Identifying Malicious Behavior in Multi-party Bipolar Argumentation Debates

Dionysios Kontarinis and Francesca Toni

Imperial College London, UK

Abstract. Lately, several works have analyzed potential uses of argumentation in multi-party debates. Usually, the focus of such works is the computation of a collectively "correct" outcome, a challenging task even when the debate's users truthfully express their beliefs. This work focuses on debates where some users may exhibit specific types of "malicious" behavior: they may lie (by making statements they do not believe to hold) and they may hide valuable information (by not making relevant statements they believe to hold). Our approach is the following: firstly, we define "user attributes" which capture different aspects of a user's behavior in a debate (how active, how opinionated and how classifiable a user has been); then, we build and test experimentally hypotheses that, from the values of these attributes, can predict whether a user has lied and/or hidden valuable information.

1 Introduction

Several works, e.g. [12, 13, 4, 10], have analyzed potential uses of argumentation in multi-party debates. Some focus on computing a "correct" collective outcome [10], given the users' claims, a challenging task even when users truthfully express their opinions. Others, e.g. [12, 13, 15], focus on user strategies.

In [12] concepts from game theory are used for the analysis of argumentation debates. It is considered that some users, in order to satisfy their preferences, may exhibit "malicious" behavior: they may lie (by making claims they do not believe to hold) and, they may hide (by not making claims they believe to hold).

In this work we attempt the analysis of argumentation debates in order to estimate which users have exhibited malicious behavior. We assume that there is an issue, which is an argument, being debated, as for example in [11]. We also assume that each user has a viewpoint over that issue, in the form of a (private) bipolar argumentation framework [1, 5], which has two types of relations over arguments: an attack relation and a support relation. Users engage in a debate, by progressively stating new attacks and supports. These debates can be seen as abstractions of opinion exchanges in social networks, in general, and in argumentation-inspired social networks such as `www.convinceme.net` and `www.quaestio-it.com` [8]. In these settings, users have no access to the private argumentation frameworks of other users, and therefore no way to assess the truthfulness of information contributed to the debate. Our work aims at helping users and debate administrators estimate user truthfulness.

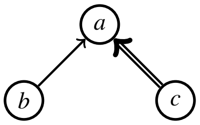
Our approach is the following: firstly, we define several *user attributes* which capture different aspects of a user's behavior (how active, opinionated, and classifiable a user has been in a debate). Then, we build and test experimentally *hypotheses* that, from the values of these attributes, predict whether a user may have lied and/or hidden valuable information. The experimental evaluation is in Java and simulates and analyzes a large number of debates. Albeit preliminary, the results seem to suggest that user attributes may indeed be good indicators of lying and hiding.

The paper is organized as follows. In Section 2 we present background on bipolar argumentation. In Section 3 we define our general debate framework. In Section 4 we define user attributes. In Section 5 we define lying and hiding in our debate setting, and we propose two hypotheses for identifying malicious behavior, which are experimentally tested in Section 6. In Section 7 we conclude.

2 Background on Bipolar Argumentation

A *Bipolar Argumentation Framework* (BAF) [5, 1] is a tuple $\langle Arg, Att, Sup \rangle$ where: Arg is a set, whose elements are referred to as *arguments*, $Att \subseteq Arg \times Arg$, referred to as *attack relation* over arguments, and $Sup \subseteq Arg \times Arg$, referred to as *support relation* over arguments. We will represent BAFs as graphs whose nodes are elements of Arg and whose edges are of two types: simple arrows, to represent attack in Att, and double arrows, to represent support in Sup, as illustrated in the following example.

Example 1. Three users take part in a debate about global warming. The issue being debated is argument a = "global warming should be addressed now, because it already affects our ecosystems." User u_2 introduces the attack (b,a) with b = "there is no conclusive proof of global warming taking place", and then user u_3 introduces the support (c,a) with c = "recent studies show that global warming effects are real". User u_1 observes the debate but, when he is able to intervene, he refuses to contribute. The debate gives rise to the BAF $\langle Arg, Att, Sup\rangle$ with $Arg = \{a,b,c\}$, $Att = \{(b,a)\}$ and $Sup = \{(c,a)\}$, represented graphically as:



Arguments in BAFs may be evaluated using a number of different methods (known as "semantics"), falling broadly within two classes: (1) methods for determining *acceptable sets of arguments*, e.g. as in [3], and (2) methods for determining a *numerical strength*, e.g. as in [5,2]. We shall focus on the latter approach, but we will not commit to any specific method until Section 6. Until then, we will use a generic evaluation function $\sigma : Arg \to [0,1]$ but assume that the addition to a BAF of a support for (attack against) an argument x increases (resp. decreases) $\sigma(x)$. In Section 6 we will choose σ from [2], for which this assumption holds.

3 A General Debate Framework

The starting point for our work is a general framework for multi-party argumentation debate focused on the evaluation of a specific argument, the *issue* of the debate, and involving users with different viewpoints with respect to that issue, represented by *private BAFs*. The evaluation of the issue, after the aggregation of all users' opinions, can be deemed a *collective goal*, shared by all users. Also, we assume that each user pursues a *personal goal*, which is either the maximization or the minimization of the issue's evaluation.

Definition 1. *Let a be the* (debate) issue. *Let $\mathcal{U}$ be a set of* users. *Each $u \in \mathcal{U}$ has a* private BAF*, denoted $AS_u = \langle Arg_u, Att_u, Sup_u\rangle$, such that $a \in Arg_u$, and a* personal goal*, which is either $max\sigma(a)$ or $min\sigma(a)$.*

A debate takes place in discrete timesteps. At each timestep, users introduce attacks against and/or supports for arguments, or pass (introducing no attack or support).

Definition 2. *A* debate *is a tuple $\mathcal{D} = \langle a, \mathcal{U}, IntroAtt, IntroSup, IntroPass\rangle$ such that:*
$IntroAtt \subseteq \{\langle t,u,(x,y)\rangle \mid t \in \mathbb{N},\ u \in \mathcal{U}\}$; $IntroSup \subseteq \{\langle t,u,(x,y)\rangle \mid t \in \mathbb{N},\ u \in \mathcal{U}\}$;
$IntroPass \subseteq \{\langle t,u,pass\rangle \mid t \in \mathbb{N},\ u \in \mathcal{U}\}$ where pass is a constant.

In the remainder of the paper, unless otherwise indicated, we will assume as given a debate $\mathcal{D}$=$\langle a, \mathcal{U}$, $IntroAtt, IntroSup, IntroPass\rangle$. The *first timestep* of a debate is 0, the *last timestep* is defined as follows:

Definition 3. *The* last timestep *of $\mathcal{D}$ is $lastTs(\mathcal{D})$ such that*
if $IntroAtt = IntroSup = IntroPass = \{\}$ then $lastTs(\mathcal{D}) = 0$, otherwise $lastTs(\mathcal{D}) = t$ such that
1. *$\exists\langle t,u,(x,y)\rangle \in IntroAtt \cup IntroSup$ or $\exists\langle t,u,pass\rangle \in IntroPass$, and*
2. *$\nexists\langle t',u',(x',y')\rangle \in IntroAtt \cup IntroSup$ and $\nexists\langle t',u',pass\rangle \in IntroPass$ such that $t' > t$.*

All users' introductions in a debate lead to the emergence of a collective opinion in the form of a common BAF, that we call *gameboard* as in [4,10]. The fact that a debate "remembers" all the introductions that users made, and when they made them, means that it is possible to compute the gameboard at every timestep, up to and including the last, as follows:

Definition 4. *Let $t \in \mathbb{N}$ be such that $0 \le t \le lastTs(\mathcal{D})$. The* gameboard *of $\mathcal{D}$,* at timestep *t, is the BAF $GB^{\mathcal{D}}_t = \langle Arg^{\mathcal{D}}_t, Att^{\mathcal{D}}_t, Sup^{\mathcal{D}}_t\rangle$, such that:*
$Arg^{\mathcal{D}}_t = \{a\} \cup \{x,y \mid \exists\langle t',u,(x,y)\rangle \in IntroAtt \cup IntroSup,\ 0 \le t' \le t\}$;
$Att^{\mathcal{D}}_t = \{(x,y) \mid \exists\langle t',u,(x,y)\rangle \in IntroAtt,\ 0 \le t' \le t\}$;
$Sup^{\mathcal{D}}_t = \{(x,y) \mid \exists\langle t',u,(x,y)\rangle \in IntroSup,\ 0 \le t' \le t\}$.

As an illustration, the BAF of Example 1 is the gameboard $GB_3^{\mathcal{D}}$ at the last timestep of the following debate:

$\mathcal{D} = \langle a, \{u_1, u_2, u_3\}, \{\langle 1, u_2, (b,a)\rangle\}, \{\langle 2, u_3, (c,a)\rangle\}, \{\langle 3, u_1, pass\rangle\}\rangle$

Debates and gameboards are motivated by and provide abstractions of a number of currently available online debate platforms, such as `www.convinceme.net` and `www.quaestio-it.com`. In these platforms, users are able to make claims and back them up with relevant arguments, expressed in natural language, as well as introduce relations between arguments, such as attacks and supports, or simply observe.

In the remainder, we will use the following notations. $Intro_u^{\mathcal{D}} = \{\langle t,u,obj\rangle \mid \langle t,u,obj\rangle \in IntroAtt \cup IntroSup \cup IntroPass\}$ denotes all the introductions by user u in $\mathcal{D}$ (similarly for $IntroAtt_u^{\mathcal{D}}$, $IntroSup_u^{\mathcal{D}}$, $IntroPass_u^{\mathcal{D}}$). Moreover, for an introduction $i = \langle t,u,obj\rangle$, where $obj = (x,y)$ or $obj = pass$, the function $ts(i)$ returns its timestep t, while the function $rel(i)$ returns (x,y) or $pass$, respectively. Furthermore, $\sigma_t^{\mathcal{D}}(x)$ denotes the evaluation, using σ as in Section 2, of argument x in $GB_t^{\mathcal{D}}$. For all notations, if $\mathcal{D}$ is clear from the context, we will drop the $\mathcal{D}$ superscript. Finally, we will refer to the set of all possible debates as Δ, and to the union of all $Arg_t^{\mathcal{D}}$ for all $t \in \mathbb{N}$ and $\mathcal{D} \in \Delta$ as $Arg_{\mathbb{N}}^{\Delta}$.

4 User Behavior Analysis

In order to analyze user behavior in multi-party argumentation debates, we define three *(user) attributes*, each capturing a specific aspect of user behavior. The first two attributes measure how active and opinionated a user has been in a debate. Thus, they describe a user's *general stance* in a debate. The third attribute estimates how similar a user's *beliefs* are to those of some known user classes.

The *activity attribute* indicates the quantity of a user's contribution in a debate. Roughly, the less pass introductions a user makes, the more active he is considered.

Definition 5. *The* activity evaluation *of a user in a debate is given by the function active* : $\mathcal{U} \times \Delta \rightarrow [0,1]$*: if* $|Intro_u| = 0$*, then* $active(u,\mathcal{D}) = 0$*, else*

$$active(u,\mathcal{D}) = \frac{|IntroAtt_u \cup IntroSup_u|}{|Intro_u|}$$

Next, the *opinionatedness attribute* indicates how one-sided a user's impact has been on an argument's evaluation (with respect to σ). Roughly, the more a user has increased (or decreased) an argument's evaluation, the more opinionated he is deemed on that argument.

Definition 6. *The* opinionatedness evaluation *of a user on an argument in a debate is given by the function* $opinionated : \mathcal{U} \times Arg_{\mathbb{N}}^{\Delta} \times \Delta \rightarrow [0,1]$*: if* $\sum_{i \in Intro_u} |\sigma_{ts(i)}(x) - \sigma_{ts(i)-1}(x)| = 0$*, then* $opinionated(u,x,\mathcal{D}) = 0$*, else*

$$opinionated(u,x,\mathcal{D}) = \frac{|\sum_{i \in Intro_u} \sigma_{ts(i)}(x) - \sigma_{ts(i)-1}(x)|}{\sum_{i \in Intro_u} |\sigma_{ts(i)}(x) - \sigma_{ts(i)-1}(x)|}$$

In the fraction above, the numerator reflects how one-sided u's impact has been on $\sigma(x)$ (either increasing it, or decreasing it), while the denominator reflects how large u's overall impact has been on $\sigma(x)$.

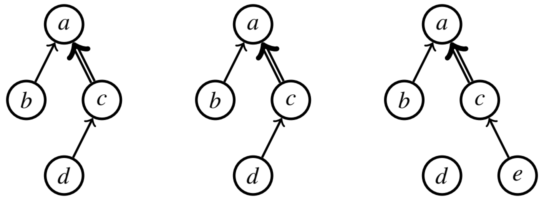
Example 2. Let us assume that in debate $\mathcal{D}$, user u has made two relation introductions, and three pass introductions. So, $active(u,\mathcal{D}) = \frac{|IntroAtt_u \cup IntroSup_u|}{|Intro_u|} = \frac{2}{2+3} = \frac{2}{5} = 0.4$. Moreover, let us assume that u's first relation introduction had increased $\sigma(a)$ from 0.2 to 0.7 and u's second relation introduction had decreased $\sigma(a)$ from 0.5 to 0. Thus, $opinionated(u,a,\mathcal{D}) = \frac{|(0.7-0.2)+(0-0.5)|}{|0.7-0.2|+|0-0.5|} = \frac{0}{1} = 0$. Notice that this is the lowest possible opinionatedness value, indicating that u is not opinionated at all towards a. This is sensible, since u has equally increased and decreased a's evaluation.

In order to define the third user attribute, we introduce the notion of *user class*. In practice, if some users think in a similar way about a topic, then we may say that they *belong to the same class*. For example, there may be a class of users who believe global warming is a threat, and another class who believe it is not. Most probably, users of the same class will agree on many points, though not on everything. For example, a scientist may consider an elaborate argument that another user will not. We define user classes as BAFs:

Definition 7. *Let $\mathcal{C}$ be a set of* classes. *For each $\kappa\in\mathcal{C}$, $AS_\kappa=\langle Arg_\kappa,Att_\kappa,Sup_\kappa\rangle$ is a BAF s.t. $Att_\kappa\cup Sup_\kappa\neq\{\}$.*

Users have personal BAFs which may be similar, but not identical, to classes, as illustrated next.

Example 3. For some $\kappa\in\mathcal{C}$, let AS_κ be as given below. Let $u_1,u_2\in\mathcal{U}$ with private BAFs AS_{u_1} and AS_{u_2} as given below. Both u_1 and u_2 may be deemed to *belong to* κ, even though AS_{u_2} is not identical to AS_κ.



AS_κ AS_{u_1} AS_{u_2}

The *classifiability attribute* estimates how *distant* a user is from classes in some given set. The notion of distance we use is inspired by the *edit distance*, used e.g. in [6], to measure the similarity of argumentation systems, albeit in our case it depends on a class, on introductions by the user alone, and on those by the other users.

Definition 8. *The function distance*: $\mathcal{U}\times\mathcal{C}\times\Delta\rightarrow[0,1]$ *is defined as* $distance(u,\kappa,\mathcal{D})=\frac{dsg_{u,\kappa}}{agr_{u,\kappa}+dsg_{u,\kappa}}$*, with* $agr_{u,\kappa}$ *(*$dsg_{u,\kappa}$*), the number of* agreements *(resp.* disagreements*) between u and* κ*, computed as follows:*

1. *Set* $agr_{u,\kappa}:=0$, $dsg_{u,\kappa}:=0$.
2. *For every* (x,y) *such that either* $(x,y)\in Att_\kappa\cup Sup_\kappa$*, or* $\exists\langle t,u,(x,y)\rangle\in IntroAtt_u\cup IntroSup_u$*, find the corresponding case in the following table, where Rel is one of Att or Sup:*

Case for (x,y)	$\exists\langle t,u,(x,y)\rangle\in IntroRel_u$	$(x,y)\in Rel_\kappa$	$\exists\langle t',u',(x,y)\rangle\in IntroRel_{u'}$ with $u'\neq u$	Considered
1	*Yes*	*Yes*	*Yes*	*Agreement*
2	*Yes*	*Yes*	*No*	*Agreement*
3	*Yes*	*No*	*Yes*	*Disagreement*
4	*Yes*	*No*	*No*	*Disagreement*
5	*No*	*Yes*	*Yes*	*Agreement*
6	*No*	*Yes*	*No*	*Disagreement*
7	*No*	*No*	*Yes*	*Agreement*
8	*No*	*No*	*No*	*Agreement*

If the column "Considered" gives "Agreement", then $agr_{u,\kappa}:=agr_{u,\kappa}+1$*, else* $dsg_{u,\kappa}:=dsg_{u,\kappa}+1$.

Since the BAF of a class cannot (by definition) be without attacks and without supports, it can be proven that the denominator of $\frac{dsg_{u,\kappa}}{agr_{u,\kappa}+dsg_{u,\kappa}}$ is always different from zero. According to the above definition of distance, a disagreement between u and κ can take place in three cases: in cases 3 and 4, where u has introduced a relation (x,y) which κ does not have, and in case 6, where u has not introduced a relation (x,y) which κ has, and no other user has introduced it either. We consider case 5 as an agreement, because it is redundant for u to re-introduce (x,y), since this introduction will not change the gameboard.

The more disagreements there are between u and κ, the greater their distance is. Then, user classifiability depends on the distance between the user and the class which is "closest" to him. The more their distance decreases (increases), the more classifiability increases (resp. decreases).

Definition 9. *The* classifiability evaluation *of a user w.r.t. a set of classes in a debate is given by the function classifiable* : $\mathcal{U}\times 2^{\mathcal{C}}\times\Delta\rightarrow[0,1]$ *such that* $classifiable(u,K,\mathcal{D})=1-\min_{\kappa\in K} distance(u,\kappa,\mathcal{D})$.

Example 4. Let $\mathcal{D}=\langle a,\{u_1,u_2,u_3\},\{\langle 1,u_1,(b,a)\rangle\},\{\langle 2,u_2,(c,a)\rangle\},\{\langle 3,u_3,pass\rangle\}\rangle$. Here, $IntroAtt_{u_1}=\{\langle 1,u_1,(b,a)\rangle\}$ and $IntroSup_{u_1}=\{\}$. Moreover, the gameboard of $\mathcal{D}$ at 3 is the BAF in Example 1. Let κ be as in Example 3. To determine $distance(u_1,\kappa)$ we consider, in turn, all the attacks and supports either introduced by u_1, or belonging in AS_κ. u_1 has introduced the attack (b,a) and it belongs to Att_κ, so we have an agreement (case 2). There is no other introduction by u_1, so we now check the relations of κ. The support (c,a) belongs to Sup_κ, u_1 has not introduced it, but another user has introduced it (u_2), so we have another agreement (case 5). Finally, the attack (d,c) belongs to Att_κ, u_1 has not introduced it, and neither has any other user, so we have a disagreement (case 6). In total, $agr_{u_1,\kappa} = 2$ and $dsg_{u_1,\kappa} = 1$, thus $distance(u_1,\kappa,\mathcal{D}) = \frac{1}{3}$. As a result, $classifiable(u_1,\{\kappa\},\mathcal{D}) = 1 - \frac{1}{3} = \frac{2}{3}$.

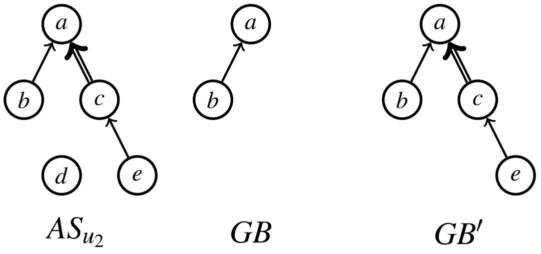
5 Malicious User Behavior

In the context of multi-party debates, *malicious* user behavior could be defined in various ways, e.g. in terms of aggressivity. In this work, we shall call a user malicious if he undermines the satisfaction of the collective goal (of evaluating the issue after aggregating all users' opinions, see Section 3), More specifically, we assume that this can happen in two ways: (1) the user may avoid to contribute towards the satisfaction of the collective goal by *hiding* attacks and supports which would affect the issue's evaluation; (2) the user may mislead the satisfaction of the collective goal by *lying* (e.g. as studied in [14]); in our setting, this amounts to stating attacks and supports he does not believe (i.e. they are not in his private BAF). Formally:

Definition 10. *$\langle t,u,(x,y)\rangle \in IntroRel$ (such that Rel is one of Att, Sup) is a* lie *if and only if $(x,y) \notin Rel_u$; $\langle t,u,pass\rangle$ is a* hide *if and only if $\exists (x,y) \in Rel_u$ (such that Rel is one of Att, Sup) such that, given $\mathcal{D}'$ obtained from $\mathcal{D}$ after deleting $\langle t,u,pass\rangle$ (from $IntroPass_u^{\mathcal{D}}$) and adding $\langle t,u,(x,y)\rangle$ (to $IntroRel_u^{\mathcal{D}}$): $\sigma_t^{\mathcal{D}'}(a) \neq \sigma_t^{\mathcal{D}}(a)$; $\langle t,u,obj\rangle$ is a* malicious action *if and only if it is a lie or a hide.*

A user may decide to perform a malicious action if this may help him achieve his personal goal (see Definition 1). From now on, users who adopt the goal of maximizing (minimizing) the issue's evaluation are said to be *PRO* (resp. *CON*). Let us give an example of strategic lying and hiding in our setting:

Example 5. Let u_2 be a *CON* user, whose private BAF AS_{u_2} is illustrated below:



Firstly, we analyze the case where the gameboard of the dialogue is *GB* above. The only relation in AS_{u_2} whose introduction can change a's evaluation in *GB* is the support (c,a). By introducing it, a's evaluation increases (by the assumptions at the end of Section 2). But, since u_2 is *CON*, he may decide to pass instead. This would be a hide. Secondly, let us analyze the case where the gameboard is GB' above. User u_2 cannot introduce any relation from his private BAF to change a's evaluation, but he may decide to introduce attack (d,c) and decrease the issue's evaluation. Since attack (d,c) is not in AS_{u_2}, that would be a lie.

Now that we have defined two types of malicious behaviour, we propose a method for identifying it. But first, in addition to assuming that users are either *PRO* or *CON*, we also assume that there is a set of classes K, such that each user belongs to one of them.

We now formulate two hypotheses about the relation between the three user attributes defined in Section 4 and malicious actions.

Let z_u denote the evaluation of an attribute, for user u, and let $z_u \uparrow$ ($z_u \downarrow$) indicate that the evaluation of that attribute for u is higher (resp. lower) than the average evaluation of that attribute for all users. Also, let $\mathcal{L}_u^{\mathcal{D}}$ ($\mathcal{H}_u^{\mathcal{D}}$) denote the percentage of introductions made by u in $\mathcal{D}$ which were lies (resp. hides). Finally, let $\mathcal{L}_u^{\mathcal{D}} \uparrow$ ($\mathcal{L}_u^{\mathcal{D}} \downarrow$) indicate that the percentage of lies by u in $\mathcal{D}$ is higher (resp. lower) than the average percentage of lies by all users. Similarly, let $\mathcal{H}_u^{\mathcal{D}} \uparrow$ ($\mathcal{H}_u^{\mathcal{D}} \downarrow$) indicate that the percentage of hides by user u

in $\mathcal{D}$ is higher (resp. lower) than the average percentage of hides by all users.

Hypothesis 1. $active(u,\mathcal{D})\uparrow$ and $opinionated(u,a,\mathcal{D})\uparrow$ and $classifiable(u,K,\mathcal{D})\downarrow\implies\mathcal{L}_u^{\mathcal{D}}\uparrow$

Hypothesis 1 says that a combination of high activity, high opinionatedness on the issue, and low classifiability indicates a liar. The intuition behind it is as follows. Firstly, a higher-than-average activity may indicate a liar, since a liar is not only introducing relations appearing in his private BAF (as honest users do), but he also "makes-up" introductions (lies) when they are useful to him. Secondly, a higher-than-average opinionatedness on the issue may indicate a liar, since lies are always introduced strategically in order to increase or decrease the issue's evaluation. Thirdly, a lower-than-average classifiability may indicate a liar. Indeed, since we have assumed that every user belongs to some class, their private BAFs are somewhat similar to the BAF of that class. When a user "disagrees" with every class (and has a lower-than-average classifiability), there are two possibilities: either (i) the "disagreement" is honest, or (ii) the user is lying. Finally, a combination of the given values for all three attributes is an even stronger indication of a liar.

Hypothesis 2. $active(u,\mathcal{D})\downarrow$ and $opinionated(u,a,\mathcal{D})\uparrow\implies\mathcal{H}_u^{\mathcal{D}}\uparrow$

Hypothesis 2 says that a combination of low activity and high opinionatedness on the issue indicates a hider. The intuition behind it is as follows. Firstly, a lower-than-average activity may indicate a hider, since a hider refrains from introducing relations which do not help achieve his personal goal (contrary to honest users). Secondly, a higher-than-average opinionatedness may indicate a hider, for the same reason. Finally, a combination of the given values for the two attributes is an even stronger indication of a hider.

Note that we do not claim that these hypotheses lead to the definite identification of liars and hiders. Nonetheless, they may be valuable to debate system administrators and to users, as they can raise "red flags" about potentially malicious users.

6 Experimental Evaluation

In order to test hypotheses 1 and 2 we conducted an experimental evaluation. To the best of our knowledge, no repository exists of debates for which the maliciousness of participants is known. Therefore, we opted for simulating debates, as follows.

- **Generation of BAFs**. Since many existing platforms model debates as trees, all BAFs in the simulation (for classes, users, gameboards) were trees, all with the issue as root. We chose for these trees to have a maximum branching factor of 4 and to contain at most 20 arguments and thus 19 relations.
- **Argument evaluation**. We used the function σ defined as the final score of [2].
- **Generation of user classes**. For each debate, we randomly generated 3 user classes (with their BAFs).
- **Generation of users**. From each user class, we randomly generated 4 users, as follows: we replicated with a 90% probability each relation (and its arguments) from the class' BAF into the user's BAF. Thus each user was structurally "similar", but possibly not identical, to one class. The 4 users comprised of
  - one *honest user* who could never lie nor hide;
  - one *liar user* who could lie as many times as he wanted, but could never hide; the possible lies of a user u were restricted to a randomly generated set of attacks and supports $PossLies_u$;
  - one *hider user* who could hide as many times as he wanted, but could never lie;
  - one *malicious user* who could lie, restricted to $PossLies_u$, and hide, as many times as he wanted.

  For each experiment, the 12 users (4 × 3 classes) were partitioned into *PRO* and *CON*, as follows: a user u was *PRO* if and only if the issue's evaluation $\sigma(a)$ in AS_u was at least 0.5, and *CON* otherwise.
- **Debate protocol**:
  - users took *turns* following an order over $\mathcal{U}$ and introducing a single relation or pass per turn;
  - users were *not allowed* to introduce an attack or support already in the debate gameboard;
  - each debate *terminated* after $|\mathcal{U}|$ passes in a row.
- **User strategies**. Each user u followed the strategy described informally below, with $t = lastTs(\mathcal{D})$:
  1. u computes the set P of all his *possible relation introductions* at $t+1$, i.e. $\langle t+1,u,(x,y)\rangle$ such that $(x,y)\in Att_u\cup Sup_u$ and $\sigma_t^{\mathcal{D}}(a)\neq\sigma_{t+1}^{\mathcal{D}'}(a)$, where $\mathcal{D}'$ is $\mathcal{D}$ after introducing any member of P;

2. if u is a *liar user* or a *malicious user* then u computes the set M of all *possible lie relation introductions* at $t+1$, i.e. $\langle t+1, u, (x,y)\rangle$ such that $(x,y) \in \mathit{PossLies}_u$ and $\sigma_t^{\mathcal{D}}(a) \neq \sigma_{t+1}^{\mathcal{D}'}(a)$ where $\mathcal{D}'$ is $\mathcal{D}$ after introducing any member of M; all elements of M are then added to P;
3. if $P = \{\}$ then the strategy returns $\langle t+1, u, \mathit{pass}\rangle$; else let $\mathcal{D}^{pass}$ be $\mathcal{D}$ after the pass introduction $\langle t+1, u, \mathit{pass}\rangle$ and let $\mathcal{D}^i$ be $\mathcal{D}$ after the relation introduction i;
    (a) if u is *PRO* let $i^* = \underset{i \in P}{\mathrm{argmax}}\,\sigma_{t+1}^{\mathcal{D}^i}(a)$; if $\sigma_{t+1}^{\mathcal{D}^{i^*}}(a) > \sigma_{t+1}^{\mathcal{D}^{pass}}(a)$ then the strategy returns i^*;
    (b) if u is *CON* let $i^* = \underset{i \in P}{\mathrm{argmin}}\,\sigma_{t+1}^{\mathcal{D}^i}(a)$; if $\sigma_{t+1}^{\mathcal{D}^{i^*}}(a) < \sigma_{t+1}^{\mathcal{D}^{pass}}(a)$ then the strategy returns i^*;
    (c) if $P \setminus \mathit{PossLies}_u = \{\}$ the strategy returns $\langle t+1, u, \mathit{pass}\rangle$;
    (d) if u is a *hider user* or a *malicious user* then the strategy returns $\langle t+1, u, \mathit{pass}\rangle$;
    (e) otherwise, the strategy returns a random member of $P \setminus \mathit{PossLies}_u$.

Intuitively, the user first tries to (greedily) choose the relation introduction which is best (or tied for best) for him (cases 3.(a) and 3.(b)). But, if all possible relation introductions are bad for him, then: if they are all lies, then he "honestly" passes (case 3.(c)), whether he is a liar, a hider, both or neither. Otherwise, there exist some truthful moves, and in this case a hider or malicious user will pass (hiding), whereas a user who cannot hide will be forced to play a bad move.

We implemented this debate setting in Java, and we generated and analyzed a number of debates as follows. For each debate, for each of the 12 users in it, we calculated whether he had lied more than the average user (or not), and whether he had hidden more than the average user (or not). Then, for each user u, we tested the prediction given by the two hypotheses: if the attribute values of u were as indicated on the left-hand side of Hypothesis 1 (Hypothesis 2), then we estimated that u was an above-average liar (resp. hider). We also tested the predictions given by reversing the hypotheses: if the attribute values of u were *not* as indicated on the left-hand side of Hypothesis 1 (Hypothesis 2), then we estimated that u was *not* an above-average liar (resp. hider). This led to the following types of estimations:

- *Correct estimations*:
    - **true positive**: u was predicted to be a liar (resp. hider), and he was;
    - **true negative**: u was predicted not to be a liar (resp. hider), and he was not.
- *Erroneous estimations*:
    - **false positive**: u was predicted to be a liar (resp. hider), but he was not.
    - **false negative**: u was predicted not to be a liar (resp. hider), but he was.

We generated $100,000$ debates, with $100,000 \times 12 = 1,200,000$ users. The results are summarised below:

Malicious action	True positives	False positives	True negatives	False negatives
Lying	163,408	73,829	726,681	236,082
Hiding	211,514	193,378	613,497	181,611

Albeit preliminary, these results seem to confirm our hypotheses: as far as lying is concerned, there were approximately 2.2 times more true positives than false positives, and approximately 3 times more true negatives than false negatives; as far as hiding is concerned, the number of false positives was relatively high (compared to lying), though still lower than the number of true positives, while there were approximately 3.5 times more true negatives than false negatives.

7 Conclusion

Malicious user behavior analysis in on-line debates has recently caught the attention of researchers, e.g. in the case of *trolls* or *flamers* [9]. At the same time, interest in multi-party debates has grown steadily. Argumentation-based debate platforms offer their users the possibility of expressing their thoughts in a structured way, e.g. by introducing arguments and relations among them. This paper is a first step towards undertaking the analysis of malicious user behavior in multi-party argumentation debates. We have identified and evaluated empirically, in a simulated debate setting, two hypotheses providing indications of potential malicious behaviour, in the form of lying and hiding. The hypotheses are formulated in terms of three user attributes, computed by observing users debate and measuring their activity, opinionatedness and classifiability. Albeit preliminary, the evaluation shows promise.

Our work has several limitations and opens the way to many directions for future work. It would be interesting to evaluate our approach in a real debate setting, rather than a simulated environment. Other user attributes, e.g. focus on specific arguments, may also provide useful information. Moreover, in addition to general-stance and belief attributes of the kinds we considered, we could consider a third category of attributes describing the relations a user has with others, for example his popularity. Our analysis focused on single debates, but it may be useful to compare the behaviour of users across debates (potentially in different platforms). We have programmed agents to follow a specific strategy, but other strategies, e.g. the ones overviewed in [15], may be interesting. Other hypotheses may provide further indications of maliciousness, and it may be interesting to *learn*, rather than guess as in this paper, relationships between user attributes and malicious behaviour. Our evaluation was restricted to a specific semantics for bipolar argumentation: it would be interesting to study the impact of different choices of semantics for prediction of maliciousness. The argumentation framework used throughout this work was just an example and it is possible to use other, more elaborate, argumentation frameworks instead, e.g. allowing for votes as in [7, 8]. Another direction for future research includes the identification of additional types of malicious users, such as trolls, flamers, or even users simply searching for “friends” and neglecting the collective goal of a debate. Also, it may be interesting to draw insights from existing work on lying [14] in order to further characterise malicious behaviour.

Acknowledgments. This research was supported by the EU project *DesMOLD (FP7/2007-2013-314581)*.

References

1. Amgoud, L., Cayrol, C., Lagasquie-Schiex, M.C., Livet, P.: On bipolarity in argumentation frameworks. International Journal of Intelligent Systems 23, 1062–1093 (2008)
2. Baroni, P., Romano, M., Toni, F., Aurisicchio, M., Bertanza, G.: Automatic evaluation of design alternatives with quantitative argumentation. Argument & Computation 6(1), 24–49 (2015), special issue: Applications of logical approaches to argumentation
3. Boella, G., Gabbay, D.M., van der Torre, L.W., Villata, S.: Support in abstract argumentation. In: Proceedings of the 3rd International Conference on Computational Models of Argument (COMMA’10). pp. 111–122 (2010)
4. Bonzon, E., Maudet, N.: On the outcomes of multiparty persuasion. In: Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS’11). pp. 47–54 (2011)
5. Cayrol, C., Lagasquie-Schiex, M.C.: Gradual valuation for bipolar argumentation frameworks. In: Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU’05). pp. 366–377 (2005)
6. Coste-Marquis, S., Devred, C., Konieczny, S., Lagasquie-Schiex, M.C., Marquis, P.: On the Merging of Dung’s Argumentation Systems. Artificial Intelligence 171, 740–753 (2007)
7. Egilmez, S., Martins, J., Leite, J.: Extending social abstract argumentation with votes on attacks. In: Proceedings of the Second Workshop on Theory and Applications of Formal Argumentation (TAFA’13). pp. 16–31 (2013)
8. Evripidou, V., Toni, F.: Quaestio-it.com: a social intelligent debating platform. Journal of Decision Systems 23(3), 333–349 (2014)
9. Hardaker, C.: Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. Journal of Politeness Research. Language, Behaviour, Culture 6, 215–242 (2010)
10. Kontarinis, D., Bonzon, E., Maudet, N., Moraitis, P.: Picking the right expert to make a debate uncontroversial. In: Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA’12). pp. 486–497 (2012)
11. Prakken, H.: Coherence and flexibility in dialogue games for argumentation. Journal of Logic and Computation 15(6), 1009–1040 (2005)
12. Rahwan, I., Larson, K.: Argumentation and game theory. In: Argumentation in Artificial Intelligence, pp. 321–339. Springer (2009)
13. Riveret, R., Prakken, H., Rotolo, A., Sartor, G.: Heuristics in argumentation: A game theory investigation. In: Proceedings of the Second International Conference on Computational Models of Argument (COMMA’08). pp. 324–335 (2008)
14. Sakama, C., Caminada, M., Herzig, A.: A logical account of lying. In: Logics in Artificial Intelligence, pp. 286–299. Springer (2010)
15. Thimm, M.: Strategic argumentation in multi-agent systems. Künstliche Intelligenz 28(3), 159–168 (2014)