OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1007/978-3-319-75553-3_12

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Kohan Marzagao, D., Murphy, J., Young, A. P., Gauy, M., Luck, M. M., McBurney, P. J., & Black, E. (2018). Team Persuasion. In *The 3rd International Workshop on Theory and Applications of Formal Argument* (pp. 159-174). (Lecture Notes in Computer Science ; No. 10757). Advance online publication. https://doi.org/10.1007/978-3-319-75553-3_12

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 13. Jan. 2025

Team Persuasion

David Kohan Marzagão[1⋆], Josh Murphy[1], Anthony P. Young[1], Marcelo Matheus Gauy[2⋆⋆], Michael Luck[1], Peter McBurney[1], Elizabeth Black[1]

[1] Department of Informatics, King's College London
[2] Institut für Theoretische Informatik, ETH Zürich

{david.kohan, josh.murphy, peter.young, michael.luck, peter.mcburney, elizabeth.black}@kcl.ac.uk, marcelo.matheus@inf.ethz.ch

Abstract. We consider two teams of agents engaging in a debate to persuade an audience of the acceptability of a central argument. This is modelled by a bipartite abstract argumentation framework with a distinguished topic argument, where each argument is asserted by a distinct agent. One partition defends the topic argument and the other partition attacks the topic argument. The dynamics are based on flag coordination games: in each round, each agent decides whether to assert its argument based on local knowledge. The audience can see the induced sub-framework of all asserted arguments in a given round, and thus the audience can determine whether the topic argument is acceptable, and therefore which team is winning. We derive an analytical expression for the probability of either team winning given the initially asserted arguments, where in each round, each agent probabilistically decides whether to assert or withdraw its argument given the number of attackers.

1 Introduction

Argument-based persuasion dialogues provide an effective mechanism for agents to communicate their beliefs and reasoning in order to convince other agents of some central topic argument [11]. In complex environments, persuasion is a distributed process. To determine the acceptability of claims, a sophisticated agent or audience should consider multiple, possibly conflicting, sources of information that can have some level of agent-hood. In this paper, we consider teams of agents that work together in order to convince some audience of a topic argument. While strategic considerations have been investigated for one-to-one persuasion (e.g. [15]), and for one-to-many persuasion (e.g. [9]), the act of persuading as a team is a largely unexplored problem.

Consider a political referendum, where two campaigns seek to persuade the general public of whether or not they should vote for or against an important proposition. Each campaign consists of separate agents, where each agent is an expert in a single argument. For example, an environmentalist might argue how a favourable outcome in the referendum would reduce air pollution. Each agent

⋆ Supported by CNPq (206390/2014-9)
⋆⋆ Supported by CNPq (248952/2013-7)

can assert its argument to the public, and each agent is aware of counterarguments that other agents can make. However, no agent can completely grasp all aspects of the campaign, for example the environmentalist may be ignorant of relevant economic issues. If the agent thinks there are no counterarguments to its argument it should keep asserting its argument, as it is beneficial for its team. While each agent wishes to further other team's persuasion goal, they do not want to risk having their argument publicly defeated by counterarguments.

From this example, we consider a team of agents to have three key properties that differentiate them from an individual agent when persuading. Firstly, each agent may have localised knowledge which is inaccessible and non-communicable to other agents in the same team. Secondly, agents may not be wholly benevolent, potentially acting in their own interest before that of their team; reconciling this conflict between individual and team goals makes strategising more complex. Thirdly, there is no omniscient or authoritative agent able to determine the actions of other members of the team, meaning each agent must act independently, making the problem a distributed one. This problem is distinct from that of an individual persuader, and therefore requires a different approach to model the outcomes of persuasion.

We approach the problem of modelling team persuasion by exploring a particular team persuasion game, in which two opposing teams attempt to convince an audience of whether some central issue, termed the *topic*, is acceptable or not. Each member of a team is individually responsible for one argument in the domain, being an expert on that particular argument. As such, each member must independently decide whether to actively assert its argument to the audience, or to hold back from asserting its argument. The persuasion game proceeds in rounds, where in each round an agent decides whether to assert its argument. An agent can decide to stop asserting its argument even if in previous rounds they had asserted it. Teams aim to reach a state in which the topic is acceptable or unacceptable according to the audience (depending on whether the agent is defending or attacking the topic), and in which no individual agent will change its decision of whether to assert its argument; in such a state the topic is guaranteed to retain its (un)acceptability indefinitely. When deciding whether to assert its argument, an agent takes into account whether the other agents are currently asserting their arguments. It aims to have a positive effect on its team's persuasion goal, but may also wish to avoid having its own argument publicly defeated (since this may, for example, negatively affect their public standing or reputation). When deciding whether to assert its argument, the agent must therefore balance the potential positive effect of this on its team's persuasion goal with the risk of its own argument being publicly defeated.

The audience determines whether they find the topic argument acceptable in a particular round by considering the set of arguments that are currently asserted. Note that the audience has no knowledge of which arguments were asserted in previous rounds; we consider the audience to be memoryless, only considering the arguments that are asserted in the current round.

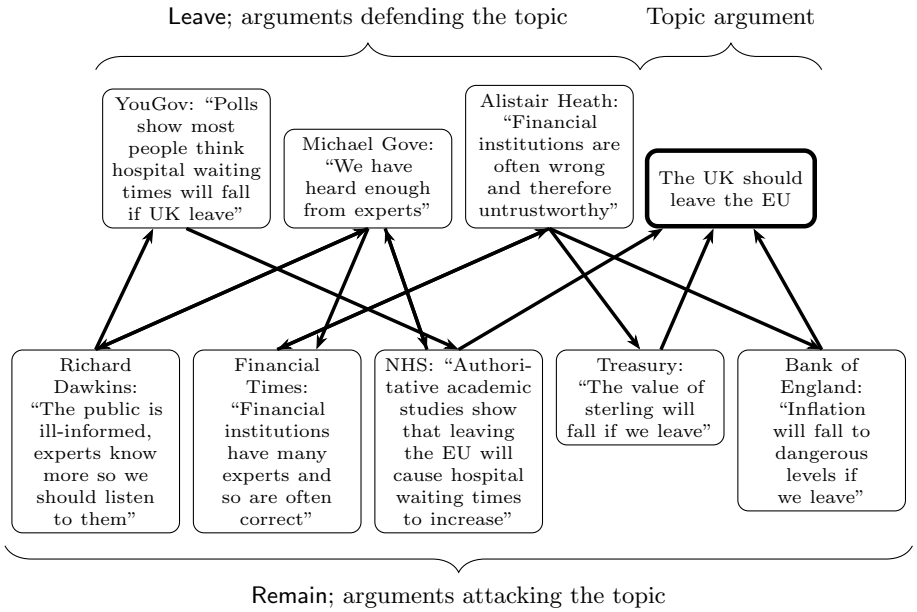

Fig. 1. An instantiated example of a bipartite argumentation framework.

For example, consider the arguments in Figure 1, in which the directed edges represent conflict between arguments. The topic argument in this example is that the United Kingdom should leave the European Union, with three arguments defending the topic and five arguments attacking the topic (some indirectly). Each argument is controlled by a particular individual or institution. The agents are organised into two teams, those defending the topic (the Leave campaign), and those attacking the topic (the Remain campaign). Consider the argument that might be asserted by the Treasury: the Treasury is motivated to assert their argument as it directly attacks the topic argument (which they are seeking to dissuade the audience of). If they are aware of the argument possibly asserted by Alistair Heath, they may decide not to assert their own argument to avoid the risk of being publicly defeated. The public decides whether leaving the European Union is acceptable depending on which arguments are currently being asserted. The contribution of this paper is the application of *team persuasion games* [10] to model public debates of this form. We answer the following:

Q1 How do we formalise the situation where one team has *definitively won*? We define such a situation to be a state where agents that are asserting their arguments will continue to do so, and agents not asserting their arguments will never do so.

Q2 What is the probability that a particular team (e.g. the Remain Campaign) has definitively won? We prove an expression for this probability, given the initially asserted arguments and the attacks between them.

In Section 2 we define a team persuasion game on a bipartite abstract argumentation framework [6], which is a special case of a *flag coordination game* [10]. In Section 3, we use our framework to answer Questions **Q1** and **Q2**. We discuss related work in Section 4, and conclude in Section 5.

2 Team persuasion games

In this section we present our model of team persuasion games. We begin by briefly reviewing the relevant aspects of abstract argumentation [6].

Definition 1. *An **argumentation framework** is a directed graph (digraph) $AF := \langle A, R \rangle$ where A is the set of arguments and $R \subseteq A \times A$ is the attack relation, where $(a, b) \in R$ denotes that the argument a attacks the argument b.*

Figure 1 is an example argumentation framework. We will only consider *finite*, non-empty argumentation frameworks, i.e. where $A \neq \varnothing$ is finite. Given an argumentation framework, we can determine which sets of arguments (*extensions*) are justified given the attacks [6]. There are many ways (*semantics*) to do this, each based on different intuitions of justification. We do not assume a specific semantics in this paper, only that all agents and the audience use the same semantics.

Definition 2. *Let AF be an argumentation framework. The set $\textsc{Acc}(AF) \subseteq A$ is **the set of acceptable arguments of** AF, with respect to some argumentation semantics under credulous or sceptical inference. An argument a is said to be **acceptable** with respect to AF iff $a \in \textsc{Acc}(AF)$.*

We model team persuasion as an instance of a *flag coordination game* over an argumentation framework [10]. A flag coordination game consists of a network of agents and an index representing discrete time. Each agent has a set of flags of different colours (representing e.g. choices or states) and a set of other agents it can see. In each round, each agent raises a coloured flag synchronously and independently, as the output of some (possibly random) decision procedure given what the agent sees other agents doing in the preceding round. Such models have been studied in the context of the adoption of new technology standards, voting and achieving consensus [10, Section 1]. We now adopt a specific instance of a flag coordination game for our purposes.

Definition 3. *A **team persuasion framework** is a tuple $\langle AF, X, \beta, \Gamma, \phi, \Lambda \rangle =: \mathcal{F}$. Let $AF = \langle A, R \rangle$ be an argumentation framework, where the nodes represent arguments, each owned by distinct agents.*[3] *Let $\phi : A \to \mathcal{P}(A)$ be **the visibility function**,*[4] *i.e. $\phi(a) \subseteq A$ is the set of arguments that a can see. Let $X := \{on, off, topic\}$ denote **the set of states**. Let $t \in A$ be a distinguished argument called **the topic (argument)**. Define **the state function** $\beta : A \to X$ such that $\beta(t) := topic$ and $(\forall a \in A - \{t\})\, \beta(a) \in \{on, off\}$.*[5]

[3] As each *argument* is owned by a distinct *agent*, we use the terms interchangeably.
[4] If X is a set, then $\mathcal{P}(X)$ is its power set.
[5] We further assume that ϕ is such that if $b \in \phi(a)$ then a can also see $\beta(b) \in X$.

Let $\mathcal{S} := X^A$ be the space of functions that assigns a state to each argument, which defines a ***configuration****. Let $\Gamma \subseteq \mathcal{S}$ be* ***the set of goal states****. For $a \in A$ let λ_a be* ***the decision algorithm of agent*** *a, that takes input β and ϕ and outputs $s(a) \in X$, for $s \in \mathcal{S}$. We define Λ as the set of algorithms for all $a \in A$.*

The team persuasion framework is such that each agent asserts a single argument, which can attack and be attacked by other asserted arguments, so it is isomorphic to an argument framework. Each of the agents can assert their argument (turning it *on*) or not assert their argument (turning it *off*). The topic is a special argument that is labelled *topic* throughout the duration of the game.

Definition 4. *Let $\mathcal{F}$ denote a team persuasion framework. Let $i \in \mathbb{N}$ denote discrete time. Consider the sequence of configurations $[s_0, s_1, ...]$, indexed by i. We call s_0* ***the initial configuration****, and s_i is* ***the i^{th} configuration****. The update rule is such that for all $a \in A - \{t\}$, $s_{i+1}(a) \in X$ is the output of λ_a given $s_i(b) \in X$ for all $b \in \phi(a)$ and possibly $\beta(a)$. Further, $(\forall i \in \mathbb{N})\, s_i(t) := topic$.* ***A team persuasion game with initial configuration*** *s_0 is the tuple $\langle \mathcal{F}, s_0 \rangle$.*

Initially, the agents start in some initial configuration defined by whether each agent asserts its argument. In each subsequent round, the agents decide using their own decision procedure whether to assert or stop asserting their argument in the next round, given the actions of other agents they see.

Both teams are presenting their arguments to an audience who are assumed to be memoryless across rounds and can only see the topic and the arguments that are being currently presented. This prompts the following definition.

Definition 5. *Given a team persuasion game,* ***the set of arguments that are on in round*** *i is $A_i^{on} := \{a \in A \,|\, s_i(a) = on\} \cup \{t\}$. The* ***induced argument framework*** *is $AF_i^{on} := \langle A_i^{on}, R_i^{on} \rangle$, where $R_i^{on} := R \cap [A_i^{on} \times A_i^{on}]$.*

The audience will therefore see a sequence of argument frameworks $(AF_i^{\text{on}})_{i \in \mathbb{N}}$ as the teams debate each other about the topic. The audience can determine which team is winning based on whether the topic is justified in a given round.

Definition 6. *In a given round $i \in \mathbb{N}$ of a team persuasion game, we say that* ***the team of defenders are winning*** *iff $t \in \textsc{Acc}\,(AF_i^{on})$ iff* ***the team of attackers are not winning****.*

In each round the acceptability of the topic may change, and hence the winner can change. We are interested in definitively winning states, as defined in Q1 in Section 1. We explore the existence of such states in Section 3.

Since we are modelling the arguments between two teams, each trying to persuade or dissuade an audience of the topic, we specialise to bipartite argumentation frameworks because no agent should attack an argument of another agent in its own team. Further, the framework is weakly connected because all arguments asserted are relevant to the debate. Further, we assume that every argument has a counterargument, and that the topic is not capable of defending itself, so it does not directly attack any argument.

Definition 7. *Our team persuasion frameworks* $\mathcal{F} = \langle AF, X, \beta, \Gamma, \phi, \Lambda \rangle$ *have an underlying argument framework* $AF = \langle A, R \rangle$ *that is bipartite and weakly connected, with the requirements that* $(\forall a \in A)\, a^- \neq \varnothing$ *and* $t^+ = \varnothing$*. As* AF *is bipartite, let* U *and* W *be the two partitions of* A *such that* $t \in U$*. We call* U ***the set of defenders of the topic****, and* W ***the set of attackers of the topic****. The* ***set of goal states*** *is* $\Gamma := \{\gamma_u, \gamma_w\}$*, where* $\gamma_u(U - \{t\}) = \{on\}$ *and* $\gamma_u(W) = \{off\}$*, and* $\gamma_w(U - \{t\}) = \{off\}$ *and* $\gamma_w(W) = \{on\}$*.*[6]

The goal states indicate that each team has the goal of unilaterally asserting their arguments and making the opposing team unilaterally withdraw their arguments. See Figure 2 for an example of γ_u, and Figure 3 for an example of γ_w. In our figures, white (resp. black) nodes are arguments that are *on* (resp. *off*).

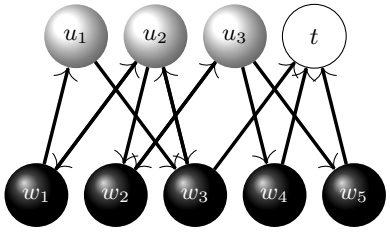

Fig. 2. The defenders' goal state γ_u; all defenders are asserting their argument.

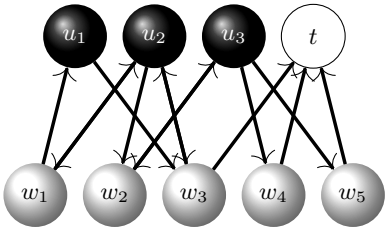

Fig. 3. The attackers' goal state γ_w; all attackers are asserting their argument.

2.1 Agent visibility

There are several possible forms of the agents' visibilities, ϕ, for example:
V1 $(\forall a \in A)\, \phi(a) = a^- := \{b \in A \,|\, (b, a) \in R\}$,
V2 $(\forall a \in A)\, \phi(a) = a^+ := \{b \in A \,|\, (a, b) \in R\}$, or
V3 $(\forall a \in A)\, \phi(a) = a^- \cup a^+$ (both).
Recall from Footnote 5 that if $b \in \phi(a)$, then a can also see the state $s(b)$ of b. It is possible to define $\phi(a) \subseteq A$ to be completely arbitrary, beyond the immediate neighbours of a. However, it is not currently clear how the behaviour of an agent might be influenced by knowledge of the states of arguments beyond the immediate neighbours especially if there is to be localised knowledge (see Section 3). In this paper, we focus on V1, leaving the rest for future work.

2.2 The agents' decision algorithm

We claim that agents with visibility of a^- can be motivated by two factors: their desire to make the topic acceptable/unacceptable to the audience (the goal of the team), and their desire not to have their argument publicly defeated (the goal of the individual). An individual does not want to have its argument publicly

[6] Recall that for a function $f : X \to Y$ and $A \subseteq X$ the *image set of* A *under* f is $f(A) := \{y \in Y \,|\, (\exists x \in A)\, f(x) = y\}$.

defeated (that is, its argument is asserted, but is not considered acceptable by the audience in the current round), as it is somehow a challenge to the agent's authority, and therefore reflects negatively on its ego. An agent can estimate how likely it is that their argument will be publicly defeated by considering how many attacking arguments the agent could see are being asserted: the more attackers that are asserted, the more likely one of the attacks will be successful, and therefore the higher the chance its argument is defeated.

- **Altruistic**: An agent which is only motivated by the team goal of making the topic (un)acceptable would always assert its argument a, regardless of the state of the arguments in a^-. We call such selfless agents *altruistic*.
- **Timid**: An agent which is only motivated by its individual goal of not having its argument being publicly defeated would never assert its argument, regardless of which arguments in a^- are being asserted. If the agent never asserts its argument, it can never be defeated, and therefore will always achieve its individual goal.
- **Balanced**: An agent motivated by both factors must find a way to balance these two goals. Such an agent is certain to assert its argument when none of its attackers are asserted, because the chance of a successful defeat is minimal. Similarly, the agent is least likely to assert when all of its attackers are asserted because the chance of successful defeat is maximised.

As a starting point for our analysis we will consider balanced agents. We define the probability of the agent not asserting its argument when all of its attackers are *on* as 1, and conversely the probability of the agent not asserting its argument when all of its attackers are *off* as 0. To begin with, we assume this probability increases linearly, proportional to the number of arguments a^- that are *on*.

Definition 8. *Let $\mathcal{F}$ be a team persuasion framework on an argument framework AF as defined in Definition 7. An agent $a \in A - \{t\}$ is **balanced** iff λ_a (Definition 3) is defined as follows. For $i \in \mathbb{N}$, λ_a outputs $s_{i+1}(a) = off$ with conditional probability*

$$\mathbb{P}\left(s_{i+1}(a) = off \middle| s_i\right) := \frac{|a^- \cap A_i^{on}|}{|a^-|} \in [0,1]. \tag{1}$$

Further, λ_a outputs $s_{i+1}(a) = on$ given s_i with probability $1 - \mathbb{P}\left(s_{i+1}(a) = off \middle| s_i\right)$. We will assume that for all $a \in A - \{t\}$, a is balanced.

Example 1. Consider Figure 4, which represents the situation in Figure 1 as a team persuasion framework with the initial configuration where the *on* arguments are u_2, u_3, w_2, and w_3, with the rest of the arguments being *off*. Consider the argument w_3. It is attacked by u_1 and u_2, which are respectively *off* and *on*. Therefore, the probability of w_3 remaining *on* in the next round is $\frac{1}{2}$.

3 Results

From the setup described in Section 2, we can now answer more precise versions of the two questions posed at the end of Section 1.

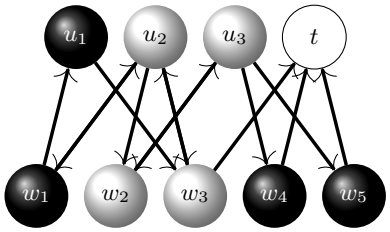

Fig. 4. An Initial Configuration $(\mathcal{F}, s_0)$ for the example in Figure 1.

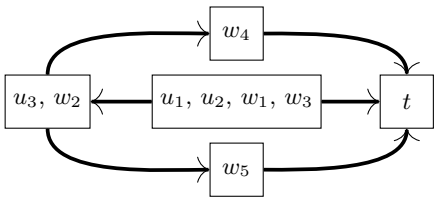

Fig. 5. Condensation graph of Figure 4, showing strongly connected components.

F1 Are there any states of the arguments (*on* or *off*) in which no agent is going to change their state in any future round according to λ_a as defined in Equation 1? We call such a state a *state-stable configuration.*[7]

F2 What is the probability of a particular team winning, i.e. achieving a state-stable configuration, where the topic is either acceptable or unacceptable?

3.1 State-Stable Configurations

We now answer Question F1, which concerns state-stable configurations.

Definition 9. *A **state-stable configuration** is a function $s \in \mathcal{S}$ such that, if attained at round $i \in \mathbb{N}$ of the team persuasion game following Equation 1, will also be the state of the game in all subsequent rounds.*

This formalises the intuition that no agent is going to change their state in any future round once a state-stable configuration is reached.

Proposition 1. *Given the setup of Section 2, the two goal states, γ_u and γ_w (Definition 7) are the only state-stable configurations.*

Proof. Please refer to Appendix A for all proofs in this paper.

3.2 Probabilities for State-Stable Configurations

We now answer Question F2. We first translate our team persuasion game into a *consensus game*. In a consensus game, the update is such that in round $i+1$, every digraph node a *copies* the colour of a randomly (uniformly) sampled neighbour in a^-, rather than adopting the opposite colour as in Equation 1 [10].

The translation is as follows. We consider the finite, weakly connected, bipartite digraph $G := \langle V, E\rangle$ which is the induced subgraph of $\langle A, R\rangle$ with nodes $V := A - \{t\}$. For each configuration $s : A - \{t\} \to X := \{\text{on, off}\}$ we define a *colouring function* $s' : V \to X' := \{0, 1\}$ such that

$$s'(a) := 1 \text{ if } [(a \in U \text{ and } s(a) = \text{on}) \text{ or } (a \in W \text{ and } s(a) = \text{off})]\,. \quad (2)$$

$$s'(a) := 0 \text{ if } [(a \in U \text{ and } s(a) = \text{off}) \text{ or } (a \in W \text{ and } s(a) = \text{on})]\,. \quad (3)$$

We intuitively associate the colour 1 with the state on and similarly, 0 with off, but notice how this association is swapped for $a \in W$. Thus, the correspondence $s \mapsto s'$ is well-defined and bijective.

[7] This is to avoid confusion with the notion of *stable semantics* [6].

Example 2. Consider the digraph in Figure 4.[8] Given this initial configuration s_0 such that $s_0(u_1) =$ off, $s_0(u_2) =$ on... etc. (see Example 1), we get a corresponding s' where $s'(\{u_2, u_3, w_1, w_4, w_5\}) = \{1\}$ and $s'(\{u_1, w_2, w_3\}) = \{0\}$, by Footnote 6. If we arrange $V = \{u_1, \ldots, u_3, w_1, \ldots, w_5\}$, we can represent s'_0 as the boolean vector $(0, 1, 1, 1, 0, 0, 1, 1)$.

We now give some definitions and results for consensus games on digraphs.

Definition 10. *Let $G = \langle V, E\rangle$ be a finite digraph. Given some fixed order of the nodes $V = \{a_1, \ldots, a_{|V|}\}$,[9] the **(row-normalised) in-matrix** of G is the $|V| \times |V|$ matrix $H := (h_{ij})$, where*

$$\text{if } (v_j, v_i) \in E \text{ then } h_{ij} = \frac{1}{|v_i^-|}, \text{ else } h_{ij} = 0. \tag{4}$$

The intuition of Equation 4 is that the i^{th} node $v_i \in V$ has a probability $h_{ij} > 0$ to copy the colour of v_j when $(v_j, v_i) \in E$.

Definition 11. *Let $G = \langle V, E\rangle$ be a digraph. Its **condensation** is the digraph $\langle \mathcal{K}, \mathcal{E}\rangle$ such that $\mathcal{K} \subseteq \mathcal{P}(V)$ is the set of strongly connected components (SCCs) of G and $(K, K') \in \mathcal{E} \subseteq \mathcal{K}^2$ iff $[(\exists a \in K)(\exists b \in K')(a, b) \in E$ and $K \neq K']$. A **source component** is a component with no in degree.*

Example 3. The condensation of Figure 4 is Figure 5. The only source component is $\{u_1, u_2, w_1, w_3\}$.

The following theorem answers Question F2 with an analytic expression of the probability of a particular team winning. We then apply this to solve our motivating example in Example 4. Intuitively, we first look at the condensation of a given bipartite AF. Since source components are not going to be influenced by any external argument, the probability of them reaching either one of the state-stable configurations is independent of the eventual state of the rest of the AF. Also, non-source components have no influence over the final outcome, since once the source components stabilise, they will be a constant influence in either defending or attacking the topic. Thus, we need all source SCCs to converge to the same state-stable configuration, otherwise a global state-stable configuration will not be reached. Finally, in order to calculate the probability of either the defender or the attacker to win in each source SCC, we find each individual agents' influence in the network.

Theorem 1. *Consider a team persuasion game on a bipartite $AF = \langle A, R\rangle$ with initial colouring s'_0. Let $\mathcal{K} = \{\{t\}, K_1, \ldots, K_m\}$ be the set of SCCs of AF (for some $m \in \mathbb{N}^+$), where $\{t\}$ is the component that contains only the topic argument. We also define $\textbf{source}_{\mathcal{K}} \subseteq \mathcal{K}$ as the set of source SCCs in the condensation of AF. Let $\mathcal{K}_{\{t\}} \subseteq \textbf{source}_{\mathcal{K}}$ denote the set of SCCs for which there is a $\mathcal{E}$-path in the condensation of AF to $\{t\}$.*

[8] Note that Figure 5 will be relevant for a following proof.

[9] In the context of team persuasion games, we write all nodes in U first and then the nodes in W, as in Example 2.

*Let μ_K be the stationary distribution of H_K, where H_K[10] is the in-matrix of the subgraph of AF induced by $K \in \mathcal{K}$ (Definition 10). Let $\mu(K) = \sum_{a \in K} \mu(a)$ for $K \subset A$. Finally, each set $K \in \mathcal{K}_{\{t\}}$ has a value g that stands for the greatest common divisor (*gcd*) of the lengths of all cycles in K. This generates a g-partite AF with partitions $\{K^1, \ldots, K^g\}$ as in Lemma 2 (Appendix A). We have that*[11]

$$\mathbb{P}(\gamma_u \textit{ is reached} \mid s_0) = \prod_{K \in \mathcal{K}_{\{t\}}} \prod_{i=1}^{g} \left(\frac{1}{\mu(K^i)} \sum_{a \in K^i} \mu_{K^i}(a) s_0'(a) \right). \tag{5}$$

Example 4. Consider the bipartite $AF = \langle V, E \rangle$ in Figure 1 and s_0 as in Figure 4. The condensation graph can be seen in Figure 5, so $\mathcal{K} = \{\{t\}, K_1, K_2, K_3, K_4\}$, where $K_1 = \{u_1, u_2, w_1, w_3\}$, $K_2 = \{u_3, w_2\}$, $K_3 = \{w_4\}$ and $K_4 = \{w_5\}$. K_1 is the only source component. Since K_1 (indirectly) influences the acceptability of the topic, we have $\mathcal{K}_{\{t\}} = \{K_1\}$. We now need to evaluate $\mu = \mu_{K_1}$, a stationary distribution of the in-matrix H_{K_1}, the induced subgraph of AF generated by K_1. Then, we have

$$\mu H_{K_1} = \mu \Leftrightarrow \mu \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{pmatrix} = \mu \Rightarrow \mu = \frac{1}{10}(1, 4, 3, 2). \tag{6}$$

Note that $g = 2$. We now use the initial configuration s_0 and the translation to s_0' according to Equations 2 and 3. We have $s_0'(u_1) = 0$, $s_0'(u_2) = 1$, $s_0'(w_1) = 0$, $s_0'(w_3) = 0$, therefore, by Theorem 1, we have

$$\begin{aligned} \mathbb{P}(\gamma_u \text{ is reached} \mid s_0) &= \prod_{K \in \mathcal{K}_{\{t\}}} \prod_{i=1}^{g} \left(\frac{1}{\mu(K^i)} \sum_{a \in K^i} \mu_{K^i}(a) s_0'(a) \right) \\ &= \left(\frac{1}{\mu(K_1^1)} \frac{4}{10} \right) \left(\frac{1}{\mu(K_1^2)} \frac{3}{10} \right) = \frac{12}{25} = 48\%. \end{aligned} \tag{7}$$

Therefore, the probability of the topic being accepted is 48%. Analogously, the probability of the topic being rejected is given by

$$\mathbb{P}(\gamma_w \text{ is reached} \mid s_0) = \left(\frac{1}{\mu(K_1^1)} \frac{1}{10} \right) \left(\frac{1}{\mu(K_1^2)} \frac{2}{10} \right) = \frac{2}{25} = 4\%. \tag{8}$$

The probability for this game not reaching a state-stable configuration is 48%.

4 Related Work

In this paper we have presented and analysed an argumentation framework for a very common form of public debate. Our work has made two novel contributions. The first contribution is the formalisation using argumentation frameworks of public policy debates where multiple parties with only local information propose

[10] Recall the row vector μ is the stationary distribution of H iff $\mu H = \mu$.

[11] We have abused notation here: we have considered γ_u to be a state configuration not on the entire AF, but just on the subgraph induced by the arguments that have a path to the topic. In other words, we exclude arguments that do not even indirectly influence the acceptability of the topic.

arguments to support (or attack) claims of interest to a wider audience, seeking to persuade that audience of a claim (or not, as the case may be). The second contribution is the use of flag co-ordination games, specifically its analysis of the dynamics of graph colouring, to understand the properties of this formal framework. Analogues of graph colouring have been used in argumentation, for example, in labelling semantics to determine acceptability of arguments [3]. However, to the best of our knowledge, interpreting such colourings as the argument having been asserted or not, and the dynamics of how such a colouring changes, have not previously been used in argumentation theory.

The general problem of two parties with contradictory viewpoints, each seeking to persuade an impartial third party of their viewpoint, has been investigated in economics, e.g. using game theory [13,14] or mechanism design [7,8]. Applying argumentation theory to study multi-agent persuasion with two teams, in which one is arguing for the acceptability of a topic and the other against, has been investigated in the work by Bonzon and Maudet [2]. They focus specifically on the problem with respect to the kinds of dialogue that occur on social websites, specifying that agents "vote" on the attack relations between arguments. One of the main differences between their work and ours is that they assume that each agent has a total view of the argumentation framework, where as we assume agents have a specific area of expertise and thus, in general, do not have complete knowledge of the structure of the argumentation framework. Furthermore, agents in their formulation do not have any motivation to act in a way that might be detrimental to their team's goal, whereas agents in our work may also be motivated by their own individual goals.

Dignum and Vreeswijk developed a testbed that allows an unrestricted number of agents to take part in an inquiry dialogue [5]. The focus of their work is on the practicalities of conducting a multi-party dialogue, concerned with issues like turn-taking, rather than in the strategising of agents participating in such a dialogue. Bodanza *et al.* [1] survey work on how multiple argumentation frameworks may be aggregated into a single framework. While this direction of work considers how frameworks from multiple agents might be merged, it removes the strategic aspect of persuasion which we are interested in here.

5 Conclusion and Future work

We have shown how to determine the probability of each team winning in a team persuasion game (Theorem 1). However, we have shown that not all games become state-stable (Appendix A, Lemma 1), having no definite winner. Considering games which do not become state-stable, we are interested in determining (1) in what proportion of rounds is the topic acceptable, and (2) what is the probability the topic being acceptable at a specific round in the future. With respect to the first question, we might determine the winning team to be the one who makes the topic acceptable/unacceptable in the majority of rounds. The second question is particularly interesting in the context of referendum-like domains, in which there is a set date (round n) in which the audience determines

whether the topic is acceptable (and thus which team wins): in this case it does not matter whether there is state-stability, only that the topic is acceptable in round n.

Future work will apply the techniques of [10] to the situation investigated in this paper. Specifically, if the team persuasion game will reach a goal state, we can calculate the expected number of rounds until that happens [10, Proposition 4]. Further, we can study the game-theoretic implications of some knowledgeable external agent "bribing" a specific agent to either assert or stop asserting its argument [10, Section 3]. We will also investigate different generalisations of the team persuasion game. There are various assumptions on the digraph that we can modify. For example, generalising from bipartite to multipartite argumentation frameworks where many teams seek to persuade the audience. Additionally, we can lift the assumption that no agent attacks its fellow agents of the same team. Such a team seems quite unlikely (and thus is not considered here), but occasionally this may occur, e.g. a campaigner who wishes to leave the EU because their environmental laws are too restrictive on UK businesses, and a campaigner who wishes to leave the EU because they do not have strong enough environmental laws; both campaigners would be on the same team, but their arguments are seemingly in conflict. Further generalisations include: consideration of the different visibility functions ϕ for each agent, or the case where each agent can assert more than one argument, or having a non-linear version of Equation 1. We will show that the results also apply to the case when the attacking arguments are weighted differently by agents in Equation 1, which we will articulate in future work. Ultimately, we hope such generalisations can give insight into situations where individual goals and societal goals conflict to a greater extent, and how this conflict can be resolved.

A Lemmas and Proofs

Proof (of Proposition 1). To show that γ_u is a state stable configuration, notice that in round $i \in \mathbb{N}$, if γ_u is attained, then for $a \in U - \{t\}$, the probability (Equation 1) a will be off in round $i+1$ is zero, because $a^- \subseteq W$ and all attackers of a are off. Therefore, a will still be on in round $i+1$. Similarly, we can show that the probability of being off for all $b \in W$ in round $i+1$ is one. Therefore, in round $i+1$, the state is still γ_u. A similar argument to this proves that if γ_w is attained in round i, then it will also be the state for round $i+1$. By induction over i, γ_u and γ_w satisfy Definition 9.

We now show that both γ_u and γ_w are the only state stable configurations. Assuming the contrary. Then, we have a configuration different from γ_u and γ_w in which no argument has a positive probability of changing their state. In this case, we would have two nodes, say u_1 and u_2, in the same partition, say U, that have different colours (otherwise we have γ_u and γ_w). Since G is weekly connected, there is a path that ignores edges' directions from u_1 to u_2. This path has even length and, therefore, since u_1 to u_2 are different, there must be at least two consecutive nodes in this path with the same colour. One it attacking the

other, therefore, the attacked one has a positive probability of changing their colour. We have a contradiction. Thus γ_u and γ_w must be the only state-stable configurations in a bipartite AF. ■

We now answer a more general version of Question F2 using the framework of consensus games and colours. We derive a formula for a colour to win the consensus game on a strongly connected digraph, given that consensus will be achieved. We then investigate the necessary conditions for consensus to be achieved, and derive an expression for the probability of failing to achieve consensus that depends on s'_0. We then generalise to the case of weakly connected graphs, and answer Question F2 via our translation back into team persuasion games.

The in-matrix H of the digraph G can be seen as a transition matrix of a time homogeneous Markov chain, where the each node v represents a state and the reversed edges represent the transitions. If the Markov chain is irreducible and finite, there is a unique *stationary distribution*, which is a row vector $\mu \in \mathbb{R}_+^V$ that satisfies $\mu H = \mu$.

Proof (of Theorem 1). The theorem follows from the following lemmas. Note that these lemmas are considering a general digraph $G = (V, E)$ and colours 0 and 1. We also denote $\mathbf{0}$ and $\mathbf{1}$ as the consensus on colour 0 and 1 respectively.

Lemma 1. *A consensus game on a strongly-connected digraph $G = \langle V, E\rangle$ reaches consensus with probability for all initial configurations 1 iff $\gcd C = 1$, where $C \subseteq \mathbb{N}$ is the set of the lengths of all cycles in G. In the case $\gcd C = g > 1$, then G is g-partite with parts $V_1, \ldots, V_g$ where all edges go from V_i to V_{i+1}.*

Proof (of Lemma 1). ($\Leftarrow$) Assuming $\gcd C = 1$. Then, given an initial configuration, a game has already reached consensus or it has not. If not, we can assume, WLOG, that there is at least one $v \in V$ coloured 0. We note that the $\gcd C_v = 1$, where C_v is the set of the lengths of the cycles passing through v. This follows from the fact that G is strongly connected. We can then show that there is a large enough $n_0 > 0$ such that for any $n \geq n_0$, we have $\mathbb{P}(s_n(u) = 0 \mid s_0) > 0$ for all $u \in V$. For that it is enough to show that there is finite n_0, such that for every $n \geq n_0$ there is a directed path from v to u of length n. The existence of such n_0 follows from Lemma 2.1 of [12]. Thus, if the game runs long enough, it will reach consensus (either $\mathbf{0}$ or $\mathbf{1}$) with probability 1.

($\Rightarrow$) We now want to prove that if the game reaches consensus with probability 1, then $\gcd C = 1$. We are going to prove this by showing that if $\gcd C > 1$, then there is a positive chance that the game never reaches consensus. Let $\gcd C = g > 1$. We start by showing that the graph must be not only a g-partite graph, but also of the form that every edge from a node in partition i points to a node in partition $i + 1 (\mathrm{mod}\, g)$. Let $v \in V$. For all $w \in V$, we define the partition that w belongs to by taking the $x (\mathrm{mod}\, g)$, where x is the length of any path from v to w.

We show that this is well defined. First, the existence of such a path is guaranteed by the strongly connectivity of G. Also, the lengths of all paths from v to w must coincide modulo g. If not, by concatenating both paths to the same

returning path from w to v, we would have created two cycles from v to v that differ in length modulo g (by assumption, all cycles must be $0(\mathrm{mod}\, g)$).

We now observe that, if the game reaches a configuration in which a partition is all 0 and another all 1, consensus will never be reached. Thus it can not be reaching consensus for sure from all possible initial configurations. We will show that no non-consensual initial configuration reaches consensus with probability 1. ■

Lemma 2. *Consider a consensus game in a strongly connected and direct graph $G = \langle V, E\rangle$ in which $\gcd C = g$. Then, we know by Lemma 1 that G is g-partite*[12] *and we denote the partitions $V_1, \ldots, V_g$. We further denote $\mu(U) = \sum_{v\in U} \mu(v)$ for $U \subset V$. In these conditions,*

$$\mathbb{P}(\textit{Colour 1 wins in } G \mid s_0) = \prod_{i=1}^{g} \left(\frac{1}{\mu(V_i)} \sum_{v\in V_i} \mu(v) s_0(v) \right) \tag{9}$$

Proof (of Lemma 2). We use a similar approach to the one in [10] and apply Theorem 1 of [4]. Note that the state of vertices of V_{i+1} in the round $n+1$, depends only in the state of vertices of V_i in the round n. We can then consider g parallel consensus games on g copies of G, where in the i-th consensus game we set the initial state of the vertices in V_i to their original initial state in the consensus game, but set the state of all other vertices to 1. Denote by p_i the probability of the i-th consensus game reaching a 1 winning state. It is then easy to see that $\mathbb{P}(1 \text{ wins in } G \mid s_0) = \prod_{i=1}^{g} p_i$.

We are left to show that $p_i = \dfrac{1}{\mu(V_i)} \sum_{v\in V_i} \mu(v) s_0(v)$. For that end, over the i-th consensus game define the random variable $X_n = \sum_{v\in V_j} \mu(v) s_n(v)$, where $j = n + i - 1(\mathrm{mod}\, g)$. We show that the process $(X_n)_{n\in\mathbb{N}}$ is a martingale with respect to the sequence s_n. We need to show that $\mathbb{E}(X_{n+1}|s_n) = X_n$. By linearity of expectation $\mathbb{E}(X_{n+1}|s_n) = \sum_{v\in V_{j+1}} \mu(v)\mathbb{E}(s_{n+1}(v)|s_n)$. Note that $\mathbb{E}(s_{n+1}(v)|s_n) = \sum_{u\in V_j} h_{vu} s_n(u)$ and by changing the order of summation we get that: $\mathbb{E}(X_{n+1}|s_n) = \sum_{u\in V_j} s_n(u) \sum_{v\in V_{j+1}} \mu(v) h_{vu}$. Due to stationarity of μ and the fact that h_{vu} is non-zero only for $v \in V_{j+1}$, we have that $\sum_{v\in V_{j+1}} \mu(v) h_{vu} = \mu(u)$, which implies that $\mathbb{E}(X_{n+1}|s_n) = X_n$.

Now, it is easy to see that $\mu(V_i)p_i = \mathbb{E}(X_\infty|X_0) = \mathbb{E}(X_0)$ and this proves that $p_i = \dfrac{1}{\mu(V_i)} \sum_{v\in V_i} \mu(v) s_0(v)$, which concludes the result. ■

Lemma 3. *Consider a consensus game played in a weekly connected digraph $G = \langle V, E\rangle$ and let $\mathcal{K} = \{K_1, \ldots, K_n\}$ be the set of strongly connected components (SCC) of G. We define $\boldsymbol{source}_{\mathcal{K}}$ as the set of SCCs that have no attack coming from the outside, i.e., if $K \in \mathcal{K}$, then $K \in \boldsymbol{source}_{\mathcal{K}}$ if for every $(a, b) \in E$ such that $b \in K$, we have $a \in K$. Then,*

$$\mathbb{P}(\textit{Colour 1 wins in } G \mid s_0) = \prod_{K\in \boldsymbol{source}_{\mathcal{K}}} \mathbb{P}(\textit{Colour 1 wins in } K \mid s_0) \tag{10}$$

[12] By 1-partite, we mean $\gcd C = 1$ and $V_1 = V$

Proof (of Lemma 3). First note that each $K \in \mathbf{source}_{\mathcal{K}}$ is independent of each other, since they are independent from anything outside each of these SCCs. Then, we cannot have consensus if they reach different consensus. It remains now to observe that, in the case they reach same consensus colours, then all the other SCCs will eventually stabilise in the same colour. That happens because of the influence they receive from components in $\mathbf{source}_{\mathcal{K}}$, so consensus cannot be achieved by any other colour. Finally, for every node not in a source component, there is a path from a source node to it, therefore there is a non-zero probability that the game achieves the sources' colour. ■

References

1. G. Bodanza, F. Tohme, and M. Auday. Collective Argumentation: A Survey of Aggregation Issues Around Argumentation Frameworks. *Journal of Argument and Computation*, 8(1):1–34, 2016.
2. E. Bonzon and N. Maudet. On the Outcomes of Multiparty Persuasion. In *Proceedings of the 8th International Workshop on Argumentation in Multi-Agent Systems*, pages 86–101. Springer, 2012.
3. M. Caminada. On the issue of reinstatement in argumentation. *Logics in artificial intelligence*, 4160:111–123, 2006.
4. C. Cooper and N. Rivera. The Linear Voting Model: Consensus and Duality. 2016.
5. F. Dignum and G. Vreeswijk. Towards a Testbed for Multi-party Dialogues. *Advances in Agent Communication*, pages 212–230, 2004.
6. P. M. Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77:321–357, 1995.
7. J. Glazer and A. Rubinstein. Debates and Decisions: On a Rationale of Argumentation Rules. *Games and Economic Behavior*, 36(2):158–173, 2001.
8. J. Glazer and A. Rubinstein. On Optimal Rules of Persuasion. *Econometrica*, 72(6):1715–1736, 2004.
9. A. Hunter. Toward Higher Impact Argumentation. In *Proceedings of the The 19th American National Conference on Artificial Intelligence*, pages 275–280. MIT Press, 2004.
10. D. Kohan Marzagão, N. Rivera, C. Cooper, P. McBurney, and K. Steinhöfel. Multi-Agent Flag Coordination Games. In *Proceedings of the 16th International Conference on Autonomous Agents & Multiagent Systems*, pages 1442–1450. International Foundation for Autonomous Agents and Multiagent Systems, 2017.
11. H. Prakken. Formal Systems for Persuasion Dialogue. *The Knowledge Engineering Review*, 21(02):163–188, 2006.
12. J. C. Rosales and P. A. García-Sánchez. *Numerical Semigroups*, volume 20. Springer Science & Business Media, 2009.
13. H. Shin. The Burden of Proof in a Game of Persuasion. *Journal of Economic Theory*, 64:253–264, 1994.
14. H. Shin. Adversarial and Inquisitorial Procedures in Arbitration. *RAND Journal of Economics*, 29:378–405, 1998.
15. M. Thimm. Strategic Argumentation in Multi-Agent Systems. *Künstliche Intelligenz*, 28:159–168, 2014.