HAL open science

Deliberation Towards Transitivity with Unshared Features

Arthur Boixel, Pierre Bisquert, Madalina Croitoru

To cite this version:

Arthur Boixel, Pierre Bisquert, Madalina Croitoru. Deliberation Towards Transitivity with Unshared Features. PRIMA 2019 - 22nd International Conference on Principles and Practice of Multi-Agent Systems, Oct 2019, Torino, Italy. pp.16, 10.1007/978-3-030-33792-6_1 . hal-02388879

HAL Id: hal-02388879

https://inria.hal.science/hal-02388879v1

Submitted on 2 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deliberation towards Transitivity with Unshared Features

Arthur Boixel[1], Pierre Bisquert[2], Madalina Croitoru[3]
[1] ILLC, University of Amsterdam
[2] IATE, INRA, France
[3] University of Montpellier, France

December 2, 2019

Abstract

We place ourselves in a decision making setting where a set of agents needs to collectively decide upon a set of alternatives characterised by their features. We introduce the notion of unshared features and show that if such features do not exist then we can reach a Condorcet consensus. We provide a deliberation protocol that ensures that, after its completion, the number of unshared features of the decision problem can only be reduced.

1 Introduction and Motivation

Social choice theory allows to study the way in which the aggregation of individual preferences can lead to the expression of a collective preference. Unfortunately, well-known impossibility results prevent the construction of simple and satisfactory preference aggregation methods [4]. In this paper we focus on a well-known topic in social choice : single-peakedness preferences and their link to the Condorcet paradox. The Condorcet paradox is a situation noted by the Marquis de Condorcet in the late 18th century [5], in which the aggregation of individual preferences *via* pairwise majority can result in cyclic collective preferences, even if the individual preferences are not cyclic. For example, if we consider three agents 1, 2 and 3 and three alternatives *bike*, *car* and *train*, one can encounter the situation where agent 1 prefers *bike* to *car* to *train*, agents 2 prefers *car* to *train* to *bike* and agent 3 prefers *train* to *bike* to *car*. In this case no alternative beats the other in pairwise majority. The *car* is strictly preferred to the *train* by a majority (agents 1 and 2) but the *train* is strictly preferred to the *bike* by an other majority (agents 2 and 3) and the *bike* is strictly preferred to the *car* by a majority (agents 1 and 3). Thus we have no Condorcet winner here (i.e. an alternative that is preferred, pairwise, to all other alternatives by a majority) and we obtain a non transitive result. Although satisfying properties

which are desirable in democracy[1] [1], in such cases, pairwise majority cannot be used to aggregate individual preferences.

One way of going around this situation is to look for necessary or sufficient conditions on the individual preferences that will ensure a well-defined result [15]. But restricting the expression of individual preferences is a non democratic way for their aggregation. Indeed, it forces our preference aggregation method to violate the universality property [1]. To take advantage of the latter conditions without violating this property, we have to better understand them. In the forties, Duncan Black studied cyclic preferences and introduced the notion of *single-peakedness* [3]. A group of agents is said to have single-peaked preferences if each agent has an ideal choice in the set of alternatives, and for each agent, alternatives that are—according to a fix order on the alternatives—further from her ideal choice are less preferred. Single-peaked preferences have the desired property of allowing for a Condorcet winner [3]. For example if we alter the preferences from the previous example and we consider that agent 3 prefers *train* to *car* and *car* to *bike* then the set of individual preferences is single-peaked according to the order $> \ : \ bike > car > train$. Therefore we can conclude that there exists a Condorcet winner, in that case, the *car* alternative. In this paper we place ourselves in a decision making setting where a set of agents needs to collectively decide upon a set of alternatives characterised by their features [6]. The agents have desired features and the satisfaction of these features by the alternatives induces agents' individual preferences. The more an alternative satisfies the desired features of an agent, the higher its rank will be in the agent's individual preferences. In this setting, as explained above, using voting rules satisfying desirable properties as a collective decision making procedure can lead to situations (such as the Condorcet Paradox) where no decision can be made. In this paper we address this problem by studying conditions on the alternatives' features that ensure the avoidance of the Condorcet Paradox. We introduce the notion of *unshared features* and show that if such features do not exist then we can reach a Condorcet consensus. Moreover, we conjecture and empirically prove that the less unshared features there are, the closer we get (with respect to well-known distance measures in the literature) to reaching a Condorcet consensus. Last, we provide a deliberation protocol that ensures that, after its completion, the number of unshared features can only be reduced.

2 Individual Desires and Preference Formation

In this section, we will explore how agents can form their individual preferences based on the amount of satisfaction alternatives can provide them and how the consequences of dissatisfaction affect preferences.

Let us consider a set N of n agents that will express preferences over a set X of possible alternatives. Each alternative x is objectively described by a set P_x of *features* that represents the satisfaction, or dissatisfaction, of several criteria.

[1] The pairwise majority aggregation method is known to be unanimous, independent to irrelevant alternatives and non-dictatorial.

More precisely, given a set of criteria $\mathcal{C}$, for each criterion $c \in \mathcal{C}$, P_x will either contain p_c (criterion c is satisfied) or $\mathbf{n}p_c$ (criterion c is not satisfied). We say that x *satisfies* p_c if $p_c \in P_x$, otherwise $\mathbf{n}p_c \in P_x$.

Inspired by the work of Dietrich et al. [6] we suppose in this work that agents' preferences are based on *desired features*. In particular, we assume that each agent $i \in N$ has a set W_i of desired features which will induce a preference relation over X, i.e. i will prefer an alternative $x \in X$ over an alternative $y \in X$ if the number of features in W_i satisfied by x is greater than or equal to the number of features in W_i satisfied by y.

Definition 1 (**Features-induced preference formation**). *Given a set W of desired features, two alternatives $x, y \in X$ and their respective set of satisfied properties P_x and P_y, x is preferred to y according to W ($x \succeq_W y$) if and only if*

$$|\{p \in P_x \text{ s.t. } p \in W\}| \geq |\{p \in P_y \text{ s.t. } p \in W\}|$$

If x is preferred to y and y is not preferred to x according to W, then x is strictly preferred to y according to W ($x \succ_W y$). Otherwise, we suppose that x and y are equivalent according to W ($x \sim_W y$).[2] *Given an agent $i \in N$ and her desired features W_i, we will denote by $\succeq_{W_i}$ or $\succeq_i$ her preferences.*

Among these desired features, some are desired by all agents in the group, while others are more personal. Some of these personal features can lead to modifications in the agents preferences which bring the collection of individual preferences (so-called *preference profile*) farther from consensus. Intuitively, the more the agents want personal features, the more heterogeneous their preferences will be and the lower the probability of obtaining a transitive result *via* pairwise majority. Following this idea we will qualify as *unshared* every feature which is not desired by the entire set of agents, the others are considered as *consensual*.

Definition 2 (**Consensual and unshared features**). *Given a set N of agents and, for each agent $i \in N$ its set of desired features W_i, we denote by*

- $W_\forall = \bigcap_{i \in N} W_i$ *the set of* consensual features,
- $W_\exists = \{p \in \bigcup_{i \in N} W_i \text{ s.t. } p \notin W_\forall\}$ *the set of* unshared features.

Hence, it is possible to consider the preferences induced by the consensual features, $\succeq_{W_\forall}$, which correspond to a ranking that can be seen as an approximation of the group's collective preferences. The aim of our work is to provide agents with a means to reach a situation where the Condorcet paradox can be circumvented thanks to a deliberative dialogue. But accounting for every particular case of induced preferences following a deliberation is nearly impossible

[2] Please note that Dietrich et al. [6] suppose that agents can have a preference relation over features and thus they can discriminate between two alternatives satisfying the same number of desired features. Intuitively, the importance given to a feature depends on the context.

Table 1: The six possible rankings for three alternatives.

#	Ranking	#	Ranking
1	$A \succ B \succ C$	4	$C \succ A \succ B$
2	$B \succ A \succ C$	5	$B \succ C \succ A$
3	$A \succ C \succ B$	6	$C \succ B \succ A$

as it depends on what the agents want and their justifications, which can be considered as infinitely diverse.

Hence, we need to introduce a notion that will help to represent disagreement within the group. Disagreement between agents is caused by diverging goals and contradicting means to satisfy them. In particular, due to unshared features, agents might distance themselves from the preference relation induced by the consensual features ($\succeq_{W_\forall}$) by swapping alternatives in this approximate ranking. We will see in Section 4.2 how to formally link these swaps—called *alternative escalations*—to the number of unshared features.

In other words, this notion aims to represent the quality of the deliberation: the smaller the number of alternative escalations, the higher the level of consensus, and the more "decisive" the deliberation has been.

3 Empirical Results

In this section, we will first define the metrics allowing to assess the distance between a given profile of preferences and some kind of idealised preference structure. We next present the experimental setup as well as the results we obtained.

3.1 Profile Distance

Single-peakedness and single-cavedness In 2004, Gehrlein [8] considers a variation of the measure proposed by Niemi et al. [13]. Consider the case of an election with three alternatives $X = \{A, B, C\}$. The individual preferences of the agents on these alternatives are limited to the six rankings in Table 1.

Let n_l be the number of agents whose individual preferences correspond to ranking $\#l$. We therefore have $n_1 + n_2$ equal to the number of agents who ranked the alternative C in last position. Similarly, $n_3 + n_4$ agents ranked B last and $n_5 + n_6$ agents ranked A last. In our case, if one of the three alternatives is never ranked last, then the preference profile will be—according to an order in which this alternative is ranked second—*single-peaked* [15]. We thus define our *measure* of proximity: if there exists a candidate rarely ranked last by the agents, then it is probably a unifying candidate in the sense that very few agents would regard her election as the worst possible result. If there are such candidates, it will be easier to find a Condorcet winner [9].

Definition 3 (**Proximity to single-peakedness**). *Given n agents, 3 alternatives, a preference profile $\mathcal{P}$ and the rankings in Table 1, let n_l be the number of agents whose individual preferences $\succeq_i \in \mathcal{P}$ correspond to ranking $\#l$. The* single-peakedness *proximity measure m_{sp} is defined as*

$$\frac{min(n_1 + n_2, n_3 + n_4, n_5 + n_6)}{|\mathcal{P}|}$$

When the value of the metric is 0, a candidate is never ranked last, so the preference profile is *single-peaked* and there is a Condorcet winner[3]. A trivial upper bound for this measure is $\frac{n}{3}$.

A similar metric m_{sc} can be set up in order to compute the proximity to *single-cavedness*, a mirror property of *single-peakedness* which is also a sufficient condition ensuring the existence of a Condorcet winner [11]. A triplet of alternatives is *single-caved* when there is an alternative that is never ranked first by the agents.

Separability into two groups In 2005, based on the work of Inada [11], Gehrlein [9] proposes another measure, variant of the previous ones. Still in the case of three alternatives, if there is a candidate rarely ranked second by the agents, then it is a polarizing candidate. Indeed, it is either very appreciated by some agents (ranked first), or very little appreciated by others (ranked last). In such a situation, it will be easier to extract a structuring dimension from the individual preferences [9]. We can then, in the same way as before, define m_{sg} as a measure of the proximity to *separability into two groups* of a preference profile. In the case $m_{sg} = 0$, there is a candidate who is never ranked second, the preference profile satisfies the condition of *separability into two groups* and a Condorcet winner exists. The same upper bound of $\frac{n}{3}$ applies to this measure.

Triple wise value restriction The previous measures can be combined to compute a proximity to *triple wise value restriction* which is a generalisation of the three previous conditions introduced by Sen [15].

Definition 4 (**Distance to triple wise value restriction**). *Given n agents, 3 alternatives, a preference profile $\mathcal{P}$ and the rankings in Table 1, let n_l be the number of agents whose individual preferences $\succeq_i \in \mathcal{P}$ correspond to ranking $\#l$. The* triple wise value restriction *proximity measure m_{tw} is defined as*

$$min(m_{sp}, m_{sc}, m_{sg})$$

In Section 3.2, we will use variants of these measures—normalised over all triplets of alternatives—in order to study the consequences that the number of unshared features may have on the proximity of preference profiles to these structural properties.

[3] In this case, the Condorcet winner is the most preferred alternative of the median voter [3].

3.2 Simulation Results

The notion of alternative escalations introduced in the previous section lets us control the *quality* of the deliberation that takes place between the agents. A small amount of alternative escalations after deliberation indicates that the agents managed to reduce the quantity of unshared features. On the contrary, a high amount of alternative escalations indicates that they did not manage to agree on a large amount of desirable features and each agent potentially has a significant amount of residual unshared features. More precisely, we will answer in this section the following question: *does proximity to interesting structural properties increase when the number of unshared features decreases?*

Experimental Settings . The experiment, which aims at simulating a deliberation outcome, is fixed by the following parameters: the number n of agents, the number k of alternatives and the maximum number e of alternatives escalations an agent can do. The experimental protocol is the following. One takes a linear order $\succeq_{W_\forall}$ of k alternatives in order to simulate the preference relation induced by the set $W_\forall$ of features considered as desirable by all the agents after deliberation. Then, each agent can do at most e random alternatives escalations in $\succeq_{W_\forall}$ (by using her residual unshared features). Once the new preferences generated, one computes the proximity of the preference profile to a given preference structure (*single-peakedness*, *separability into two groups*, *triple wise value restriction*): the proximity measure is the ratio between the number of triplets of alternatives satisfying the preference structure and the total number of triplets. A measure of $m_s = 1$ indicates that all the triplets are satisfying the structure $s \in \{sp, sg, tw\}$. On the contrary, a measure $m_s = 0$ indicates that all the triplets are problematic and thus the preference profile is not satisfying the preference structure, which will give rise to a non-transitive result in most cases. In order to treat and harmonise the different cases that we can encounter, each point on the graphs corresponds to an average performed on $10,000$ repetitions.

Single-peakedness . The first proximity measure we want to observe is the one to *single-peakedness*. Figure 1 shows the results of the experiment for $n = 200$ agents. For $k = 3$ to $k = 10$ alternatives, the proximity to *single-peakedness* has been computed according to e.

Proximity to *single-peakedness* increases when e (and thus the number of unshared features) decreases. This result supports the hypothesis that agreeing on features allows agents to restructure their individual preferences in an interesting way. However, this increase in proximity is at different speeds depending on the number of alternatives. Indeed, with three alternatives, it is easy to obtain a non *single-peaked* preference profile by modifying very little the same ranking[4]. On the other hand, with more alternatives, and thus more triplets to consider, more personal modifications from the agents in their new preferences

[4]Case of the Condorcet paradox for example.

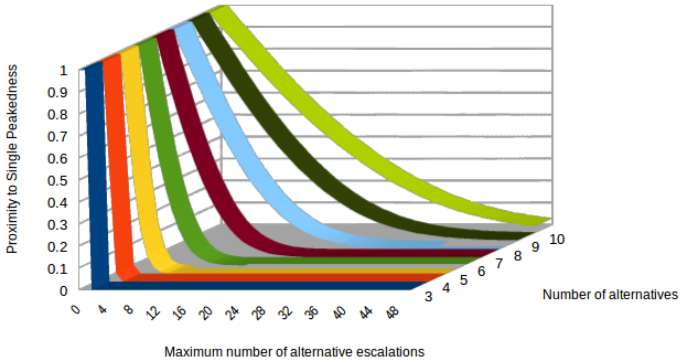

Figure 1: Proximity to *single-peakedness* : 200 agents.

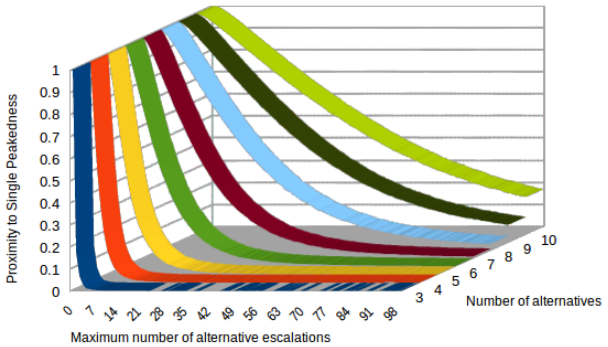

Figure 2: Proximity to *single-peakedness* : 20 agents.

are necessary to move away from *single-peakedness*. Thus, for a fixed number of modifications, the probability of obtaining a good proportion of *single-peaked* triplets increases with the number of alternatives.

Please note that good proximity to *single-peakedness* does not guarantee a transitive result *via* pairwise majority. We can still hope, in these cases, to obtain a Condorcet winner even if the overall ranking of alternatives is not totally transitive. Indeed, even if there is a cycle in the ranking, a Condorcet winning alternative may exist and may dominate this cycle. This is why the search for a unifying candidate is interesting.

Figure 2 shows the results of the same experiment with a set of 20 agents. The proximity to *single-peakedness* decreases less rapidly here. This can be explained by the fact that, for a triplet of alternatives, the probability that it is not *single-peaked* increases with the number of times it has to be considered. The more agents there are, the more unlikely their individual preferences will be *single-peaked*. Deliberation seems therefore to be more efficient with a reduced

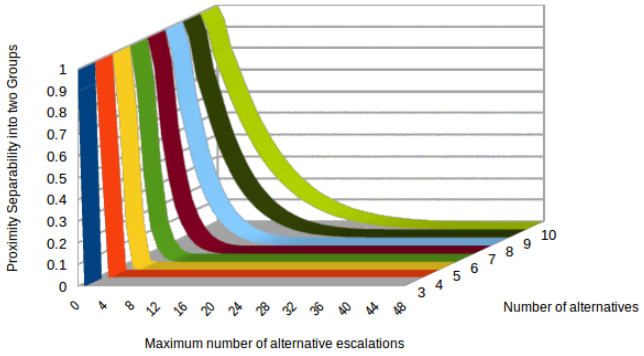

Figure 3: Proximity to *sep. into two groups* : 200 agents.

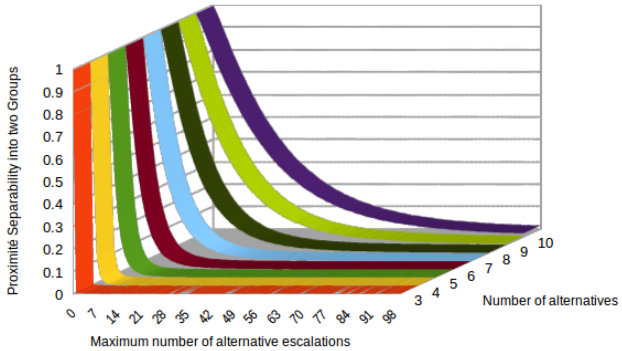

Figure 4: Proximity to *sep. into two groups* : 20 agents.

number of agents. Although intuitive, this result is interesting because the implementation of a deliberation protocol in real life situations seems difficult if it is necessary to consider a large number of agents.

Separability into Two Groups . The second experiment aims to study the proximity to *separability into two groups* as introduced in Section 3.1. The results obtained are given in Figure 3. The general shape of the curves is the same as before, so we can conclude that the proximity to *separability into two groups* increases when e (and therefore the number of unshared features) decreases. However, the proximity value to this property decreases faster than the proximity to *single-peakedness*, this can be explained by the way in which the new individual preferences are generated. The alternatives ranked first and last in $\succ_{W_\forall}$ are half as likely as the others to change their position when performing alternatives escalations (the first one cannot go up and the last one cannot go down). In case the alternative ranked first moves, then it has to go down to the

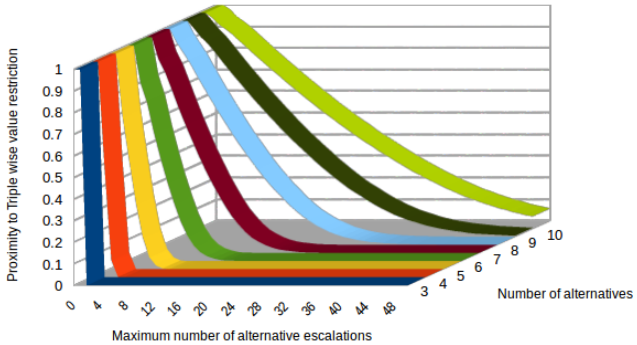

Figure 5: Proximity to *Triple wise value rest.* : 200 agents.

second place, increasing the probability that a triplet of alternatives containing this alternative will not satisfy the *separability into two groups* condition. The same holds for the alternative ranked last in $\succ_{W_\forall}$. The final preference profile will therefore be more likely to be *single-peaked* or *single-caved* than to satisfy the *separability into two groups* condition.

Triple Wise Value Restriction . In the same way as before, we look at the proportion of triplets of alternatives satisfying this condition: checked if the triple is *single-peaked*, *single-caved* or satisfies the *separability into two groups* condition. As expected, the shape of the curves remains the same. Proximity to *triple wise value restriction* increases as the number of residual unshared features decreases. There is also again a rapid drop in the value of proximity when few alternatives are considered. The same experiment was performed for $n = 20$ agents, the results are given in Figure 6. As for the proximity measure to *single-peakedness* (but to a lesser extent since the measure of proximity to *separability into two groups* comes into play), with less agents the proximity decreases less drastically and this regardless of the number of alternatives considered. Another positive observation is the fact that for a large number of alternatives, a significant proportion (around 25%) of triplets of alternatives always satisfy one of the three conditions that make up the *triple wise value restriction* even with a very large and seemingly unrealistic number of residual unshared features.

Now that we observed the critical impact that alternative escalations can have on the deliberation outcome, it is necessary to study means to reduce their amount. In the next section, we will assess how a simple deliberation protocol can be used to achieve this goal.

4 Deliberation Around Unshared Features

In this section, we will first define a simple yet effective deliberation protocol, we will then assess its ability to impact the number of possible alternative

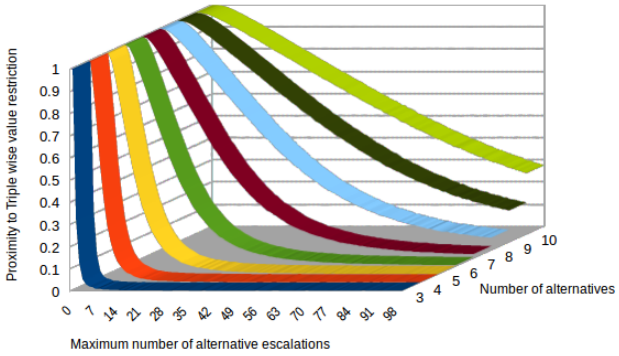

Figure 6: Proximity to *Triple wise value rest.* : 20 agents.

escalations an agent will be able to do once the deliberation is over.

4.1 Protocol Definition

The deliberation protocol will take place in two phases. Before the first phase, each agent $i \in N$ has a preference relation $\succeq_i$ over the alternatives based on (i) her desirable features in W_i and (ii) the *a priori* knowledge she has on the alternatives.

The goal of the protocol is to provide the agents with a way to revise these preferences by refining their set of desired features and the knowledge they have about the alternatives. Intuitively, agents will share knowledge and opinions through arguments. During this process, agents might discover new features that they were not aware of before as well as new arguments leading them to change their opinion on a particular question. Hopefully, this discussion will give agents more insights on the situation and will allow them to refine their preferences, namely their set of desired features and their knowledge about the alternatives. At the end of the second phase, each agent will have modified preferences $\succeq'_i$ over the alternatives which she will use to vote.

During the **first phase**, agents will deliberate to collectively agree upon a—final—set $W'_\forall$ of features, considered as relevant by the group, that they will use in order to choose among the alternatives. Throughout the deliberation, the goal for an agent i is to make her desired features from W_i accepted as relevant by the group. To achieve this goal, agents will assert arguments to justify their preferences. During this first phase, by referring to accepted arguments stored in her *commitment store*, each agent will change her desired features as she gets eventually convinced by other agents' arguments. In this phase, agents will talk in a round robin manner (one after another) but each agent can, if desired, skip her turn. The first phase ends after n successive turn skips. At the end of the first phase, the set $W'_\forall$ of consensual features after deliberation is obtained.

The goal of the **second phase** is to determine, for each alternative $x \in X$,

and at least for each criterion $c \in \mathcal{C}$ such that $p_c \in W'_\forall$ or $\mathbf{n}p_c \in W'_\forall$, whether or not x satisfies the criterion c. Agents will use arguments to justify their position. This process results in an attribution for each alternative $x \in X$ of a set P_x of satisfied features. Based on this outcome, the agents will then revise their preferences over the alternatives by considering their own desired features and the features considered as relevant by the group. Once again, agents will deliberate in a round robin manner and can skip their turn. The second phase ends after n successive turn skip.[5]

The way an agent can revise her preferences is to agree that the features chosen by the group are desirable for her too as she is part of the group. Then, each agent i can merge the features contained in $W'_\forall$ with her own set of desired features W_i constructing thus a new set W'_i of desired features, being careful not to leave opposite features in W'_i. In case of a conflict, the feature extracted from $W'_\forall$ is more relevant than its opposite one initially in W_i. The new preference relation of agent i ($\succeq'_i$) is then induced by W'_i and the collective knowledge gained on the alternatives.

Definition 5 (**Desired properties after deliberation**). *Let N be a set of agents, each agent i in N has a set of wanted properties W_i. Let $W'_\forall$ be the set of consensual features obtained after deliberation. For all i in N, the new wanted features of i are defined as follows :*

$$W'_i = (W_i \cup W'_\forall) \backslash (\{p \in W_i \mid \boldsymbol{n}p \in W'_\forall\} \cup \{\boldsymbol{n}p \in W_i \mid p \in W'_\forall\})$$

Let us formally define how the agents can interact with each other. In order to get a desired feature or a piece of knowledge accepted as relevant by the group, agents must justify their claims. To achieve this goal, arguments will be used. Classically, in this paper, we assume that agents possess a logical language allowing the construction of arguments. Here, an argument is a triple containing a set of premises, a set of rules and a conclusion which is derived from the premises using the rules [7, 2]. Agents can use arguments for different purposes according to possible actions. We define these actions as a set of possible speech acts [14]. This set contains the following locutions:

- ASSERT(.) :
  - **Meaning :** an agent uses this locution to formally prove a claim.
  - **Usage :** ASSERT(i, arg) where i is in N and arg is an argument whose conclusion is the statement agent i wants to prove.
- REJECT(.) :
  - **Meaning :** an agent uses this locution to formally reject a statement another agent made before. To achieve this goal, the agent uses an argument which proves that the statement she wants to reject is false.

[5] Please note that for simplicity purposes, we assume here that both phases always succeed, *i.e.* all the agents manage to agree on a set of desired features and on the features satisfied by the alternatives.

- **Usage :** REJECT(i, j, F, arg) where i and j are in N and F is a set of premises previously used by j to prove a statement. Note that the conclusion of the argument arg and the rejected premises F are logically incompatible.

- CHALLENGE(.) :
  - **Meaning :** an agent uses this locution to ask another agent to justify some premises she used in order to prove a statement.
  - **Usage :** CHALLENGE(i, j, F) where i and j are in N and F are premises used by j in order to prove a statement.

- RETRACT(.) :
  - **Meaning :** an agent uses this locution when she is unable to justify some premises she used to prove a statement.
  - **Usage :** RETRACT(i, arg) where i is in N and arg is an argument using premises i is unable to prove.

- CONCEDE(.) :
  - **Meaning :** an agent uses this locution to explicitly accept a statement made by another agent.
  - **Usage :** CONCEDE(i, j, arg) where i and j are in N and arg is an argument previously asserted by j whose conclusion is accepted by i.

These locutions allow agents to justify their positions. We will now see how they can use it to deliberate constructively.

Reply Structure In order to maintain a coherent dialogue, agents have to use locutions in a constructive way. Speech acts are subject to a particular reply structure ensuring that each locution is used for a correct purpose. This reply structure is described in Table 2. We can see, for example, that the RETRACT(.) locution might be used by an agent to respond to a CHALLENGE(.) locution. This CHALLENGE(.) locution, in turn, might be used by someone else to ask for a justification about a claim made earlier using the ASSERT(.) or the REJECT(.) locution.

Correctness Conditions In order to avoid misuses of the locutions, they are subject to a set of conditions which comply with the reply structure. The conditions that must be satisfied in order to use a specific locution are listed in Table 3. These conditions ensure that agents deliberate in a focused manner.

Table 2: Locutions and their respective attacks and surrenders.

Locutions	**Attacks**	**Surrenders**
ASSERT(.)	REJECT(.) CHALLENGE(.)	CONCEDE(.) RETRACT(.)
REJECT(.)	CHALLENGE(.) REJECT(.)	RETRACT(.)
CHALLENGE(.)	ASSERT(.)	RETRACT(.)
RETRACT(.)	$\emptyset$	$\emptyset$
CONCEDE(.)	$\emptyset$	$\emptyset$

Commitment Store Effects Due to the previous conditions, each action performed by an agent is done for a particular purpose. In order to track the effects of these actions on the dialogue, the use of some speech acts is subject to post conditions which imply commitment store modifications. For each agent $i \in N$, $CS(i)$ is the set of arguments that i has explicitly accepted, her *commitment store*. The effects of each locution on agents' commitment stores are shown in Table 4. By applying these modifications, each agent can know, at any time, the status of her preferences. This is necessary as agents have to revise their preferences after the deliberation ends.

Now that the deliberation protocol has been formally defined, we can assess its quality for reaching consensus by characterising its impact on the number of possible alternative escalations agents will be able to do.

4.2 Impact of the Protocol on the Number of Possible Alternative Escalations

At the end of the deliberation process, each agent has a new set of desired features built according to Definition 5. As a consequence, we can make a first trivial observation: if there is a consensus on which features are desirable and as they all share the same knowledge over the alternatives, then all the agents will have the same preferences and the application of pairwise majority will always give a transitive result.

Lemma 1 (**Absence of unshared features**). *When deliberation ends, if no agent desires an unshared feature then the result of pairwise majority on the induced agents' preferences is transitive.*

We have seen that absence of unshared features benefits transitivity, it is thus interesting to track their evolution during the deliberation phase. The Lemma 1 guarantees a transitive result in case of consensus. The following result ensures that the deliberation process allows agents to move closer to such consensus. The proof consists in verifying that (i) an *unshared* feature can only become a *consensual* one and (ii) a *consensual* feature cannot become an *unshared* one.

Table 3: Locutions and their using conditions.

Locutions	**Using conditions**
ASSERT(i, arg)	The arg argument was never asserted before.
REJECT(i, j, F, arg)	The rejected premises in F were used by j to prove some statement that i is rejecting. The conclusion of argument arg and the premises in F are logically incompatible with respect to the logical language considered.
CHALLENGE(i, j, F)	The challenged premises in F were used by j to prove some statement that i has not yet accepted.
RETRACT(i, arg)	The argument arg was asserted by i.
CONCEDE(i, j, arg)	The argument arg was asserted by j.

Lemma 2 (**Diminution of unshared features**). *The number of unshared features can only decrease during the deliberation.*

At the end of the deliberation, the agents agree on a set $W'_\forall$ of desirable features for the group that induces a preference relation $\succeq_{W'_\forall}$. Then, the unshared features of each agent $i \in N$ will let her modify $\succeq_{W'_\forall}$ in order to obtain her new individual preferences $\succeq'_i$. Let us try to identify the *alternative escalations* that an agent can perform on $\succ_{W'_\forall}$ according to her residual unshared features. Let $x, y \in X$ be two alternatives such that $x \succ_{W'_\forall} y$ and $i \in N$ an agent. In order to obtain $y \succ'_i x$ by changing $\succeq_{W'_\forall}$, i needs a certain number of unshared features. Let us suppose that x satisfies exactly one more feature of $W'_\forall$ than y. Then, i will strictly prefer y over x only if there exist two unshared features (not in $W'_\forall$) p_1 and p_2 in W'_i desired by i such that y satisfies both and x satisfies none. One can generalise this result for scenarios in which the difference $di(x, y)$ of number of features in $W'_\forall$ satisfied by x and y is greater than 1. In such cases, at least $di(x, y) + 1$ unshared features satisfied by y but not by x must be desired by i to obtain $y \succ'_i x$.

Based on this observation, we can now use Lemma 2 to obtain the following theorem which links the *alternative escalation* parameter used in the experiments of Section 3.2 to the number of *unshared features*.

Theorem 1 (**Maximum number of possible modifications**). *Given Max the maximum value of $di(x, y)$ among all pairs of alternatives $(x, y) \in X^2$, an*

Table 4: Locutions and their effects on Commitment Stores.

Locutions	**Commitment store effects**
ASSERT(i, arg)	Add arg to $CS(i)$.
REJECT(i, j, F, arg)	Add arg to $CS(i)$.
CHALLENGE(i, j, F)	-
RETRACT(i, arg)	Remove arg from $CS(i)$
CONCEDE(i, j, arg)	Add arg to $CS(i)$.

agent $i \in N$ with u_i unshared features will be able to do at most $\lfloor \frac{u_i}{Max+1} \rfloor$ alternative escalations in $\succeq_{W'_\forall}$.

Hence, during the deliberation, the more the agents are able to bring effective arguments, the less they will get away from the preference relation $\succeq_{W'_\forall}$ induced by $W'_\forall$, which will bring them closer to consensus.

5 Discussion

In this paper, we tried to answer the following question: *is it possible to assess formally the possibility of defining a deliberation protocol moving agents' preferences closer to particular structures (single-peakedness, triple wise value restriction, etc.) ensuring a transitive result under pairwise majority?* We started by defining agents' preference formation based on the notion of *desired features* and used the notion of *alternative escalations* to represent how agents might diverge from the preference relation induced by the group's desired features. Using these notions, we proposed an experimentation showing that less alternative escalations leads to agents' preferences being closer to useful preference structures. Finally, we defined a simple deliberation protocol and characterised it in terms of its impact on the number of possible alternative escalations. While the presented work answer our initial problem, it raises many other questions. We present them in the following paragraphs.

Preferences formation Although the results of the experiment seem to confirm the hypothesis that the deliberation protocol improves the agents' preferences structure[6], we observed that our simulation choices have some consequences. In particular, the way in which individual preferences are generated through alternative escalations impacts the *separability into two groups* measure. In our experiment, agents are assumed to be completely independent of each other and to perform alternative escalations in a random way. This way to deal with agents is reminiscent of how one can generate individual preferences

[6]This observation is in line with the experimental results obtained by List et al. [12] in 2012.

under the impartial culture assumption [10]. Unfortunately, impartial culture is known to be unrealistic and seems likely to maximise the probability of obtaining majority cycles [16]. For these reasons, it would be interesting to study other ways of generating preferences and disagreements between agents.

Deliberation protocol We deliberately chose to leave aside the argumentation part of the deliberation protocol. However, considering the argument exchange part it is necessary to decide on several points. For instance, what would happen if two agents have rational justifications for opposite features? Or if they desire the same feature but with contradictory justification? Considering argumentation systems [7] during the deliberation could allow the resolution of such conflicts and help agents deciding which justifications should be taken into account.

Measures generalisation In the experiment carried out, we have chosen to generalise the various measures of proximity to the whole set of triplets of alternatives by observing the proportion of triplets satisfying the desired condition. Several reasons motivated this choice, such as the fact that for extreme values the measure remains consistent (a measure of 1 means that the condition is true for the preference profile as a whole, a measure of 0 means that the condition is absolutely unverified).

That being said, it would be interesting to consider other approaches to measure the efficiency of the deliberation, for instance by studying the link between the distribution of unshared features and the probability of obtaining a non-transitive result using pairwise majority.

Real-life situation Finally, setting up a real-life experimentation would allow to focus on other aspects of deliberation. Indeed, in addition to confirming or refuting the experimental results of this work, it would let us assess to which extent people are able to identify desirable features and to defend them using rational arguments.

References

[1] Arrow, K.J.: Social choice and individual values. Wiley (1951)

[2] Besnard, P., Hunter, A.: A logic-based theory of deductive arguments. Artificial Intelligence **128**(1-2), 203–235 (2001)

[3] Black, D.: The median voter theorem. The Journal of Political Economy **56**(1), 23–34 (1948)

[4] Brandt, F., Conitzer, V., Endriss, U., Lang, J., Procaccia, A.D.: Handbook of Computational Social Choice. Cambridge University Press, New York, NY, USA, 1st edn. (2016)

[5] de Condorcet, M.J.A.N.C.M.: Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. L'imprimerie royale (1785)

[6] Dietrich, F., List, C.: Where do preferences come from ? International Journal of Game Theory **42**(3), 613–637 (2013)

[7] Dung, P.M.: 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)

[8] Gehrlein, W.V.: Consistency in measures of social homogeneity: A connection with proximity to single peaked preferences. Quality and Quantity **38**(2), 147–171 (2004)

[9] Gehrlein, W.V.: Probabilities of election outcomes with two parameters: the relative impact of unifying and polarizing candidates. Review of Economic Design **9**(4), 317–336 (2005)

[10] Guilbaud, G.T.: Les théories de l'intérêt général et le problème logique de l'agrégation. Économie appliquée **5**, 501–584 (1952)

[11] Inada, K.i.: A note on the simple majority decision rule. Econometrica: Journal of the Econometric Society pp. 525–531 (1964)

[12] List, C., Luskin, R.C., Fishkin, J.S., McLean, I.: Deliberation, single-peakedness, and the possibility of meaningful democracy: evidence from deliberative polls. The journal of politics **75**(1), 80–95 (2012)

[13] Niemi, R.G.: Majority decision-making with partial unidimensionality. American Political Science Review **63**(2), 488–497 (1969)

[14] Searle, J.R.: Speech acts: An essay in the philosophy of language, vol. 626. Cambridge university press (1969)

[15] Sen, A.K.: A possibility theorem on majority decisions. Econometrica: Journal of the Econometric Society pp. 491–499 (1966)

[16] Tsetlin, I., Regenwetter, M., Grofman, B.: The impartial culture maximizes the probability of majority cycles. Social Choice and Welfare **21**(3), 387–398 (2003)