].

Definition 8 (VBAF) *A Value-based Bipolar Argumentation Framework $\mathcal{VBAF}$ is a tuple $(\mathcal{A}, \mathcal{R}^a, \mathcal{R}^s, V, val, P)$ where $\mathcal{A}$, $\mathcal{R}^a$ and $\mathcal{R}^s$ are as before, V is a set of values, val is an assignment $\mathcal{A} \longrightarrow V$ and P is a set of "possible audiences" where $p \in P$ is a ranking on V*

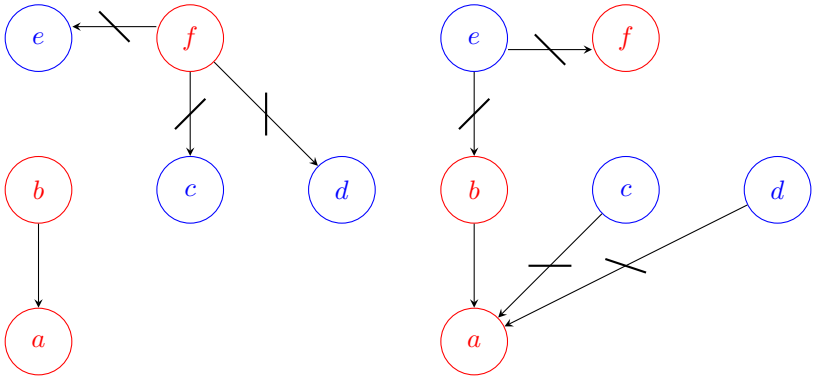
Given a set V of values (e.g. freedom, non-violence etc.), arguments are associated to them by means of the function val. A possible audience p represents the specific ranking an individual or a group assigns to such values. Relative to a specific audience an argument a can properly attack or support b only when the value of a is greater or equal to the value of b. More formally, the following definition applies.

[8]This doesn't mean that arguments for different sides are always associated with different values. Quite often, indeed, to make a "good" move in a debate is to attack the opposite side with an argument who has value for the other side.

Definition 9 (Strong attack and strong support) *For all $a, b \in \mathcal{A}$ and $p \in P$*

(i) *a* strongly attacks *b for audience p iff $a\mathcal{R}^a b$ and not $val(a) <_p val(b)$*
(ii) *a* strongly supports *b for audience p iff $a\mathcal{R}^s b$ and not $val(a) <_p val(b)$*

Going back to our main example in Figure 1, we can easily show how this may generate polarization of two different audiences. Suppose we have only two values, which we label by two different colors, e.g. red and blue. Our V is then $\{red, blue\}$. We also suppose that $val = \{(a, red), (b, red), (c, blue), (d, blue), (e, blue), (f, red)\}$. Finally, we assume that agents belong to two audiences p_1 and p_2 where $blue <_{p_1} red$ and $red <_{p_2} blue$.



(a) Audience p_1 prefers red values (b) Audience p_2 prefers red values

Figure 5. Two values

When the situation at the final stage of the debate is as in Figure 1 the two audiences may adopt two different solutions based on their value rankings. Audience p_1 will come out to the $\mathcal{VBAF}$ represented in Figure 5(a) while audience p_2 will converge to the one represented in Figure 5(b). For p_1 a belongs to an admissible solution while this is clearly not the case for p_2.

All in all, there are many possible explanatory clues for group polarization. BAFs and their dynamics are an adequate tool for capturing most of them.

4. Conclusions and future work

Group polarization is a very complex phenomenon and this paper constitutes only an initial stage of a formal research on this problem. Our main aim was to show that bipolar argumentation frameworks are an adequate tool for framing the steps of a polarization process. We have shown that in some simple scenarios polarization may be captured at a very intuitive level by a simple process of argumentative update. However much work in many directions is left to do in future research. First, we have left out all the quantitative aspects which are a fundamental ingredient of group polarization. Indeed, polarization of attitudes means that argumentative updates induce an increase of the likelihood that individuals

will settle an issue in one way or another. A measure of such likelihood is therefore needed. *Probabilistic Argumentation Frameworks* [21] and *Graded Semantics* [14] are a useful tool for providing such measures and to investigate how likelihood is influenced by argumentative dynamics. Further insights for implementation can be provided by *Social Argumentation Frameworks* [20] and [5]. Such structures are an extension of Argumentation Frameworks meant to model and assess online debates, where *pro* and *contra* votes are associated to arguments. As a most interesting aspect, [20] provides a fine-grained semantics to compute one arguments strength as a function of the structure of the graph and the social opinion expressed through the votes.

Argumentative dynamics are a second main field of inquiry to understand polarization. In our examples, we adopted union of graphs as a straightforward policy of argumentative update. However, as stressed in Section 2, this only works under specific conditions. It won't work in more complex situations where participants receive information which is inconsistent with their prior belief state. To handle such situations more complex operations of graph merging are needed, which are provided by [6],[7] and [8].

References

[1] T.J. Bench-Capon. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3): 430–448, 2003.

[2] M. Caminada and C. Sakama. On the Issue of Argumentation and Informedness. *2nd International Workshop on Argument for Agreement and Assurance (AAA 2015)*, 2015.

[3] 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.

[4] 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.

[5] M. Correia, J. Cruz and J. Leite On the Efficient Implementation of Social Abstract Argumentation. *ECAI 2014*: 225–230, 2014.

[6] 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.

[7] J. Delobelle, S. Konieczny and S. Vesic. On the Aggregation of Argumentation Frameworks. *IJCAI 2015*: 2911–2917, 2015.

[8] J. Delobelle, A. Haret, S. Konieczny, J. Mailly, J. Rossit. and S. Woltran. Merging of Abstract Argumentation Frameworks. *KR 2016*: 33–42, 2016.

[9] 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.

[10] 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.

[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] J. Leite and J. Martins Social Abstract Argumentation. *IJCAI 2011*: 2287–2292, 2011.
[21] H. Li, N. Oren and T.J. Norman Probabilistic Argumentation Frameworks *Lecture Notes in Computer Science* 7132: 1–16, 2011.
[22] 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.
[23] 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.
[24] M. Mäs and A. Flache Differentiation without Distancing. Explaining Bi-Polarization of Opinions without Negative Influence. *PLoS ONE* 8 (11): e74516, 2013.
[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. Yardi, D. Boyd, Dynamic Debates: An analysis of group polarization over time on Twitter *Bulletin of Science, Technology and Society* 30 (5), pp. 316–27, 2010.
[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.