# Polarization and Bipolar Probabilistic Argumentation Frameworks

Carlo Proietti

Lund University

**Abstract.** Discussion among individuals about a given issue often induces *polarization* and *bipolarization* effects, i.e. individuals radicalize their initial opinion towards either the same or opposite directions. Experimental psychologists have put forward Persuasive Arguments Theory (PAT) as a clue for explaining polarization. PAT claims that adding novel and persuasive arguments *pro* or *contra* the debated issue is the major cause for polarization. Recent developments in abstract argumentation provide the tools for capturing these intuitions on a formal basis. Here Bipolar Argumentation Frameworks (BAF) are employed as a tool for encoding the information of agents in a debate relative to a given issue $a$. A probabilistic extension of BAF allows to encode the likelihood of the opinions pro or contra $a$ before and after information exchange. It is shown, by a straightforward example, how these measures provide the basis to capture the intuitions of PAT.

## 1 Introduction

The work by [9] unveiled the phenomenon of *risky shift*, i.e. the tendency of a group to take decisions that are more extreme than the average of the individual decisions of members before the group met. Subsequent research in social psychology showed that a similar pattern applies, more generally, to change of attitude and opinion after debate. This phenomenon goes under the name of *Group polarization*. A side effect of polarization are *bipolarization effects*, i.e. the tendency of different subgroups to radicalize their opinions towards opposite directions. One explanation of polarization is provided by *Persuasive Arguments Theory*, PAT for short, developed by social psychologists in the 1970s [12]. PAT assumes that individuals become more convinced of their view when they hear novel and persuasive arguments in favor of their position.[1]

[1] PAT is not the only explanation for polarization. A relevant tradition in social psychology has regarded polarization and bipolarization as a byproduct of social comparison [8]: to present themselves in a favorable light in their social environment, individuals take a position which is similar to everyone else but a bit more extreme. Both social comparison and PAT seem to play an important role in different polarization contexts. It is beyond our purposes to adjudicate between the two. We only remark that PAT is fundamental in many such scenarios. To stress this point, if social comparison was the only decisive factor, we should expect to find more pronounced polarizing effects in face-to-face discussion rather than, e.g., web-mediated exchange. But there is conflicting evidence thereof.

Some recently developed tools in Dung-style abstract argumentation [3] allow to capture this phenomenon on a formal ground. Indeed, Bipolar Argumentation Frameworks (BAF) ([1, 2]) allow to encode the argumentative knlowledge base of an agent and the arguments both in favor and against a debated issue $a$. Moreover, a probabilistic extension of BAF allows to encode the likelihood with which opinions pro and contra the debated issue are entertained by the agent. The probabilistic extension of BAF will be named PrBAF and is defined along the same lines of [5] and [4]. In Section 4 PrBAF are employed to validate PAT. There it is shown, by an example, how the exchange of arguments in a debate modifies the likelihoods of pro and contra opinions. Furthermore, it is indicated how polarization and bipolarization can be encoded in a way that tracks the change of likelihood of pro and contra opinions.

## 2 Bipolar Argumentation Frameworks

Bipolar Argumentation Frameworks (BAF for short) [1] are defined as follows

**Definition 1 (BAF).** *A Bipolar Argumentation Framework $G$ 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 $b$.

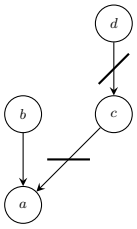


**Fig. 1.** An example of BAF. Labelled nodes represent arguments. Relations of support between arguments are indicated with a plain edge, while relations of attack are indicated with a barred one.

The following definition provides some important relations between arguments in a BAF.

**Definition 2 (Defeat, defence, support and neutrality).**

**(i)** *$a$ defeats $b$ if there is a path $a\mathcal{R}_1 \dots \mathcal{R}_n b$, $\mathcal{R}_i \in \{\mathcal{R}_s, \mathcal{R}_a\}$ $\forall i \leq n$, $n \geq 1$ such that the path contains an odd number of attacks.*

**(ii)** *$a$ defends $b$ if there is a path $a\mathcal{R}_1 \dots \mathcal{R}_n b$, $\mathcal{R}_i \in \{\mathcal{R}_s, \mathcal{R}_a\}$ $\forall i \leq n$, $n \geq 1$ such that the path contains an even number $k$ of attacks with $k > 0$.*

**(iii)** *a supports (resp. weakly supports) b if there is a path* $a\mathcal{R}_1 \ldots \mathcal{R}_n b$, $\mathcal{R}_i = \mathcal{R}_s$ $\forall i \leq n$, *such that* $n \geq 1$ *(resp.* $n \geq 0$*).*

**(iv)** *a is neutral w.r.t. b if there is no path (including paths of length 0) between a and b.*

The BAF of Figure 1 helps to illustrate these notions. Argument $b$ supports $a$, while $c$ defeats $a$ and $d$ defends $a$.

**Fact 1.** *Provided that there is a unique path from a to b, we have either that a defeats b, or that a defends b, or that a (weakly) supports b, and these options are mutually exclusive.*

*Opinions* held by the participants are represented as sets of arguments. The opinions we are interested in are those which are either *pro* or *contra* a given issue $a$. Based on Definition 2 we identify the positions pro $a$, $\mathcal{P}(a)$, and contra $a$, $\mathcal{C}(a)$ as the following families of sets.

**Definition 3 (Pro and contra opinions).**

- *a set* $S$ *belongs to* $\mathcal{P}(a)$ *iff* $S \neq \emptyset$ *and* $\forall b \in S$ *either b defends a or b weakly supports a.*
- *a set* $S$ *belongs to* $\mathcal{C}(a)$ *iff* $S \neq \emptyset$ *and* $\forall b \in S$ *b defeats a.*

In this framework, the acceptability of an argument depends on its membership to some set, usually called *extensions* (or *solutions*). 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.[2] 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.

**Definition 4 (Properties of extensions).**

**(i)** *A set* $S$ *is conflict-free if there is no* $a, b \in S$ *s.t. a defeats b.*

**(ii)** *A set* $S$ *defends collectively an argument a if for all b such that b defeats a there is a* $c \in S$ *s.t. c defeats b.*

**(iii)** *A set* $S$ *is admissible if it is conflict-free and defends all its elements.*

Admissibility is a key to define the likelihood measures of $\mathcal{P}(a)$ and $\mathcal{C}(a)$ in the next section.

## 3 Probabilistic BAFs

One Argument $b$ may be more or less persuasive in support or against a debated issue $a$. Persuasivity can be factorized into two main components: (a) the intrinsic

[2] A stronger notion of coherence is also provided in [1] under the name of 'safety'. However, we only need to introduce conflict-freeness for our present purposes.

strength of $b$ which supports or defeats $a$ and (b) the strength of the link between $b$ and $a$. Both these components are taken into account by the approach of [5] and can be straightforwardly extended to the bipolar case. For simplicity, and without any loss for our point, we restrict our attention to component (a) as done by [4].

**Definition 5 (PrBAF).** *A Probabilistic Bipolar Argumentation Framework $G$ is a tuple $(\mathcal{A}, \mathcal{R}_a, \mathcal{R}_s, p)$ where $(\mathcal{A}, \mathcal{R}_a, \mathcal{R}_s)$ is a BAF and $p : \mathcal{A} \to [0, 1]$ is a probability function over arguments.*

$G' = (\mathcal{A}', \mathcal{R}_a', \mathcal{R}_s')$ is a **spanning subgraph** of $G = (\mathcal{A}, \mathcal{R}_a, \mathcal{R}_s, p)$, denoted as $G' \sqsubseteq G$, if $\mathcal{A}' \subseteq \mathcal{A}$, $\mathcal{R}_a' = \mathcal{R}_a \otimes \mathcal{A}'$ and $\mathcal{R}_s' = \mathcal{R}_s \otimes \mathcal{A}'$, where $\mathcal{R} \otimes \mathcal{A} = \{(a, b) \in \mathcal{R} \mid a, b \in \mathcal{A}\}$.

**Definition 6.** *Let $G = (\mathcal{A}, \mathcal{R}_a, \mathcal{R}_s, p)$ be a PrBAF and $G' \sqsubseteq G$. The probability of $G'$, denoted $p(G')$ is*

$$p(G') = (\prod_{a \in \mathcal{A}'} p(a))(\prod_{a \in \mathcal{A} \setminus \mathcal{A}'} (1 - p(a)))$$

It is a standard result that $\sum_{G' \sqsubseteq G} p(G') = 1$ and therefore $p$ defines a probability assignment over the set of spanning subgraphs.

Let $G$ be a PrBAF, $G' \sqsubseteq G$, and $S \subseteq \mathcal{A}$ a set of arguments. $G' \models S$ denotes that $S$ is admissible in $G'$, for short $G'$ *entails* $S$ (see [4]). We can now provide the measures of the likelihood of the pro and contra opinions relative to $a$.

**Definition 7 (Measures of likelihood).**

1. $p(\mathcal{P}(a)) = \sum_{G' \sqsubseteq G} p(G')$ *such that* $G' \models S$ *and* $S \in \mathcal{P}(a)$
2. $p(\mathcal{C}(a)) = \sum_{G' \sqsubseteq G} p(G')$ *such that* $G' \models S$ *and* $S \in \mathcal{C}(a)$

## 4 Change of likelihood, polarization and bipolarization

Figure 2 serves to illustrate the change of likelihood of $\mathcal{P}(a)$ and $\mathcal{C}(a)$ from an initial knowledge base of an agent (Figure 2(a)) where no information is available except for the debated issue $a$. At the subsequent stages arguments are added with different configurations (Figure 2(b), Figure 2(c) and Figure 2(d)).

In Figure 2(a) the spanning subgraphs $G' \sqsubseteq G$ are $G_1 = (\{a\}, \emptyset, \emptyset)$ and $G_2 = (\emptyset, \emptyset, \emptyset)$, and both have probability 0.5. Here $G_1 \models \{a\}$, and $\{a\} \in \mathcal{P}(a)$. On the contrary $G_2$ entails nothing else than the empty set $\emptyset$, which does not belong to $\mathcal{P}(a)$. Therefore $p(\mathcal{P}(a)) = 0.5$. None of $G_1$ and $G_2$ instead entails any set in $\mathcal{C}(a)$ and therefore $p(\mathcal{C}(a)) = 0$.

When we add $b$ to the debate the argumentative knowledge base transforms into that of Figure 2(b), where the spanning subgraphs are $G_1 = (\{a\}, \emptyset, \emptyset)$, $G_2 = (\emptyset, \emptyset, \emptyset)$, $G_3 = (\{b\}, \emptyset, \emptyset)$, and $G_4 = (\{b, a\}, \emptyset, \{(b, a)\})$. Here $p(G_1) =$

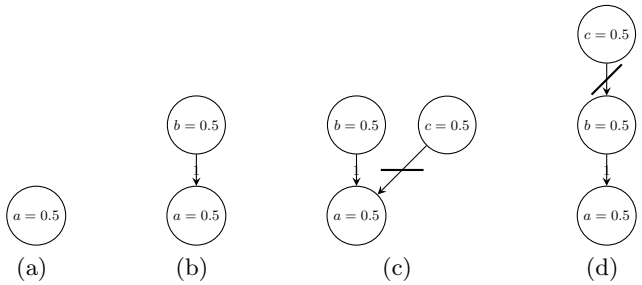


**Fig. 2.** Change of likelihood for pro and contra after new arguments are added.

$p(G_2) = p(G_3) = p(G_4) = 0.25$. Since all of $G_1$, $G_3$ and $G_4$ entail a set in $\mathcal{P}(a)$, it follows, by the same reasoning as before, that $p(\mathcal{P}(a)) = 0.75$, while we still have $p(\mathcal{C}(a)) = 0$. Therefore, our measures capture the intuition that adding one argument in favor of $a$ augments the likelihood of the opinion pro $a$.

By adding $c$, as in Figure 2(c), we obtain eight spanning subgraphs $G_1 = (\{a\}, \emptyset, \emptyset)$, $G_2 = (\emptyset, \emptyset, \emptyset)$, $G_3 = (\{b\}, \emptyset, \emptyset)$, $G_4 = (\{b, a\}, \emptyset, \{(b, a)\})$, $G_5 = (\{c\}, \emptyset, \emptyset)$, $G_6 = (\{b, c\}, \emptyset, \emptyset)$, $G_7 = (\{c, a\}, \{(c, a)\}, \emptyset)$ and $G_8 = (\{b, c, a\}, \{(c, a)\}, \{(b, a)\})$. All of them have probability 0.125. It can be seen that $G_1, G_3, G_4, G_6$ and $G_8$ entail some set in $\mathcal{P}(a)$ and therefore $p(\mathcal{P}(a)) = 0.625$. By the same process we see that $G_5, G_6, G_7$ and $G_8$ entail some set in $\mathcal{C}(a)$ and therefore $p(\mathcal{C}(a)) = 0.5$. Here we see that adding an argument against $a$ not only increases the likelihood of $\mathcal{C}(a)$, but also decreases that of $\mathcal{P}(a)$ and therefore acts as a balance.

Most interestingly, we can see that the configuration of the arguments in the graph influences strongly the likelihoods of $\mathcal{P}(a)$ and $\mathcal{C}(a)$. Indeed if we add the defeater $c$ as in Figure 2(d), its impact changes w.r.t. the case of Figure 2(c). There are always eight spanning subgraphs, but now it can be checked that $p(\mathcal{P}(a)) = 0.375$ and $p(\mathcal{C}(a)) = 0.5$. This helps to appreciate how different argumentative strategies (attack of a supporter vs. direct attack) may have a strong impact on opinion formation. Although the likelihoods of $\mathcal{P}(a)$ and $\mathcal{C}(a)$ are correlated, it is often the case that $p(\mathcal{P}(a)) + p(\mathcal{C}(a)) > 1$ (see example of Figure 2(c)). Therefore, adding new arguments in a debate may increase the likelihood of both $\mathcal{P}(a)$ and of $\mathcal{C}(a)$, thus providing more grounds for both opinions.

The question remains open how to understand polarization and bipolarization in this framework. According to PAT, polarization is to be viewed as a change of the degree of our attitude towards $a$ after an exchange of information. In this framework the likelihoods $p(\mathcal{P}(a))$ and $p(\mathcal{C}(a))$ are the essential components of such degree. The latter can be encoded by a weighting function $w$ whose arguments are $p(\mathcal{P}(a))$ and $p(\mathcal{C}(a))$. Many such functions are possible candidates. The most natural is the simple arithmetic difference, i.e. $w(p(\mathcal{P}(a)), p(\mathcal{C}(a))) = p(\mathcal{P}(a)) - p(\mathcal{C}(a))$. Another option is provided by the measures defined by [6]. It is also likely that different subjects $i$ and $j$ have dif-

ferent ways of weighting, i.e. $w_i$ and $w_j$. Following this thread, polarization of an individual $i$ is measured as the change of the value of $w_i(p(\mathcal{P}(a)), p(\mathcal{C}(a)))$ after information exchange with others. Analogously, we have a bipolarization effect between two subjects $i$ and $j$ when we register a change of $w_i(p(\mathcal{P}(a)), p(\mathcal{C}(a)))$ and $w_j(p(\mathcal{P}(a)), p(\mathcal{C}(a)))$ towards opposite directions after information exchange.

Some general considerations are in order to conclude. Group polarization is a widespread phenomenon which, to some extent, speaks against theories attributing to collective discussion a positive role in improving the global performance of a group. Two most notable cases are Surowiecki's wisdom of crowds [11] and Mercier and Sperber's argumentative theory of reasoning [7]. However, the contrast is often only apparent. For example Surowiecki stresses very clearly that the wisdom of crowds applies mostly to *cognition*, *coordination* and *cooperation* tasks (see [11], introduction p. XVIII). Now, many contexts in which polarization arises, like political debate [10], dont fall under these categories. Furthermore, Surowiecki stresses three necessary conditions for a crowd to be wise, namely *diversity*, *independence* and *decentralization*. As he himself recognizes ([11], chapter 9) polarization often arises where diversity of opinion fails. The set of conditions suggested by Surowiecki provide interesting tests for future research.

## References


1. 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.
2. 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.
3. 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.
4. A. Hunter. Some foundations for probabilistic abstract argumentation. *COMMA 2012*: 117–128, 2012.
5. H. Li, N. Oren and T.J. Norman. Probabilistic Argumentation Frameworks *Lecture Notes in Computer Science* 7132: 1–16, 2011.
6. P. Matt and F. Toni. A Game-Theoretic Measure of Argument Strength for Abstract Argumentation. *Lecture Notes in Computer Science* 5293: 285–297, 2008.
7. H. Mercier and D. Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34: 57–111, 2011.
8. G.S. Sanders and R.S. Baron Is social comparison irrelevant for producing choice shifts? *Journal of Experimental Social Psychology* 13: 303–314, 1977.
9. J.A. Stoner. A comparison of individual and group decision involving risk MA thesis, Massachusetts Institute of Technology, 1961.
10. C. Sunstein. *Why societies need Dissent.* Cambridge, Harvard University Press, 2003.
11. J. Surowiecki. *The Wisdom of Crowds.* New York: Anchor Books, 2005.
12. 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.