UNIVERSITY OF AMSTERDAM

MASTER THESIS

THIS THESIS IS A WORK IN PROGRESS

Modelling Meta-Agreement through an Agent-Based Model

Author:
Amir Sahrani

Examiner:
Dr. Fernando P. Santos
Supervisor:
Prof. Dr. Ulle Endriss
Assessor:
Dr. Davide Grossi

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computational Science

in the

Computational Science Lab
Informatics Institute

June 10, 2025

Declaration of Authorship

I, Amir Sahrani, declare that this thesis, entitled 'Modelling Meta-Agreement through an Agent-Based Model' and the work presented in it are my own. I confirm that:

- □ This work was done wholly or mainly while in candidature for a research degree at the University of Amsterdam.
- □ Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- □ Where I have consulted the published work of others, this is always clearly attributed.
- □ Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- □ I have acknowledged all main sources of help.
- □ Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Your signature

Date: June 10, 2025

“The majority, standing in for the people, wills everything and therefore wills nothing”

Joshua Cohen

Abstract

Include your abstract here Abstracts must include sufficient information for reviewers to judge the nature and significance of the topic, the adequacy of the investigative strategy, the nature of the results, and the conclusions. The abstract should summarize the substantive results of the work and not merely list topics to be discussed.

Length 200–400 words.

Acknowledgements

Thank the people that have helped, supervisors family etc.

Contents

List of Figures

List of Tables

List of Algorithms

ABBREVIATIONS

PBS **P**olicy-**B**ased (ideology) **S**core

PsT **P**roximity **t**o **S**ingle-peakedness

PsT-V **P**roximity **t**o **S**ingle-peakedness (through) **V**oter (deletion)

PsT-C **P**roximity **t**o **S**ingle-peakedness (through) **C**andidate (deletion)

SYMBOLS

N	The set of all voters
X	The set of all alternatives
$\succ$	A preference relationship
$\mathcal{D}$	A domain of possible profiles
D	A deterministic deliberative procedure
DB	A deliberative procedure with biased voter
$\mathcal{L}(A)$	Set of all possible preference order over A
R	Set of a preference relations over all candidates
$\boldsymbol{R}$	Set of preferences of all voters
f	A function mapping a strict profile to a candidate
$\triangleleft$	A geometric order over candidates
Ψ	Vector of all policies
ψ	An instance of a policy
S	Vector of support for each policy
Σ	matrix of shape $\lvert A\rvert \times \lvert\Psi\rvert$, estimating support of policies for each alternative

CHAPTER 1

INTRODUCTION

CHAPTER 2

PRELIMINARIES

We first proceed by giving a short introduction of social choice. We outline the basic model, and restate well known results relevant to the following chapters.

2.1 The basic model

To model elections, or more generally voting games, we represent voters by the set N consisting of n voters. The possible outcomes of an election, we represent with the set A consisting of $|A|$ possible outcomes, from now on we will refer to the outcomes of an election as alternatives. Each voter can represent their preference on alternatives through a preference relation $\succcurlyeq_i$, for example if voter 2 prefers outcome a to outcome b, we write $a \succcurlyeq_2 b$. If, however, this voter wants to make it clear a is strictly better than b, we instead write $a \succ_2 b$. When a voter specifies their preferences on the entire set of alternatives we call this a (weak) linear order. We call the set of possible linear orders over the alternatives $\mathcal{L}(A)$, the set of weak linear orders is denoted by $\hat{\mathcal{L}}(A)$. Thus, for an election, all voters report a (weak) linear order, the set of each voters preference is called a profile, denoted by $\boldsymbol{R}$. Finally, a rule f decides the outcome of the election based on the profile. We discuss the specifics of these rules in Section section 2.2.

The last general tool we will need is the *majority relation*. Given some profile $\boldsymbol{R}$ we can construct a majority relationship as follows: for each pair of alternatives x, y, we ask how many people prefer x to y; if the number of people who prefer x to y is greater than the other way around we write $x \succ_{\text{maj}} y$, if we have an even number of voters, these two number can be equal and this becomes a weak preference, we simply write $x \succcurlyeq_{\text{maj}} y$ (defaulting to lexicographical order). We proceed with an example.

EXAMPLE 1: *Majority relation*

1	2	3
a	b	a
b	c	c
c	a	b

Given the profile on the left, we first start by comparing a to b, both voters 1 and 3 prefer a to b, thus the majority has prefers a to b. Comparing b to c,the majority prefers b to c. Finally, comparing a to c, a is preferred again. Thus, the majority relation is $a \succ_{\text{maj}} b \succ_{\text{maj}} c$

A majority relation can be a-cyclic, and transitive, though neither are guaranteed. An a-cyclic majority profile is simply a majority relation without any cycles, meaning there does not exists a series of alternatives $a_1, \ldots a_n$ such that $a_1 \succ a_2 \succ \ldots \succ a_n \succcurlyeq a_1$. Transitivity is very similar, stating that the preferences between alternatives are transitive in that for any triplet of alternatives x, y, z if $x \succcurlyeq y$ and $y \succcurlyeq z$ then $x \succcurlyeq z$. These notions are similar, but transitivity is a stronger requirement, as it includes indifference.

2.2 Social Choice Functions

In order to decide the outcome of an election, we pick a social choice function f, this function should map all possible profiles to an outcome, thus $f : \boldsymbol{R} \to A$. A famous example of a SCF is the plurality rule, which simply elects the alternative voted into first place most often, though simple, it can also lead to a tie. Since the outcome of our SCF is only allowed to be a single alternative, the plurality rule needs to be equipped with a tie breaking mechanism in order to be a valid SCF, we require the tie-breaking to be deterministic.

2.2.1 Axioms

Though any voting rule that outputs one alternative for each profile is valid, for real elections organizers likely will want to ensure the rule has certain nice properties, such as not favoring some alternatives of others. In social choice these properties are called axioms, and the procedure of designing a rule based on axioms is called the axiomatic approach. The name of the property just described is the axiom of neutrality, stating that the voting rule should be neutral with respect to the alternatives. In this work three main axioms are of importance.

Axiom of Surjectivity. A SCF f is surjective, if for every alternative, there exists a profile R such that $f(R)$ elects it.

Axiom of Non-Dictatorship. A SCF f is non-dicatorial, if there does not exist a voter i such that $f(R) = \text{top}(i, R)$ for all profiles R, where $\text{top}(i, R)$ extracts voter i's most preferred alternative from profile R.

Axiom of Strategyproofness. A SCF f is strategy proof if, for any voter $i \in N$, i cannot report an untruthful preference, and thereby cause the outcome of the elective to improve for them.

Axiom of Anonymity. A SCF f is anonimous if, when the labels of voters are shuffled, the winning alternative stays the same.

Axiom of Neutrality. A SCF f is neutral if, when the labels of the alternatives are shuffled, the winning alternative is the alternative who is ranked the same by each voter as the original winning alternative.

Another way to interpret strategyproofness is that the SCF should maximize the outcome for all voters, as such if a voter reports something which is not their true preference, the outcome will maximize the wrong preference and thus result in an outcome that is worse for you.

There are many more axioms on could reasonably argue for, however, these are enough to lead to the main impossibility result this works focuses on.

2.3 Negative results

Classic social choice theory has many negative results, one such example is the Condorcet cycle. This is a specific profile that results in a cycle in the majority relation, as shown in the following example.

EXAMPLE 2: *Condorcet cycle*

1	2	3
a	b	c
b	c	a
c	a	b

Voters 1 and 3 prefer a to b, forming a majority, next voters 1 and 2 prefer b to c, forming another majority. However, voters 2 and 3 prefer c to a forming a majority, and thus creating a cycle.

It is not hard to convinces oneself that under weak preferences the Condorcet cycle can occur anytime there are 3 or more alternatives and voters. While under strict preferences this can occur anytime the number of alternatives is odd and greater than 3, with the number of voters being a multiple of the number of alternatives. As we will show later, this profile can be the cause of some impossibility results.

One of the major negative results in social choice is that of the Gibbard Satherswaite theorem [1, 2].

Theorem 2.1. [Gibbard-Satherswaite] There exists no resolute Social Choice Function for elections with $|A| \geq 3$ that is surjective, strategyproof, and non-dictatorial.

Put plainly, it is impossible to have a voting rule that incentivizes voters to report their preferences truthfully, when we also want at least three candidates to be able to win, unless we accept a dictatorship.

2.4 Domain Restrictions

Many negative results are a consequence of a few ill-behaved profiles, if one can argue such profiles do not occur in the real election, there is some hope of constructing SCF's satisfying our axioms. To speak more formally about profiles "not occuring", we introduce Domain restrictions, for this we use the definition by Elkind et al. [3].

DEFINITION 1: *Domain*

Given a set of voters N, alternatives A, and conditions C, the domain $\mathcal{D}$ of an election is the set of all profiles R such that all conditions C are satisfied.

This definition is different from usual definitions in social choice in so far as it talks about allowed profiles instead of allowed votes.

As stated earlier, the Condorcet profile is one such ill-behaved profile, as each alternative, holds a majority preference over another alternative. Naturally one might consider if this profile might even come up in practice, since though conceivable it seems generally unlikely that there exists a perfect split in opinions. Quite naturally one of the first "solutions" one might consider is when the number of voters is not a multiple of the number of alternatives, though this is hardly a useful solution since it only prevents Condorcet cycles, it is the first example of a domain restriction, we define it as follows

DEFINITION 2: $\mathcal{D}_{\text{No-tie}}$

Let X be the set of alternatives and N be the set of voters, of size n such that $n \neq k \cdot |X|$ for any $k \in \mathbb{N}$. We call this domain $\mathcal{D}_{\text{No-tie}}$.

This allows us to state our first proposition.

Proposition 2.2. The plurality rule never returns a $|X|$-way tie between alternatives when applied to $\mathcal{D}_{\text{No-tie}}$

Proof. Assume, for the sake of contradiction, the plurality in fact does return a tie this must mean that all alternatives were ranked first an equal number of times, call this k, necessarily then, we have need exactly $k \cdot |X|$ voters, but this leads to a contradiction, as this would no longer be inside $\mathcal{D}_{\text{No-tie}}$. □

This is a simple result, but it serves as an example on how we can use the properties of the domain to prove things about the election. Gaertner [4] establishes 2 ways in which a domain can be restricted. Firstly we can restrict the domain to a number of voters or alternatives, which is what we did in $\mathcal{D}_{\text{No-tie}}$. Secondly, the domain can be restricted to have a certain structure, such as being single-peaked.

2.4.1 Single-Peaked profiles

In a election the alternatives might represent a axis, such that a voters is prefers an alternative more if they are closer to them on the axis. For example, if the alternatives represent free-trade vs regulation, we can imagine that a voter that is of the opinion that free trade is of ultimate importance will prefer alternatives more the more the are on the side of free trade. More generally, we call a profile single-peaked if there exists an axis on which we can place the alternative such that all voters' preferences have a single peak on this axis. Definition 3 makes this notion formal.

DEFINITION 3: *Single-Peaked Profiles*

A profile P is single-peaked, if given some ordering $\triangleleft$ over the alternatives, it holds that for all voters i, and all $a, b, c \in X$, if $a \triangleleft b \triangleleft c$, then either $a \succ_i b$ or $c \succ_i b$, but never both.

CHAPTER 3

LITERATURE REVIEW

In this chapter we explore previous results, as well as introducing relevant concepts.

3.1 Condorcet Domain

If our goal is to prevent Condorcet cycles, or in general have transitive majority relations, the best we could hope to do is to apply our domain restriction such that our domain contains all profiles P such that P has a (weak) Condorcet winner. We call this domain $\mathcal{D}_{\text{Condorcet}}$. Under this domain, let $f_{Condorcet}$ be the Condorcet Rule, which picks a Condorcet winner. Then $f_{Condorcet}$ is strategyproof over $\mathcal{D}_{\text{Condorcet}}$ [3].

> *Proof*. (Elkind et al. [3]). Assume, for the sake of a contradiction, we have profiles $P = (\succ_1 \ldots \succ_i \ldots \succ_n)$ and $P' = (\succ_1 \ldots \succ_{i'} \ldots \succ_n)$ such that:
>
> $$f_{Condorcet}(P) = a, \quad f_{Condorcet}(P') = b, \quad \text{and } a \neq b$$
>
> Then under P a strict majority $N' \subseteq N$ have $a \succ b$, but $i \notin N'$, thus in P', N' is still a majority preferring a to b, but this is in contradiction to b winning in P'. □

This result is strengthened by Campbell and Kelly [5, 6], showing that for an odd number of alternatives, $f_{\text{Condorcet}}$ is the only voting rule over $\mathcal{D}_{\text{Condorcet}}$ that is Strategyproof, Surjective and Non-dictatorial.

When Surjectivity is strengthened to Neutrality, and Non-dictorship to Anonymity, $f_{\text{Condorcet}}$ is the only Strategyproof voting rule over $\mathcal{D}_{\text{Condorcet}}$ for an odd number of voters [7].

3.1.1 Hereditary Domains

Though this result is positive, we might wonder how stable it is, for this we need to define a notion of stability. On natural way to think about it is as follows: suppose one of the alternatives or voters drops out, do we keep the nice structure of the domain? If this is true we call a domain *Hereditary*.

DEFINITION 4: *Hereditary* (Elkind et al. [3])

A domain restriction onto $\mathcal{D}$ is *hereditary* if, for every profile $P \in \mathcal{D}$, and every profile P', that can be obtained by deleting voters and alternatives from P, P' is also in $\mathcal{D}$

$\mathcal{D}_{\text{Condorcet}}$ is not hereditary, this is easy to see through an example:

EXAMPLE 3: $\mathcal{D}_{\text{Condorcet}}$ *is not hereditary*

v_1	v_2	v_3	v_4
a	b	c	a
b	c	a	c
c	a	b	b

We can see that in this example, a is the weak Condorcet winner, as it beats b and is tied with c, however if we remove voter 4, we return to the original Condorcet cycle.

A domain not being hereditary means that the nice properties of the domain can be unstable, as the number of voters and alternatives might not be known or could be manipulated. Instead, we might want to look at hereditary domains. We present the single-peaked domain, will also be the main focus of this thesis. This is the domain of all single-peaked profiles.

Proposition 3.1. (Elkind et al. [3]). $\mathcal{D}_{\text{SP}}$ is hereditary.

Proof. (Voter Deletion). If we remove a voter, this does not affect the other voters, thus the profile is still single-peaked. ✓

(Alternative Deletion). Consider any voter i and their single-peaked vote, if we remove some alternative x, to this voter all alternatives which voter i preferred to x stay in the same position, while all other alternatives move up one rank, thus preserving the order, and single-peakedness. ✓ □

As the goal is to ensure we find ourselves in nicely structured domains, we need some mechanism through which we can ensure this is the case. Deliberation is one possible

pathway towards this. We will now provide a brief overview of the literature surrounding deliberation.

3.2 The History of Deliberation and Meta-Agreement

We have provided an overview of different domain restrictions and their properties, showing they avoid Condorcet cycles. **?**] argue however, that Condorcet cycles are empirically rare. The next section is dedicated to explaining how deliberation might explain this is so through examining the historical ideas around deliberation and deliberative democracy, as well as that of Meta-Agreement.

3.2.1 Deliberation

Though deliberation is intuitively familiar, namely the process of multiple people talking through a problem with the goal of coming to an agreement, compromise or solution. Providing a definition that is both clear and consistent with the literature in Political Science, Philosophy and Social choice is difficult. As this intuition leaves some of the reasons for and goals of deliberation, as stated in the literature, unmentioned.

Freeman [8] gives an overview of deliberative democracy, in which he shares the intuitive idea that a deliberative democracy contains open discussion, open legislative deliberation and a pursuit of the common good. He also notes that there is no common agreement on the central features of a deliberative democracy, one account is that of deliberative democracy simply involving discussion among the public before voting. Another similar account is that this voting must not only be preceded by deliberation, but also general communication, all of which intended to change people's preferences. He further proceeds to give a more detailed conception of deliberative democracy, according to which a deliberative democracy is one in which political agents or their representatives

1. Aim to collect, deliberate and vote
2. Represent their sincere and informed judgements
3. Vote and deliberate on measures beneficial to the common good on the citizens
4. Are seen and see each other as political equals
5. Have Constitutional rights and their social means enable them to participate in public life
6. Are individually free, such that they have their own freely determined conceptions of the good
7. Have diverse and disagreeing conceptions of the good

8. Recognize and accept their duty as democratic citizens, and do not engage in public argument on the basis of their particular moral views incompatible with public reason.
9. Agree reason is public, in so much as it is related to and advances common interests of citizens
10. Agree that their common interest lies primarily in freedom, independence and equal status as citizens.

Firstly, why suddenly talk about deliberative democracy? how is this different from deliberation. Secondly, does this imply that this is already the case? Or should we aim to achieve a deliberative democracy?

These features allow us to be more precise when we talk about a deliberative democracy, and in turn be more careful about what deliberation must entail. Cohen [9] further argues that deliberation is needed for democratic legitimacy. By this he means that without deliberation, a democracy is simply the will of the majority, but since majority rule is unstable, it is simply a reflection of the particular institutional constrains at the time. He further goes on to describe the *ideal deliberative procedure* as follows

What does it mean to be unstable in this context? Elaborate on "particular institutional constraints"

1. Ideal deliberation is *free*, participants regard themselves as only bound by the results of the deliberation, and the preconditions thereof. Participants act in accordance with the decision made through deliberation, and it being agreed on is sufficient reason to do so.
2. Ideal deliberation is *reasoned*, parties are required to state their reasons for advancing proposals.
3. In ideal deliberation, parties are *equal*, both formally and substantively. There are no rules that single individuals out, and existing distributions of power to no lend a party the opportunity to contribute to deliberation.
4. Ideal deliberation aims to arrive at *consensus*, which can be rationally defended.

3.2.2 Meta-Agreement

Consensus, sometimes referred to as substantive agreement, then seems like a natural goal for deliberation. Elster [10] argues that this is not only the goal, but through unanimous agreement this process completely replaces voting, thereby circumventing Arrow's impossibility theorem: "Or rather, there would not be any need for an aggregation mechanism, since a rational discussion would tend to produce unanimous preferences." (p.

112). Though it would be desirable to circumvent Arrow's impossibility theorem, in practice people, even after deliberation, might not and indeed often do not come to full substantive agreement. List [11] instead proposes another perspective on deliberation based on Meta-Agreement

Under *Meta-agreement* individuals do not need to agree on their most preferred outcome, instead they only need to agree on the dimensions of the problem. To contrast this with Substantive-agreement, under which individuals do not need to conceive of the problem in the same way, all they need is to agree on the same outcome. This means that under substantive agreement, voters can agree outcome $a \succ b$ for different reasons, while under Meta-Agreement, if voters disagree on $a \succ b$ it must be for the same reason.

According to List [11] there are three hypotheses that need to be satisfied for deliberation to induce meta-agreement:

- D1 Deliberation leads people to discover a single *issue*-dimension
- D2 Deliberation lets people place all possible alternatives in this *issue*-dimension
- D3 After this deliberation, people update their preferences by picking a preferred outcome, and all other rankings are based on the distance to this outcome in the *issue*-dimension

All these are necessary conditions for *meta-agreement*, from this is it also clear to see that, given that there is exactly 1 *issue*-dimension, single-peaked profiles are, by definition, a direct consequence. This is the main reason meta-agreement is desirable, as it lets us circumvent the Gibbard-Satterthwaite theorem [1, 2] through restricting the domain of preference profiles to the single-peaked domain $\mathcal{D}_{\text{SP}}$

List et al. [12] provide empirical evidence for this theory of deliberation, showing deliberation increases proximity to single-peakedness through voter deletion (PtS-V), which they define as $S = \frac{m}{n}$ where $n = |N|$ and m is the largest subset of voters such that their profile is single-peaked. Furthermore, they also introduce the notion of salience, which represents to what extent a topic is salient in the voting population. In order to test whether deliberation increases single-peakedness *through* meta-agreement, they test the following four hypotheses: (H1) deliberation increases PtS-V. (H2')[1] high salience issues show less increase in PtS than low salience issues. (H3) Effective deliberation, in the sense that more is learned during deliberation, results in bigger increases of PtS. (H4) All things equal, the increase is largest for issues with natural *issue*-dimensions. They find support for all these hypothesis, showing that on low-moderate salience issues PsT increases following deliberation.

[1] This is a test for a corollary. H2 states that the rate of increase of PtS-V decreases. This is not experimentally testable, however since high salience means some sort of deliberation has happened before, they expect this to approximate this affect.

It is important to note that these claims simply predict what will happen, there is not much explanatory power to these claims. Little is known about to process be which voters signal the issue dimensions, nor how they decide on which ones to present.

TODO Read the paper again Furthermore, Ottonelli and Porello [13] show meta-agreement to be a stronger requirement than it may seem at a first glance. Firstly for (D1) to hold, the *issue*-dimension must hold some semantic meaning, as it is unclear how people can exchange conceptualization of the problem otherwise. Furthermore, the issues must consist of 2 semantic issues, with only 1 issue voters simply reach substantive agreement. A further restriction on these two dimensions is that they need to be opposite, with opposite justifications. If this is not the case, a voter can agree with both justifications, and thereby introduce a new dimension "balance", which then violates the conditions under which single-peaked profiles guarantee the existence of fair, strategyproof voting rules. D2 requires that all voters share the exact same semantic understanding of the dimension, and the outcome associated with each alternative. Finally D3 requires D1 and D2 to have happened before in order. D3 seems to be the easiest hypothesis to satisfy.

Thus, meta-agreement as a means for single-peaked profiles is still quite restrictive, needing multiple forms of unanimity, and only applying to problems with certain properties. Nonetheless, meta-agreement could still provide explanatory power to deliberation.

3.3 Related Work

Rad and Roy [14] model deliberation and its effect on single-peakedness, though they argue single-plateauedness is a more accurate term. To this end, they model deliberation as the process of all voters announcing their preferences, and all other voters updating their current preference towards that of the announced preference, in doing so they might have a bias towards their own preference, as such they try to minimize the distance between their current preference and the announced one. This process repeats until all voters have announced their opinion once, which constitutes one "round" of deliberation. The preference a voter adopts when updating must lie between their current profile and the announced profile, which profiles are considered to be "between" is defined by the distance metric used. They considered three distance metrics, Kemeny-Snell (KS) [15], Duddy-Piggins (DP) [16], and Cook-Seiford (CS) [17]. Both KS and DP depend on the judgement set resulting from the voters preferences, which is contains, for each pair of alternatives a, b, where $a \neq b$, a proposition $(a \succ b)$ or $\neg(a \succ b)$. The KS distance is then defined as the number of binary swaps a judgement set needs to undergo before it becomes the target judgement set, an example for such a swap would be

going from $(a \succ b)$ to $\neg(a \succ b)$. The DP distance is defined on the graph of judgement sets, where 2 sets share an edge if there is no judgement set between them. Since KS and DP share their notion of betweenness, we introduce their betweenness as follows.

DEFINITION 5: *J-Betweenness*

A judgement set J_i is between preferences J_j and J_k if for every proposition about $x, y \in A$, J_i either agrees with J_j or J_k.

From this definition it is clear that this could only result in a voter updating their original opinion in which they have $(a \succ b)$ to a new opinion where $\neg(a \succ b)$ only if the announced opinion contains $\neg(a \succ b)$.

Figure 3.1 shows a graph used for the DP distance in the case of 3 alternatives, for simplicity the associated preferences are used to label the judgement sets.

The CS distance is simpler and is simply defined as the number of positions two voters disagree on, and a preference is between two others if for each position it agrees with one of the two preferences.

Each distance has different trade-offs, CS is the simplest, but might exaggerate the distance when there are many alternatives, for example if 2 voters agree on the relative ranking of all but 1 alternative, which one voter happens to rank first, thereby shifting the opinion of voter 2 right by one, the CS distance would conclude that these voters are in full disagreement, while reasonably one could conclude their opinions do not differ much. The KS distance, using judgement sets instead of raw profiles captures this more effectively, while still being relatively easy to compute, but in cases of more disagreement, it is likely to over count the distance, since the binary changes do not capture logical necessities. For example, swapping $(a \succ b)$ to $\neg(a \succ b)$ must result in $(b \succ a)$ becoming true (in the case of strict preferences), thus one might reasonably conclude this should only count as 1 step. DP improves upon this, but in doing so becomes much harder to compute, mainly through the cost of constructing the full graph of judgement sets, which grows in $f_n = 1 + \sum_{j=1}^{n-1} \binom{n}{j} f_{n-j}$ in the number of vertices, where n is the number of alternatives [**?**]. This can easily be verified by noting that the number of judgements sets over n corresponds to the number of weak preference rankings over n alternatives, which is define as alternatives, and a binary choice on each proposition.

Apart from these distances, they define a voter as a tuple of a (weak) preference and a bias (towards their current position) $v = \langle r, b \rangle$, with $b \in \mathbb{R}_{[0,1]}$. Finally, a deliberation step $D_s : V \times r \to V$, with V being a set of voters and s being one of the spaces (KS,

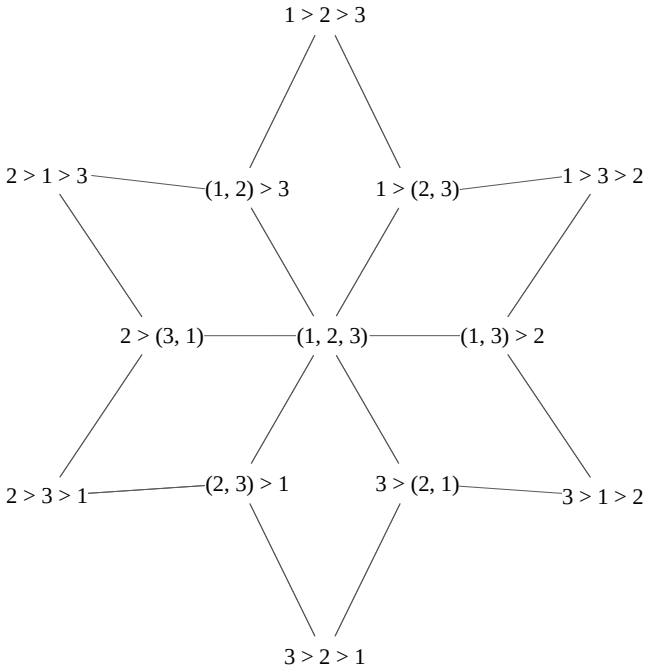

FIGURE 3.1: The graph of judgement sets for all preferences over three alternatives, brackets indicate ties.

DP, CS). A round of deliberation consists of n deliberation steps, such that each voter announces their opinion once. We formulate this procedure in the following program:

---

input : Set of Voters V, metric space s
output: Updated set of Voters V

$V_u \leftarrow V$ `// Set of unannounced voters (references to` V`)`
while $|V_u| > 0$ **do**
  Select a random $v \in V_u$
  $V_u \leftarrow V_u \setminus \{v\}$
  $V \leftarrow D_s(V, v.r)$ `// Update voters based on` v`'s preference`

The deliberation step D_s then updates all voters such that their new preference minimize the following formula.

$$r = \sqrt{b d_s(r_i, r')^2 + (1-b) d_s(r_j, r')^2} \tag{3.1}$$

Where r_i, r_j are the voters and the announced preference, respectively, and r' is the voters new preference.

We present a replication and extension of their work Chapter 6. Furthermore, we present novel (negative) results based on this model in Chapter 4.

Though this model is simple and captures some communication of preferences, if we attempt to use it to model meta-agreement, it seems to be lacking in at least two important ways. Firstly, agents do not conceive of anything relating to the structure of the problem. They simply announce their preferences, and all other listen and update accordingly, thereby moving to some sort of substantive agreement. Secondly, the model presupposes that all opinions are equally defensible, and that each voter is equally able to formulate this defense. To address this we formulate a new model in Chapter 4.

3.3.1 Deliberative experiments

We now present some empirical studies showcasing the effects of deliberation in voting populations, focusing on some meta analysis of deliberative interventions, and the AMERICA IN ONE ROOM experiment.

3.3.1.1 Meta-Analysis

3.3.1.2 America in One Room

Fishkin et al. [18] conducted a large scale experiment, during which they brought together American Voting-eligible citizens to deliberate about policies leading up to the 2020 presidential elections. They conducted a questionnaire on these people measuring the knowledge of the current state of politics, the opinions on 4 issue domains (Climate, Migration, Economy, Health Care, Foreign Policy), and their political affiliation (E.g. Who they would likely vote for, whether they considered themselves more liberal or conservative). This questionnaire was also conducted to a control group of people who did not participate in the deliberation. They found deliberation to increase the likelihood of voting, improve the opinion on their political rivals, increase the likelihood of voting for president Biden, among other effects. They explain these effects through, what they call, "Civil awakening". This states that previously uninformed and uninvolved voters become involved through an increase in self-efficacy as well as their knowledge. These were still measurable one year after the intervention. Though they did not measure full preference rankings over the possible parties, these results do indicate both an increase in Meta-Agreement, and Substantive-agreement. Namely, in terms of their opinions, opinions tended to shift more moderate, which more conservative voters changing their opinions most. The authors also note that moderate voters become more likely to voter for Biden, indicating some change in how voters conceptualize of the Candidates' positions.

CHAPTER 4

THEORY

In the model of deliberation by Rad and Roy [14], outlined in Section 3.3, they aim to model deliberation and show that deliberation results in nicely structured profiles which allow for strategy proof voting rules. One important caveat, given by the authors as well, is all participants should honestly and truthfully participate in deliberation. We now provide a formal statement, showing deliberation does not prevent strategic behavior.

Proposition 4.1. The process of deliberation over $|A| \geq 3$ through deterministic deliberation procedure $D : \mathcal{L}(A)^n \to \mathcal{L}(A)^n$, followed by voting with voting rule f cannot be surjective, strategyproof and non-dictatorial.

Proof. Assume, towards a contradiction, such a pair of deliberative procedure (D) and voting rule (f) exists. Any deterministic deliberation procedure D could, in principle, be embedded into a voting rule $f'(\boldsymbol{R}) = f(D(\boldsymbol{R}))$, such that the voting rule simulates D before applying f, which would result in voting rule f' being surjective, strategyproof and non-dictatorial. This is a contradiction, by the Gibbard-Satterthwaite theorem [1, 2]. □

We extend upon this result, showing the inclusion of biases in voters does not mitigate the negative result. For this we define DB as follows:

DEFINITION 6: *Biased Deliberation*

A deliberative procedure with biases DB : $\mathcal{L}(A)^n \times \mathbb{R}^n_{[0,1]} \to \mathcal{L}(A)^n$ is an extension on a standard deliberative procedure. DB has access to the bias each voter has towards their own opinion.

We now proceed with a corollary on Proposition 4.1. Towards this we assume biases are true, in the sense that a voter cannot help but be 'convinced' by the presented profiles as much as their bias allows for this. We think this assumption is weak and natural in the light of the current model. Furthermore, a violation of this assumption would not imply the following corollary to be false, instead the bias itself becomes a point of strategy, allowing voters to pretend to be more hardheaded than they in fact are.

Corollary 4.2. A deliberative procedure with biases, followed by voting with any voting rule f, cannot be surjective, strategyproof and non-dictatorial

The proof of this follows from a reduction of the biased Deliberation DB to general deliberation D.

Proof. Take any election consisting of biased deliberation DB and voting rule f, since biases $\boldsymbol{b}$ are true by assumption, they must be fixed, meaning that $\boldsymbol{b}$ is not reported but some fact of the matter. If this election was immune to strategic manipulation, then a deliberative procedure D could embed this b, and simulate biased deliberation DB, resulting in $D'(\boldsymbol{R}) = \text{DB}(\boldsymbol{R}, \boldsymbol{b})$. As a direct corollary to Proposition 4.1, such a D' cannot be surjective, strategyproof and non-dictatorial, showing a contradiction. □

This result is independent of the metric space chosen. From here we now show that even if we take the deliberation procedures on its own, it still not immune to strategic manipulation. For this we restate strategyproofness as follows:

DEFINITION 7: *Strategyproofness of Deliberation*

A deliberation procedure is strategyproof if there exists no voter i such that there is a profile $\boldsymbol{R}$, in which i misreporting their preference R_i as R'_i results in the profile after deliberation $D(\boldsymbol{R})$ is further from the i's original preference than if they had reported R'_i. This distance is measured as

$$\text{Dist}(R_i, D(\boldsymbol{R})) \geq \text{Dist}(R_i, D(\boldsymbol{R}')).$$

Where the Dist function is simply the sum of all distances between R_i and all preferences in $\boldsymbol{R}$.

One important note is that in the final profile, the preferences of voter i might not be the same as it was before the deliberation. That is why the distance is calculated w.r.t. i's original preference. Intuitively this could be read as i misreporting their preference to prevent even their own mind from being changed. Using this definition, we show that the deliberative procedures, under the metric spaces *KS*, *DP*, *CS* are not strategyproof. Stated as follows:

Proposition 4.3. Deliberation under distance measures *KS*, *DP*, *CS* is not strategyproof, for $n \geq 2$ and $m \geq 3$.

We provide a proof by construction, we show how to do this for the KS and DP distance measures, as they share the same profiles for this proof. The proof for the CS distance measure is laid out in Appendix A.

Proof. Assume the following population: we have voter 1 whose bias is 1, and all other voters $j \neq 1$ have bias 0.5. Furthermore, we have $\text{Dist}(R_1, R_j) = 2$ for all j. Voter 1 now has the option to report R'_1 instead, which has $\text{Dist}(R'_1, R_j) = 4$ and $\text{Dist}(R'_1, R_1) = 2$. If voter 1 reports R'_1, then all j will update towards 1's true preference, as using equation (3.1) we get $r(R_j, R'_1, R_1) = 4$, while $r(R_j, R'_1, R_j) = r(R_j, R'_1, R'_1) = 16$.

Resulting in $\text{Dist}(R_1, D(R_1, \boldsymbol{R}_{-1})) = 2(n-1) > \text{Dist}(R_1, D(R'_1, \boldsymbol{R}_{-1})) = 0$.

Since 1 has a bias of 1, the order of the deliberation has no effect.

We now show that for distance measures KS and DP, there exists these 3 preference orderings such that the necessary profile can be constructed. We use the following profiles:

$$R'_1 = a \succ c \succ b \succ \cdots \succ m,$$
$$R_1 = a \succ b \succ c \succ \cdots \succ m,$$
$$R_j = b \succ a \succ c \succ \cdots \succ m.$$

As we are only allowing strict preferences, both distance metrics behave the same locally, with the distance of two profiles being 2 whenever one is 1 swap of alternatives away from the other. This means that R_i and R_j have a distance of 2, as well as R'_1 and R_1 having a distance of 2. In this case the total distance from R'_1 to R_j is simply the sum of the local distances for both distance metrics, thus satisfying our requirements. □

These results show it is likely frivolous to attempt to design a strategy proof deliberation procedure of the likes shown. Instead, focus is now brought to modeling 'ideal' deliberation, as laid out in Section 3.2.2. We provide the following mathematical formulations to the four tenants laid out. *Freedom*: voters can report any preference, *Reason*: voters

are rational, *Equality*: no voter has special rights, *Consensus*: voters deliberate aim to reach consensus. Which we extend with *Honesty*: Voters represent their true beliefs and preferences only.

4.1 Our model

In an attempt to model meta-agreement through deliberation, our model needs to make a proper distinction between the 'substantive level' and the 'meta level'. In order to do so, we propose the following, let $\Psi = \{\psi_1, \cdots \psi_k\}$ denote the set of policies that could be implemented. A voter $i \in N$, has support for these policies, represented as a number on an interval over $\mathbb{R}$. At a meta level, a voter has an understanding of which policies are supported by which alternatives. This is modelled as matrix, representing the estimated support for each policy for a candidate, thus voter i has Σ^i, where $\Sigma^i_{j,x}$ represents this voters' estimated support of ψ_j by alternative x.

This model does not explicitly model $D1$, the discovery of a common issue dimension, on the one hand, if the alternatives can be reduced to a line, this model should be able to capture this, even if this one line crosses through multiple issue dimension. For example if all issues are strongly (negatively) correlated on the side of the alternatives, but not on the voters, this model allows for the voters to recognize this by properly estimating the alternatives' support matrices, while voters themselves can keep an uncorrelated support vector. In the case that the actual issue dimension is simply not included in Ψ, our model would not be able to discover this new dimension, even if human deliberation feasibly could. More straightforwardly, if we the measured support is irrelevant to the true issue dimension(s), our model cannot recover the true issue dimension.

Our model adapts the DeGroot learning model, which originally models probability distributions. In that model, a voter is a node in a graph, and deliberation can be modeled as a Markov chain. In our model, we keep voters as nodes on a graph, as well as a Markov chain, however, instead of a probability distribution, a voter has a support vector $S_i \in \mathbb{R}^{|\Psi|}_{[0,1]}$, and estimated support matrix $\Sigma_i \in \mathbb{R}^{|A|\times|\Psi|}_{[0,1]}$.

Note that this does not mean that all policies have to have any (estimated) support, nor that an alternative can only support a specific number of policies, in principle there can be alternatives that represent the status quo, and thus do not support any policies, and there can be alternatives that are estimated to support all policies. Let $\boldsymbol{S} = [S_1, \ldots, S_n]^T$ denote the population opinion, which has shape $|N| \times |\Psi|$.

In order to extract a ballot from this matrix, we assume a voter ranks the alternatives such that the most preferred alternative has the smallest distance between the estimated support matrix for that alternative and her own. We further allow this distance to be

weighted, such that a voter may have one or more policies their think are more important.

Next we define the deliberative procedure in terms of the trust matrix of the DeGroot model.

Firstly, a deliberative step can be modelled using a transition matrix T, defined as follows:

$$T = \begin{bmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nn} \end{bmatrix}$$

Here each t_{ij} represents how much voter i trusts the opinion of voter j, in order for this to be a proper stochastic matrix, all rows must sum to one, and have non-negative entries. Although this last requirement could be seen as unrealistic, as a voter might actively distrust another voter and update away from their opinion.

Using this, we can now model the opinions of voters after a deliberative step as a matrix multiplication on some matrix M:

$$M^{(1)} = TM^{(0)} \tag{4.1}$$

Each entry in the matrix then is simply a linear combination of the other entries in that same column in $M^{(0)}$. In the case of $M = \Sigma$, this means that voter i's support vector becomes a linear combination of all support matrices, weighted by the trust in each voter. Deliberation can now be modelled by taking powers of the trust matrix, T^t, representing t deliberation steps. This matrix now represents how much each voter i has learned from the other voters, and can then be used to right multiply both the support and the estimated support matrix to calculate a voters beliefs after deliberation.

Finally, we provide an example of the first deliberation round in example 4.1, since it is identical for both S and Σ, we only show it for Σ. The example also shows how voters can initially agree on their support for policies, while disagreeing on their preferred candidates, using meta-agreement to come to a consensus.

EXAMPLE 4: *DeGroot deliberation*

We have voters $N = \{1, 2\}$, events $\Psi = \{\psi_1, \psi_2\}$, and candidates $A = \{a, b\}$. The voters both think that $\psi_1 = 1, \psi_2 = 0$, meaning that they fully support the first policy and reject the second, they estimate the support by alternatives as:

1	ψ_1	ψ_2
a	0.5	0
b	0.5	1

2	ψ_1	ψ_2
a	1	0.9
b	1	0.1

Interpreting this matrix for both players on ψ_1 shows, voter 2 thinks a and b fully support ψ_1, while voter 1 thinks that a and b support ψ_1 less. We can encode this into the estimated support matrices as follows:

$$\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 1 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0.9 \\ 1 & 0.1 \end{bmatrix}$$

This results in voter 1 preferring candidate b over candidate a, while voter 2, prefers a. Intuitively, since voter 1 thinks ψ_1 is equally supported by each alternative, while ψ_2 is not supported by a, it makes sense for them to prefer candidate a. Looking at the distances, we see that the absolute distance between voter 1 and alternative a is 0.5, while for alternative b it is 1.5. For voter 2 we see that the distance to a is 0.9, while for alternative b is it 0.1. Thus, voter 2 prefers b to a.
For the deliberation, we assume the following trust matrix:

$$T = \begin{bmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{bmatrix}$$

We get the following updated opinions:

$$\begin{aligned} \mathbf{\Sigma}^{(1)} &= T\mathbf{\Sigma}^{(0)} \\ &= T\left[\Sigma_1 \Sigma_2\right]^T \\ &= \left[(0.3\Sigma_1 + 0.7\Sigma_2) \quad (0.2\Sigma_1 + 0.8\Sigma_2)\right]^T \\ &= \left[\begin{bmatrix} 0.85 & 0.63 \\ 0.85 & 0.37 \end{bmatrix} \quad \begin{bmatrix} 0.9 & 0.72 \\ 0.9 & 0.18 \end{bmatrix}\right]^T \end{aligned}$$

These new estimates are not yet in full consensus, meaning Meta-Agreement has not yet been reached. Looking at their corresponding ballots, however, shows there is consensus on their most preferred alternative, as they both agree that alternatives support ψ_1 equally, while b supports ψ_2 less.

4.1.1 Consensus

Using this model of deliberation, meta-agreement can be seen as some shared estimated support matrix over all policies. If the goal of deliberation is meta-agreement, then the study of interest becomes the dynamics of convergence towards a unified estimate.

We present a summary of results relating to strongly connected graphs, as well as graphs for which there exists only closed and strongly connected subsets of nodes. For other results we refer to Golub and Jackson [19]. Firstly we focus on the strongly connected graphs.

Proposition 4.4. (Golub and Jackson [19]). For a strongly connected matrix T, the following properties are equivalent:

- o T is Convergent
- o T is Aperiodic
- o There exists a left eigenvector $\boldsymbol{s}$ for matrix T, with corresponding eigenvalue 1, whose entries sum to one, such that for every P_i, we have

$$\left(\lim_{t\to\infty} T^t \boldsymbol{P}\right)_i = \boldsymbol{s}\boldsymbol{P}$$

This result is positive for studying the convergence dynamics, as no knowledge of the initial distribution is needed to determine convergence, it allows us to simply verify one of these three properties on the network. Though strongly connected graphs might be a strong requirement, in the case of small scale (in person) deliberation, this might be realistic. Fortunately, even outside this setting it might be possible to reach convergence. For this we first define what a closed set of nodes is.

DEFINITION 8: *Closed set of Nodes*

A set of Nodes $C = \{1, \ldots, n\}$ is closed if for each $i, j \in C$ we have $T_{ij} \geq 0$ and for each $i \in C, j \notin C$ we have $T_{ij} = 0$

Using this definition, if each node is part of a closed set, we can form the following proposition

Proposition 4.5. (Golub and Jackson [19]). If for each $i \in N$, i is a member of a closed set in the graph, and each closed set is strongly connected, T is convergent.

4.1.2 Voter Mapping

One might want to expand this model to capture larger scale group dynamics, such as social networks. For this a reasonable approach could be to gather data regarding the opinion of the general population, and to map this onto a graph representing the communication in the population. For this we might want to find a bijection between the voters and the nodes such that the difference between the shortest paths in the graph and the opinion distance is minimized.

We show that mapping voters to a graph as just described is NP-Hard, and the decision variant of the problem to be NP-Complete. We call this problem Distance-based Voter Mapping, and define it as follows.

PROBLEM 1: *δ-DBVM(S)*

Given: $A, B \in S^{n\times n}, k \in \mathbb{R}_{\geq 0}$
Decision: Does there exist some bijection $f : [n] \to [n]$, such that:

$$\delta(A, f(B)) \leq k$$

Here we take $f(B)$ to mean the matrix B' that is created when we take each $B'_{i,j} = B_{f(i),f(j)}$ and δ is some distance function, $\delta : S^{n\times n} \times S^{n\times n} \to \mathbb{R}_{\geq 0}$.

We will be needing the Quadratic assignment problem (QAP), we formulate a decision variant of QAP as follows.

PROBLEM 2: *QAP-Decision*

Given: $A, B \in S^{n\times n}, k \in \mathbb{R}_{\geq 0}$
Decision: Does there exist some bijection $f : [n] \to [n]$, such that:

$$\sum_{i,j} A_{i,j} \cdot B_{f(i),f(j)} \geq k$$

Theorem 4.6. δ-DBVM(S) is NP-Complete for $\delta \in \{\ell_1, \ell_2\}$ and $S = \{0, 1\}^n$

Proof. ($\Longrightarrow$ NP-Hard) The proof follows from a reduction to the Quadratic Assignment Decision Problem.

Let A be the matrix of pairwise distances between voters, and let B be the matrix of shortest-path distances in the graph G, and k be the δ achieved by the optimal

bijection. ℓ_2-DBVM(S) requires finding a bijection f that minimizes the ℓ_2 objective:

$$\sqrt{\sum_{i,j}\left(A_{i,j} - B_{f(i),f(j)}\right)^2}.$$

Since the square root is a strictly increasing function, minimizing the expression above is equivalent to minimizing the sum inside:

$$\sum_{i,j}(A_{i,j} - B_{f(i),f(j)})^2.$$

Expanding the square gives:

$$\sum_{i,j} A_{i,j}^2 - 2A_{i,j}B_{f(i),f(j)} + B_{f(i),f(j)}^2.$$

The terms $\sum A_{i,j}^2$ and $\sum B_{f(i),f(j)}^2$ are independent of f (the former is fixed, the latter is a permutation of a fixed matrix), so the optimization reduces to:

$$\max_f \sum_{i,j} A_{i,j}B_{f(i),f(j)},$$

which is the standard form of the Quadratic Assignment Decision Problem. Note, $\max_f$ is a consequence of the sum being subtracted from the constants, thus we are still minimizing the total distance.

Now we note that when A and B are in $S = \{0,1\}^{n\times n}$, the ℓ_1 and ℓ_2 norms are identical. We also note that this binary domain would constitute a special instance of QAP, know as 0-1 Max-QAP, and is NP-Hard [20]. Thus solving δ-DBVM(S), on the binary domain, is equivalent to solving 0-1 Max-QAP, and thus NP-Hard. ✓

($\implies$ NP-Complete) Given some f and k, the decision can be made in $O(n^2)$. ✓

□

A concern with Theorem 4.6, might be the matrices containing certain patterns that might lead to an easier solution, though this proof concerns itself with the worst-case and thus this possibility of this problem being easier in practice is not issue. For this problem such patterns seem unlikely to be of much help. We show one example to give an intuition for this.

Take the case in which all voters hold one of 2 opinions, thus we can split them into two groups of sizes n, m. Then the mapping algorithm effectively requires finding a partition in the graph, that results in two sub-graphs with exactly n and m nodes each. This is the size-constrained graph partitioning problem, which is NP-Hard.

Thus, given that even under such a strong assumption the problem remains computationally difficult, we suspect that patterns in the data are unlikely to allow for easier exact solutions. This does leave room for approximation algorithms, we do not present an overview of these, however under our constraint of one of the matrices satisfying the triangle inequality, namely the voter distance matrix. There exists a $\frac{2e}{e-1}$-approximation algorithm [20].

Despite these negative results, we attempted to enlist the help of a QAP-solver [21] to find (approximate) solutions, using the Fast Approximate QAP Algorithm [22]. Though, we find the solver does not consistently find better solutions than random assignment, and is unable to handle large enough instances for the experiments presented in the following chapters.

CHAPTER 5

METHODS

We proceed with the methods used to replicate the paper by Rad and Roy [14], as well as the experimental setup of our own model. Links to the data used for these experiments can be found in Chapter 8. The programs are implemented using `OCaml`, and `Python`.

5.1 Replication

We implement the model as described in Section 3.3. Agents are limited to strict preferences over all candidates. All experiments are done with 3 alternatives, and 51 voters. The number of voters is chosen to be an odd number, as this prevents ties between alternatives. We measure evaluations relating to strict preferences, namely the proportion of cyclic Profiles, the Number of Unique profiles and the proximity to single-peakedness by voter deletion (PtS-V), as reported by Rad and Roy [14]. In addition to those we also measure the number of Condorcet winners. We do not measure the PtS-C, as any profile with 2 candidates is single-peaked, thus given the simulation will have 3 candidates this metric would be of little added value, as all values would be either 1 or $\frac{2}{3}$.

5.2 Experiments

We aim to replicate the findings by the AMERICA IN ONE ROOM experiments [18] in-silico. To this end we use the adaption to the DeGroot model as laid out in Section 4.1 The dataset contains a control group as well as an experimental group. In the dataset, the control group shows no change in opinion over time, thus this group is best modelled by using the identity matrix I^n as the trust matrix. The experimental group is modelled as a densely connected network. The distribution of the trust we control through 3 methods.

5.2.1 Modelling Trust

We propose three different mechanisms through which we will the trust matrix, as well as the intuitive and theoretical appeal.

Knowledge. Firstly we consider knowledge, this can be used to inform both the trust in others, and your bias towards yourself. For this we can imagine a vector $\boldsymbol{k}$, where each k_i contains some knowledge score for voter i. In modeling, we now have 2 options, firstly, does a voters knowledge affect their bias towards their own opinion. Intuitively one could reasonably argue either way. Two plausible ways of reasoning are, "A knowledgeable voter knows more facts and is therefore harder to convince", or "A Knowledgeable voter realizes the complexity of the topic and is therefore less certain". The first line of reasoning seems more general, as it seems independent of the topic at hand. However, the second line of reasoning seems to capture something like the Dunning-Kruger effect, which states that people can have "meta-ignorance", meaning they do not realize what they do not realize.

As for the trust a voter places in their peers, a similar argument can be made, where the voter could either place more trust on people that are more knowledgeable, and thereby might be able to provide more facts. Or less knowledgeable voters might be more persuasive in making strong an bold claims, as without strong knowledge on the subject voters are more likely to hold strong opinions SOURCE

Similarity. A voter might trust people more if they are similar to them, in this work we take similarity to mean a similarity in substantive opinion. It is however not hard to conceive of similarity influence trust in other ways such as social status.

Ego. Finally, a voter might be less inclined to change her opinion if more people value it.

Given these different options, the right selection of methods becomes question for empirical observation, which we present in the next Chapter.

Firstly, and most simply, we give all voters a bias. This bias reflects how much of their trust they place on themselves. For example a bias of 1 represents them placing equal trust on themselves as all other voters combined. The actual weight on the self loop is calculated as the sum of all incoming edges multiplied by the bias. Secondly, we have knowledge-based trust, in which a voter trusts voter j more if voter j is more knowledgeable. We get the knowledge scores from the AMERICA IN ONE ROOM dataset by taking the proportion of knowledge questions they answered correctly. The interpretation is that more knowledgeable voters would be more persuasive and thus be more influential on other voters' opinions. Thirdly, we have credibility-based trust, where the trust

a voter places on another voter is proportional to the number of outgoing edges that second voter has. This method becomes equivalent to placing uniform trust in all voters when all voters are situated in a fully connected graph. If we do not use credibility- or knowledge- based trust, we call this uniform trust, meaning that they treat all neighbors the same. Importantly, this does not imply any specific bias value.

Given these matrices we can define the full model in terms of matrix and Hadamard products, where Hadamard products are entry wise multiplications of matrices. First we define T_{out} as follows,

$$T_{\text{out}} = A \odot K \odot S \tag{5.1}$$

Where A is the adjacency matrix, without any self loops, K is the matrix of knowledge scores, and S is the matrix of similarity between each voter. We use the indicator functions $\mathbb{1}$ for both knowledge and similarity.

In order to determine the bias of each voter we use the matrix $T_{\text{out trust}}$, to generate T_{in} as follows,

$$T_{in} = (T_{\text{out}} b) \odot \text{diag}(K) \odot E \tag{5.2}$$

Where b is a vector of length n containing the bias factor in each entry, K is the knowledge matrix, from which we extract the diagonals, and E is the ego matrix defined as $T_{\text{out}}^{T}[1]^{n}$, thereby getting the sum of all incoming edges.

Through some abuse of notation we can now define the final trust matrix T as the diagonal matrix obtained from T_{in} and its element wise addition with T_{out}:

$$T = \text{norm}\left(\text{diag}(T_{\text{in}}) \oplus T_{\text{out}}\right) \tag{5.3}$$

Were norm normalizes the matrix such that each row sums to exactly 1.

Given this formulation, we define an instance of our model through shaping the matrices, such as shown in example

EXAMPLE 5: *DeGroot deliberation Instance*

Say we have the following matrices:

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad K = \begin{bmatrix} 0.5 & 1 & 2 \\ 0.5 & 1 & 2 \\ 0.5 & 1 & 2 \end{bmatrix} \quad S = \begin{bmatrix} 0 & 0.5 & 1 \\ 0.5 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \tag{5.4}$$

Note that A has no self loops, K repeats its rows, since each voter has 1 knowledge score, and S must be symmetric, as the similarity of voter i to voter j must be same as the other way round.
Now we want to create a trust matrix T, that uses knowledge for the outgoing trust, but not the similarity. For the bias it used a bias factor of 2, and uses Ego, it does not use self knowledge. To achieve this we define the following matrices,

$$K' = \begin{bmatrix} 0 & 1 & 2 \\ 0.5 & 0 & 2 \\ 0.5 & 1 & 0 \end{bmatrix} \quad S' = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{5.5}$$

If we now use K' and S', we see that T_{out} is not affected by the change from K to K' since the diagonals remain 0, and S' now has no influence on the trust. As a result we get:

$$T_{\text{out}} = \begin{bmatrix} 0 & 1 & 2 \\ 0.5 & 0 & 2 \\ 0.5 & 1 & 0 \end{bmatrix} = K'$$

As a result, T_{in} now becomes:

$$T_{\text{in}} = \begin{bmatrix} 6 \\ 5 \\ 3 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 12 \end{bmatrix}$$

The final trust matrix T is then:

$$T = \begin{bmatrix} 6 & 1 & 2 \\ 0.5 & 10 & 2 \\ 0.5 & 1 & 12 \end{bmatrix}$$

Which we then normalize to be:

$$T = \begin{bmatrix} \frac{2}{3} & \frac{1}{9} & \frac{2}{6} \\ \frac{1}{25} & \frac{20}{25} & \frac{4}{25} \\ \frac{1}{27} & \frac{2}{27} & \frac{24}{27} \end{bmatrix}$$

5.2.2 DeGroot extension

The first experiments we perform concern the DeGroot model. These experiments consist of two parts. Firstly we search the parameter space to identify parameters that best replicate the data, using Bayesian Parameter Estimation. For this we use data from the AMERICA IN ONE ROOM experiment as described in Section 3.3. Though this data does not provide full preference rankings over the candidates, it does provide data on voters' opinions on 6 different topics of political discussion, such as climate change and immigration. Using these opinions, we are able to generate potential candidates, this is done either by selecting a voter and creating a candidate with identical opinions, or by pooling 10 voters[1] and creating an average of their opinions. Using these simulated candidates we are able to create preference rankings using the ℓ_1-norm. To model the difference between the deliberation and control group we change the topology of the graphs voters in the respective groups are situated in. As mentioned before, the deliberation groups will be embedded in a fully connected graph, while the control groups will be placed on the graph of academic citations in physics [23], this graph is small enough to allow sampling of the graph for each simulation. Since the original data provides group numbers for candidates who participated in the deliberation, we also experiment with replicating these groups as opposed to randomly grouping voters together.

We measure whether the final profiles are cyclic, whether they have a Condorcet winner, home many unique profiles there are, and their proximity to being single-peaked. Proximity to single-peakedness is measured in two ways. When the simulation size allows for it, we measure the proximity in terms of the number of voters that would need to be removed for the full profile to become single-peaked. This particular notion is NP-complete [24], though it allows for a 2-approximation. We use the method based on an ILP solver, as implemented in `PrefTools` [25]. The other notion of proximity is the proximity in terms of the number of candidates that need to be removed for the profile to become single-peaked. This can be done in $O(|V| \cdot |C|^3)$[26], though the implementation we use is that of the `PrefTools` library [25], which implements a slower $O(|V| \cdot |C|^5)$ algorithm [24].

Given the best configurations, we will analyze the behavior of the model to understand the convergence on opinion. To this end, we measure the change in the trust matrix, as well as the distance between each voter's pre- and post-deliberation preferences using the KS and CS distances. We first aim to find the number of deliberative steps needed for convergence, which we define as the moment where the largest change in the trust

[1] This is arbitrary, and it might be good to look into this, but in my opinion this is low priority for now. It might also be useful to keep the candidates constant over the course of an experiment

Parameter	Description
`Number of Voters`	The number of voters in the simulation, representing either the deliberation group, or the control population.
`Number of Candidates`	The number of candidates to be voted on.
`Candidate Generator`	The way the candidates are generated. Either a random voter is selected for each candidate, or 10 random voters (sampled with replacement) get averaged into one candidate.
`Bias`	The bias all voters have towards their own opinion.
`Time steps`	The number of deliberation "steps" the voters undergo.
`Group`	Use the original groups.
`Similarity`	Distribute trust based on credibility.
`Knowledge`	Distribute trust based on knowledge.
`Ego`	Scale voters' bias according to the trust other people have in them
`Self-Knowledge`	Scale voters' bias according to their knowledge

TABLE 5.1: The parameters of the DeGroot learning based model, as well as their descriptions

matrix is smaller than some ϵ. Then we hope to understand how individual voters' opinions change by looking at the final state of the trust matrix.

Finally, we use sensitivity analysis to investigate which parameters have the strongest effect on the model, using Sobol indices to get the first and second order effects.

CHAPTER 6

RESULTS

We first present a full replication and extension of the work by Rad and Roy [14]. Then we present the simulations based on our model of meta-deliberation, as well as the results of the sensitivity analysis on both models. All code for the replication, main experiment and visualizations can be found in this Repository. Appendix B contains all the values and ranges used for the experiments, as well as supplementary figures.

6.1 Replication

We are able to fully replicate the results found by Rad and Roy [14], in Figure 6.1 we see that for the biases less than 0.73, all metric results in acyclic preferences. We also replicate the behavior of the KS metric, where biases in the range of 0.73-0.85, show that even initially acyclic profiles can become cyclic. Figure 6.2 Further explains this by showing that within this range we always observe 3 unique profile for the KS metric, while DP and CS have already settled on 6 profiles, thereby representing all possible preferences. Figure 6.3 shows KS introduces ambiguity in the case that there was a Condorcet winner, resulting in losing the original nice profile. Finally, the proximity to single-peakedness shows a slightly more positive note for the KS metric, showing that while the DP and CS bottom out to the minimum proximity to single-peakedness, KS stays relatively close. Though this should be taken with a grain of salt, as it is likely a consequence of the unique preferences being smaller.

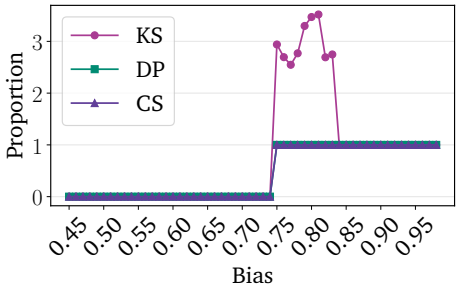

FIGURE 6.1: The proportion of cyclic profiles remaining, 0 indicating that no cyclic profiles were present after deliberation.

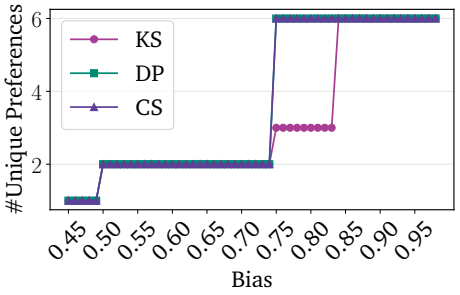

FIGURE 6.2: Number of unique preferences at the final step of deliberation.

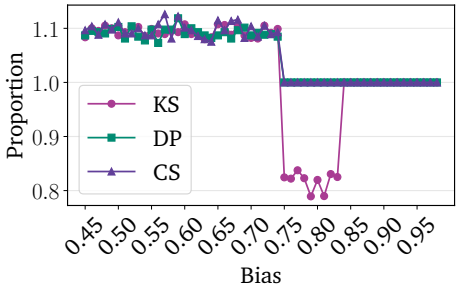

FIGURE 6.3: The proportion of Condorcet winners left after deliberation, value above one indicate Condorcet winners emerging during deliberation

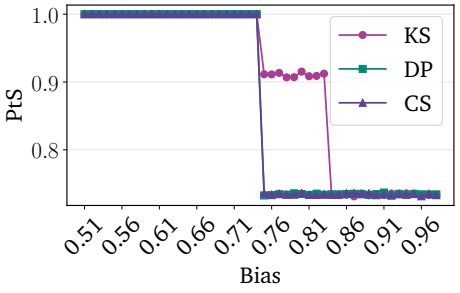

FIGURE 6.4: Proximity to single-peakedness after deliberation. Proximity to single-peakedness as defined in Section 3.3.

6.2 DeGroot Model

We present the results based on the DeGroot model. The model is informed by the data from the AMERICA IN ONE ROOM experiment, which was used to construct the support vectors S as well as the estimated support matrices Σ. We follow the original paper, focussing on the most polarizing questions, as mentioned in Section 3.3.1.2, the policy-based ideology score (PBS) is the average of the 26 most polarizing questions, where a low PBS corresponds to more liberal answers, and high PBS indicates more conservative answers.

We remove all participants with missing responses to any pre- or post-deliberation measurements, retaining only voters with complete pre- and post-deliberation data. As a result, only 247 out of the original 523 opinions remain after this selection. Though this removes a large fraction of voters, given that this model makes quite strong assumptions on voting behavior for which we do not have data, we limit our testing to voters

of which we can be sure that we know their true opinion. The support vectors S correspond to the voters' reported opinions, based on measured by several policy questions rated from 0 to 10 (inclusive). Each voter's estimated support matrix Σ is generated by adding normally distributed noise($\mu = 0$, $\sigma = 1.37$) to the candidates' true opinions. The mean of 0 ensures we do not bias the model towards preferring candidates with higher or lower average scores, as otherwise people would consistently be over or underestimating candidates' support. The standard deviation is chosen to match voter PBS distribution before deliberation.

To generate a deliberation groups, we opt for two approaches. Either using the original deliberation groups, selecting a group at random and using the voters from that group. Given the restriction of voters with complete data these groups will tend to be smaller than in the original study, where these groups averaged 13 voters, in our subsection the average is 7. Or we generate new groups by picking n voters uniformly at random without replacement and placing them into a single group. Disregarding any similarity to the original structure the groups might have had.

To evaluate model performance, we predict each voter's post-deliberation opinion and compare it to the observed data. Additionally, we group voters into m bins based on their initial PBS and compare the average predicted opinion within each group to the actual group average. This effectively model substantive agreement and thus does not incorporate *meta-agreement*. However, it allows for the evaluation of the model without assumptions on how to infer the final "preferences" of the voters, or the opinions of candidates. After this assessment, we investigate the convergence of the model, as well as its sensitivity to the choice of parameters.

Finally, we extend the model to incorporate meta-agreement through deliberation on the trust matrices. Assessing its effect on voters' final preferences, using the metrics introduced in Section 6.1.

6.2.1 Policy-Based Ideology Scores

We first proceed with analyzing the performance of the DeGroot model with respect to substantive agreement. Figure 6.5 shows the PBS of both the deliberation and control group, and the simulation results for both instances. As expected the model performs poorly at predicting the control group, as there was no significant change for control group members in the original data. Within the deliberation group, a voter's initial PBS remains a strong indicator of their final PBS. We observe that the models predictions get more accurate after the first time step, with prediction errors increasing over time. This is because the model causes voters to converge too strongly, thereby eliminating

most extreme opinions, contrary to the real data. The implications of this depend on the nature of long term deliberation. If, as suggested by **?**], deliberation is able to reach full consensus, the model might offer a plausible approximation of this process. However, if full consensus is not typically reached—as is precisely the motivation for incorporating meta-agreement into the model—then the DeGroot model should be seen as overly simplistic in its assumption that individuals converge toward a weighted average of the opinions presented to them.

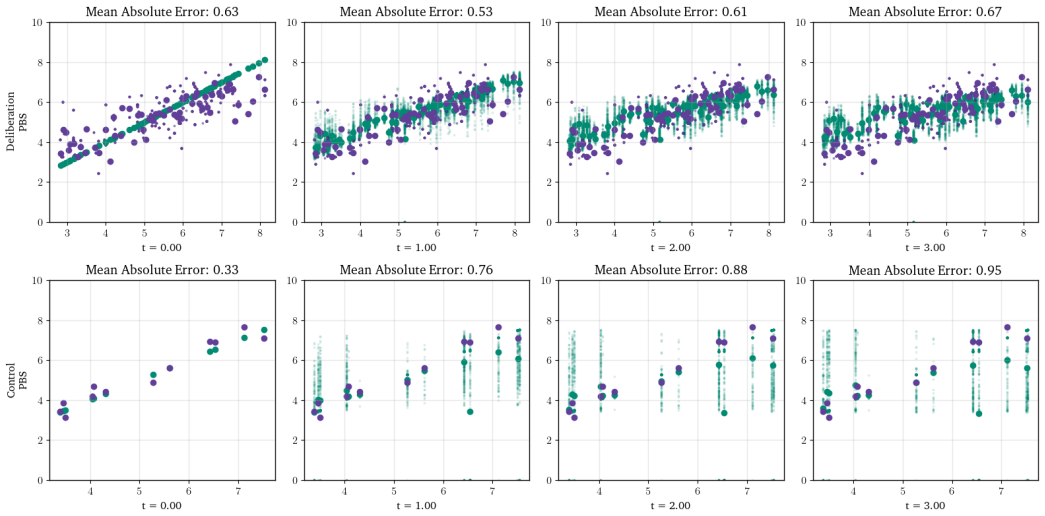

FIGURE 6.5: PBS, purple indicating the PBS after deliberation in the original data, green indicates the results of the simulation in that time step. Large dots indicate the binned data, smaller dots indicate individual voters.

Figure 6.6 depicts the change in PBS within the deliberation group. In the original data, we see that most changes occur among participants with high initial PBS, who tend to moderate their views. The model, by contrast, predicts the most significant changes among those with low PBS in later time steps.

One possible explanation for this discrepancy is the correlation between PBS and political knowledge. As shown by Fishkin et al. [18], voters with more extreme PBS also tend to be more knowledgeable. Our filtered dataset supports this, showing a weak negative correlation of -0.05 ($p < 0.5$), Figure B.1 in Appendix B shows the distribution of political knowledge across different PBS ranges. Since political knowledge in our sample is skewed toward voters with high PBS, incorporating knowledge-based trust into the model amplifies their influence, resulting in larger prediction errors.

However, Figure 6.7 shows that the model performs better when knowledge is excluded from the trust calculation. This suggests that political knowledge, at least as measured in this dataset, is a poor predictor of persuasiveness. It should be recalled, that the knowledge questions assess factual knowledge of the U.S. government, such as knowing

which party holds a Senate majority, which may not correlate well with persuasiveness on substantive issues such as immigration or economics.

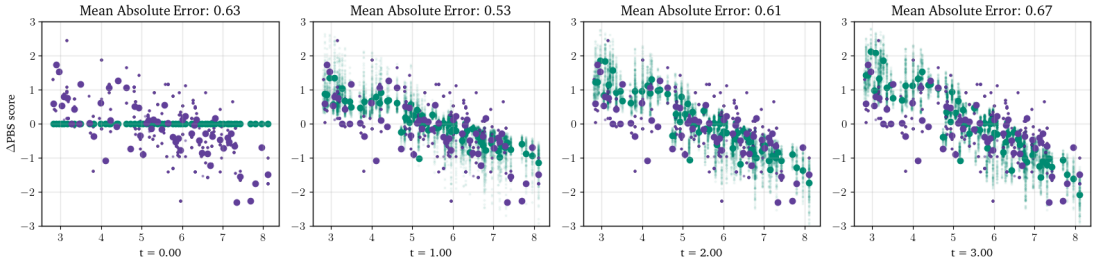

FIGURE 6.6: Change in PBS, relative to the original, pre deliberation, measurement. The control is omitted as there was no significant change.

We note that these slightly positive results appear only when the voters are grouped by their original PBS, thereby giving the model reasonable predictive power over a population of voters. This holds even for different number of bins. Figure 6.7 shows the progression of errors over time when the error is calculated on a per-individual basis, and we find the model consistently overestimates the change in PBS, and thereby gives a worse prediction than the initial PBS.

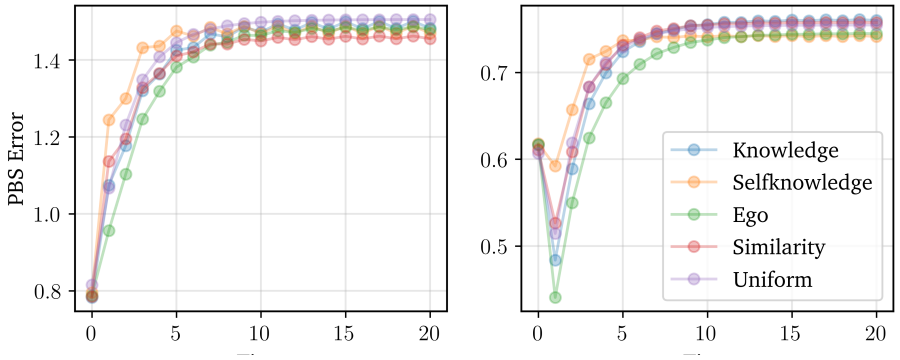

FIGURE 6.7: Prediction error of the model as a function of time, binned relative to the original PBS.

Figure 6.8 shows the relation between the bias factor and the PB score, showing that the bias does not improve the model's predictive power. As one might expect a bias is "slowing down" the model. Because of this the model is slower to diverge away from the true opinions.

We suspect ego improves predictive accuracy for two reasons. First, by assigning individual-specific biases, the model better reflects heterogeneous deliberative behavior. Second, increased self-bias slows down convergence, preventing the model from over-correcting.

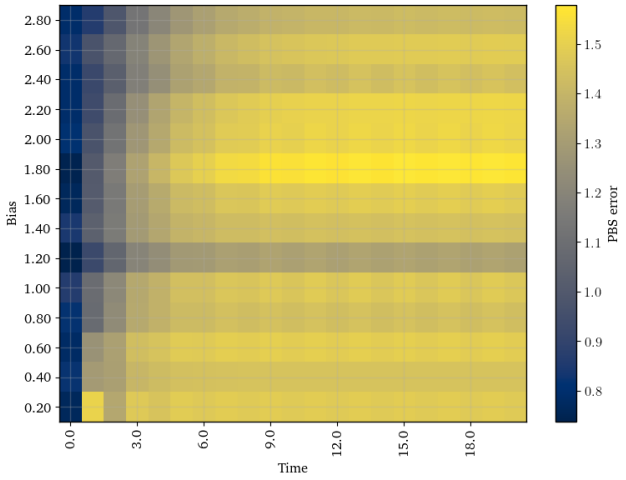

FIGURE 6.8: PBS Errors as a function of bias and time. Bias acts as a damper: when bias is higher the model take longer to over-estimate the change in opinion.

6.2.2 Convergence

From Chapter 4, we have seen that in the limit some matrices are convergent, while some are not, in particular if the matrix is aperiodic, it is convergent. As we model the deliberation group as having fully connected matrices, with self-loops, the matrices are aperiodic, and thus convergent. We look at the distance between the estimated support matrix, and the true support matrix, to get a sense of the rate of convergence. The distance is defined as the ℓ_1 norm.

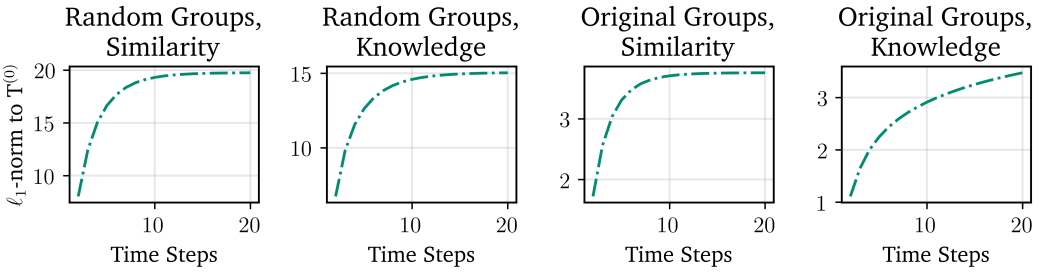

FIGURE 6.9: Convergence of trust matrices, as measured by the ℓ_1-norm between the trust matrix at the start and trust matrix at the current time step.

In Figure 6.9, we see that all configurations converge at a similar rate, slowing down the rate of change around t = 15. Since using the original groups leads to generally smaller groups, the absolute difference in the matrices is smaller. When using knowledge-based trust there is a lower rate of convergence

6.3 Sensitivity Analysis

We perform sensitivity analysis on the predicted PBS of the model. We do not use the original groups, as this allows us to vary the number of voters. Figure 6.10 shows the sensitivity indices. As for direct effects, as shown in the first order indices, the *number of voters* is clearly the biggest factor in the variance of the model. As expected the *bias* does not directly contribute to the variance in the model. *Knowledge* informed trust and *knowledge* informed bias (self knowledge) both are significantly impacting the variance of the model. The second order indices show *number of voters* interacts with *knowledge*, *self knowledge*, and *similarity*, contributing a large portion of their explained total variance induced by the *number of voters*. There is also an interaction between *ego* and *similarity* and *self knowledge*. As for the Total order indices, we see that variables contribute significantly to the variance in the model.

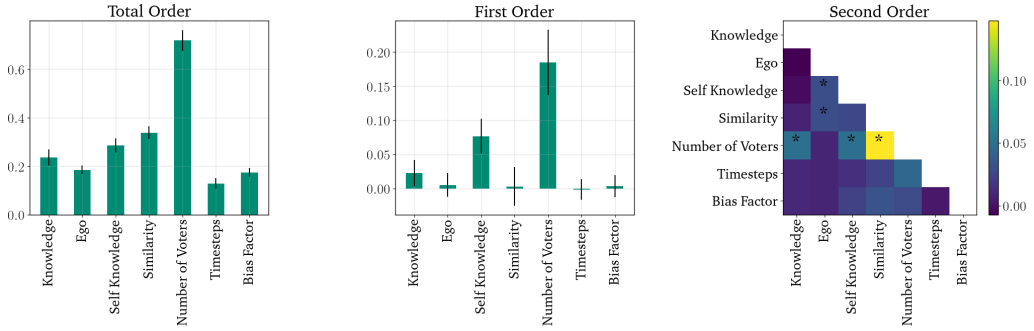

FIGURE 6.10: First, Second and Total sensitivity indices on the PBS prediction error. The stars in the heat map for the Second order sensitivity indices indicate significant interactions.

6.4 Adding Meta-Agreement

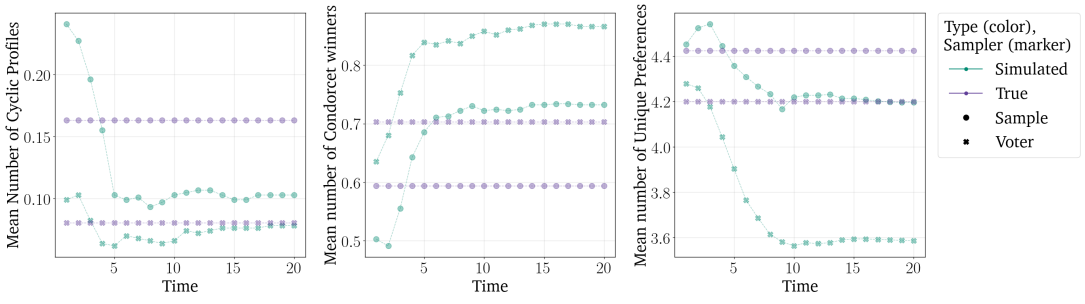

FIGURE 6.11: The proportion of cyclic profiles remaining, 0 indicating that no cyclic profiles were present after deliberation.

Firstly, when comparing different voter generation mechanisms, we find that generating a candidate by copying the opinion of a single voter performs best—both in minimizing the number of cyclic profiles and in maximizing the frequency with which a Condorcet winner exists. Though this result may seem unintuitive, we suspect the reason is that pre-deliberation opinions were relatively polarized. As a consequence, constructing candidates as averages of 10 voters tends to produce alternatives that are too similar, making it difficult for any one to stand out.

In contrast, a single voter's opinion is more likely to fall near a large cluster of voters, making that candidate closer—on average—to the majority. In such cases, that candidate is more likely to become a Condorcet winner. Put simply, averaged candidates tend to represent moderate positions, leading to greater voter indifference between them. In these situations, small errors in perceived support can have disproportionately large effects. Meanwhile, candidates based on a single voter's opinion—especially in a polarized society—are more likely to be distinct and strongly preferred.

Looking at the evaluation metrics used in the model, we observe a pattern similar to that found in the substantive agreement analysis. The simulation initially starts far from the true scores, gradually moves toward them, overshoots, and finally begins to converge.

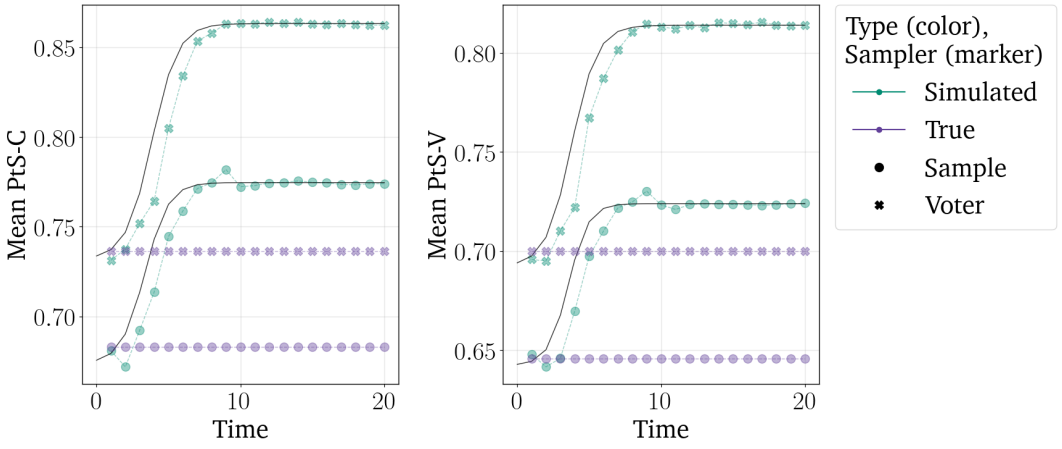

FIGURE 6.12: Proximity to single-peakedness after deliberation via candidate deletion (left) and voter deletion (right). The black line is a fitted sigmoid curve.

Figure 6.12 shows similar dynamics across simulation time for both notions of proximity to single-peakedness. Although candidate deletion and voter deletion represent two fundamentally different approaches to measuring this property, they yield a consistent conclusion: voters rapidly become more single-peaked early in the simulation, after which the rate of change slows and eventually plateaus. This behavior is well captured by a sigmoid curve, where the diminishing rate of change corresponds to the trust matrix stabilizing at its convergent state.

CHAPTER 7

DISCUSSION

7.1 Conclusion

We have shown deliberation to be susceptible to strategic manipulation under various notions of strategic manipulation. Most importantly, we have defined a new notion of strategyproofness specifically for deliberation over preference profiles. As a result, we caution that even when deliberation succeeds in producing single-peaked profiles, it does not guarantee fairness. In the sense that if all voters had been honest during deliberation, the resulting profile might have differed. Intuitively, this is expected, as individuals may pretend to hold more extreme views to shift the general consensus.

Given the idea that deliberation encompasses more than preferences alone, we extended our model to deliberation over opinions of voters and alternatives, through an adaptation of the DeGroot model. We also demonstrated that finding a voter graph that minimizes the difference between opinion distance and graph distance is NP-Hard. Suggesting that, in practice, modeling voters on sparse graphs requires collecting both communication patterns and opinion data.

Though we were able to replicate the results be Rad and Roy [14], we note that the model is too restrictive, as it only concerns itself with full preferences. We find that for certain ranges of bias the model behaves chaotically, and introduces cyclic profiles, even if before the deliberation these were acylic for the KS distance measure. Given that this model strictly models preferences, we extend it to a DeGroot learning model to incorporate opinions in general.

The DeGroot model successfully predicts opinion shifts at the population level but performs poorly at the level of individual voters. Extending the simulations to include deliberation on candidate positions, that is, on meta-opinions, yields profiles with more similar characteristics to the true preferences. However, after a few iterations, preferences converge too strongly, resulting in profiles that are less cyclic and more single-peaked than reality. We note that the development of proximity to single-peakedness (for both candidate and voter deletion) follows a sigmoid curve, characterized by a gradual start, a phase of rapid change, and eventual convergence.

Our sensitivity analysis shows that all model parameters influence outcomes. However, only parameters tied to knowledge and voter count exert strong direct effects — likely because they introduce new information, whereas other parameters only modulate existing dynamics. In terms of predicting final PBS, we found that Ego-based bias and knowledge-based trust performed best. Thereby indicating that more knowledgeable people are more convincing in deliberation, while people become less likely to change their minds if many people value their opinion. This is an intuitive result, but as mentioned Chapter 6, we caution that the reason for Ego-based bias performing well might be (partly) explain by simply reducing the change in opinions of voters.

In general political elections might elect candidates that will bring about many changes, most voters however will not have a strong opinion on all of these. Though we investigated the change in opinion on a per-topic basis, we were unable to incorporate this in the preferences over candidates. Given that most change in opinion was on immigration and healthcare, it might be the case that these topics were considered more important for this election and thus more time was spent discussing these. As a result the preferences of voters over candidates might be influence more strongly be these topics.

7.2 Limitations

7.2.1 Applicability of results

While the DeGroot model has been shown to be a more accurate representation of human belief updating than full Bayesian updating, it does not take into account why people hold certain beliefs, not does it constraint what kinds of beliefs a voter can hold at the same time. To remedy this, one might consider a framework such as abstract argumentation theory (source (DUNG 1995)), as this is able to model the arguments with the deliberative groups. Though, this seem theoretically nice, as it allows for formal description on why opinions and preferences are held, not just descriptions of these facts. From a simulation based perspective, such a model introduces major validity question. Firstly the framework requires a map on the relation of all arguments, for this one

does not only need qualitative data, i.e. reported arguments by participants, but also a method of reliably and accurately transforming these qualitative reports to argumentative graphs. Secondly, the abstract argumentation framework does not pose an updating mechanism, thus the method through which participants would updates their believes using this framework is unclear.

The choice of the DeGroot model assumes that people linearly interpolate between opinions presented to them. While it has been shown that in some circumstances this is a good heuristic, especially when compared to full Bayesian updating. It does limit the behavior of voters substantially.

Furthermore, our negative results on strategic manipulation might be remedied in human deliberation. As fellow participants might be able to sense that someone is being dishonest. It is also not unreasonable to think that someone pretending to hold a different opinion, might be worse at defending this opinion, and therefore be less persuasive.

In the testing of our model, we have consulted a single dataset containing a large amount of information need. However, as a result of not having a single data set which can fully inform the values in our model, multiple strong assumptions have been made in order to test the feasibility of the model. We mention these explicitly once more, as well as how what kind of data might be collected to inform this and similar models. We formulate this in terms of missing information on voters and candidates respectively, and present some of the most important limitations of each.

7.2.2 Voter information

Firstly, we have no access to a data set containing opinions before and after deliberation as well as the corresponding preference orders. This means having to infer preferences over candidates, though we chose to do this by minimizing the distance between the voter and the alternatives, reasonable alternatives exist. For example, people might use different heuristics to locate a few alternatives they like best, such as “Agrees with me on important topics”, thus putting weight on certain issue dimensions. Or they might simplify each comparison to "Agrees or Disagrees" with me, with some range of opinion they consider to be in agreement with theirs.

Furthermore, the way we encode voters’ information on alternatives might not accurately reflect true voters’ information. One might expect voters to be more familiar with candidates close to them in opinion, and thus have less noisy estimates of these candidates’ positions.

Finally, the error of voters' estimates might not be normally distributed. In a polarized election, it is not unreasonable to expect errors on the "opposing" party to skew further way from that voters' opinion.

7.2.3 Candidates

Though datasets such as those by Ipsos might contain the scores of political parties, these datasets cannot (easily) be combined with the data used in this study. This is mostly due to the inconsistent formats of the questions and the included topics. As a result, we are required to generate candidates manually. Thereby not only introducing another modeling choice, but also discarding an important piece of information. The AMERICA IN ONE ROOM dataset does contain voters' most preferred candidate, which is either the Democratic or Republican Party or Independent (participants are not asked to further specify). But lacking information on the candidates true positions as well as the ranking over the others, this information is hard to incorporate within the model.

Instead, we chose a simple approach, either selecting a single voter, or grouping some voters together as a single candidate. In the real world, however, candidates might arise in different ways and forms. For example, they could bring new idea's, not measured in the poll, or gather like-minded people instead of catering to the entire voting population, indeed the latter seems to be point of representative democracies. In representative democracies candidates represent specific demographics of the population, and advocate for their interests.

7.2.4 Extentions

Given the weak performance of the model, a better computational model is needed to understand deliberation and inform the design of deliberative interventions. We propose some extensions to the model, which might better capture human dynamics.

When humans deliberate, the amount of trust placed on each person is likely not fixed, for example, if someone has convinced a voter to change their mind on multiple topics, this voter might be more likely to trust them on a new topic as well. Or on the contrary, if two voters consistently disagree, and diverge in their opinions, they might come to trust each other less. Furthermore, the development of trust will likely differ between people, both on their baseline and development of their trust. Though we do not propose a specific updating scheme for this, the model is flexible enough to allow for the updating of the trust matrix over time.

If we extend this model to model large population, for example using a social network, it is crucial to be able to assess the effects disruptive events such as a national health crisis, an economic recession or a national safety threat.

As a result of such an event subset of the voters might become more informed on the position of the candidates, as the event might cause information dispersal, for example through news networks. This limitation is related to the notion of *Salience* as described by List et al. [12], stating that topics with high salience benefit less from deliberation, as participants have likely received more information on this topic. Though this could, at least in principle, be resolved by constructing trust matrices for individual topics. This would raise further questions as to the validity of these matrices.

CHAPTER 8

ETHICS AND DATA MANAGEMENT

A new requirement for the thesis is that there must be a short section in which you reflect on the ethical aspects of your project. This requirement is related to one of the final objectives that a graduated student of the Master of Computational Science must meet: "The graduate of the program has insight into the social significance of Computational Science and the responsibilities of experts in this field within science and in society". You don't need to devote an entire chapter to this; a short section or paragraph is sufficient.

I acknowledge that the thesis adheres to the ethical code (https://student.uva.nl/en/topics/ethics-in-research) and research data management policies (https://rdm.uva.nl/en) of UvA and IvI.

The following table lists the data used in this thesis (including source codes). I confirm that the list is complete and the listed data are sufficient to reproduce the results of the thesis. If a prohibitive non-disclosure agreement is in effect at the time of submission "NDA" is written under "Availability" and "License" for the concerned data items.

Short description	**Availability**	**License**
America In One Room	https://doi.org/10.7910/DVN/ERXBAB	CC0 1.0

APPENDIX A

EXTENDED PROOFS

Finally, for CS, R_1 and R_j stay the same, while $R_1^{'} = c \succ a \succ b \succ \cdots \succ m$, resulting in $\text{Dist}_{\text{CS}}(R_1^{'}, R_j) = |2-2| + |1-3| + |3-1| = 4$.

APPENDIX B

NOMINAL VALUES AND SUPPLEMENTARY FIGURES

B.1 Additional Figures

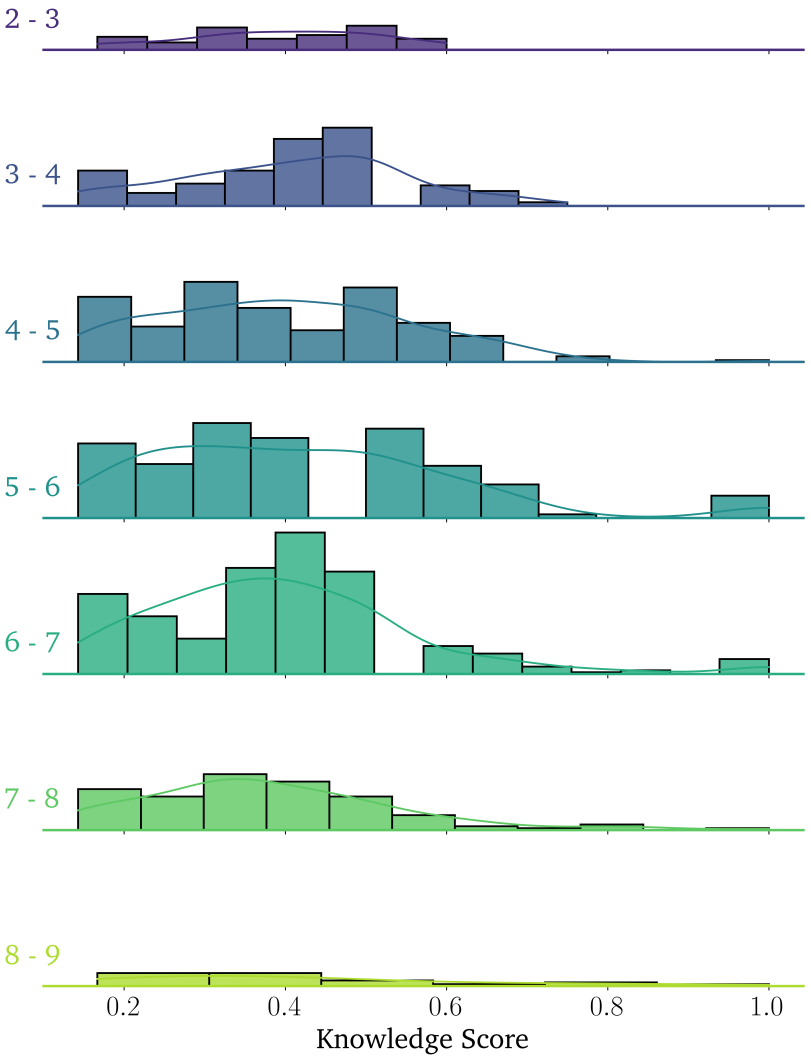

FIGURE B.1: The distribution of knowledge scores for different ranges of policy-based ideology scores.

BIBLIOGRAPHY

[1] Allan Gibbard. Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4):587–601, 1973. ISSN 0012-9682. doi: 10.2307/1914083.

[2] Mark Allen Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, April 1975. ISSN 0022-0531. doi: 10.1016/0022-0531(75)90050-2.

[3] Edith Elkind, Martin Lackner, and Dominik Peters. Preference Restrictions in Computational Social Choice: A Survey, May 2022.

[4] Wulf Gaertner. Domain restrictions. In *Handbook of Social Choice and Welfare*, volume 1 of *Handbook of Social Choice and Welfare*, pages 131–170. Elsevier, January 2002. doi: 10.1016/S1574-0110(02)80007-8.

[5] Donald E. Campbell and Jerry S. Kelly. Non-monotonicity does not imply the no-show paradox. *Social Choice and Welfare*, 19(3):513–515, 2002. ISSN 0176-1714.

[6] Donald E. Campbell and Jerry S. Kelly. Correction to "A Strategy-proofness Characterization of Majority Rule". *Economic Theory Bulletin*, 4(1):121–124, April 2016. ISSN 2196-1093. doi: 10.1007/s40505-015-0066-8.

[7] Donald E. Campbell and Jerry S. Kelly. Anonymous, neutral, and strategy-proof rules on the Condorcet domain. *Economics Letters*, 128:79–82, March 2015. ISSN 0165-1765. doi: 10.1016/j.econlet.2015.01.009.

[8] Samuel Freeman. Deliberative Democracy: A Sympathetic Comment. *Philosophy & Public Affairs*, 29(4):371–418, 2000. ISSN 1088-4963. doi: 10.1111/j.1088-4963.2000.00371.x.

[9] Joshua Cohen. Deliberation and Democratic Legimitimacy. In *Debates in Contemporary Political Philosophy*. Routledge, 2002. ISBN 978-0-203-98682-0.

[10] Jon Elster. The market and the forum: Three varieties of political theory. In *Debates in Contemporary Political Philosophy*. Routledge, 2002. ISBN 978-0-203-98682-0.

[11] Christian List. Two Concepts of Agreement. *The Good Society*, 11(1):72–79, 2002. ISSN 1538-9731.

[12] Christian List, Robert C. Luskin, James S. Fishkin, and Iain McLean. Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls. *The Journal of Politics*, 75(1):80–95, January 2013. ISSN 0022-3816, 1468-2508. doi: 10.1017/S0022381612000886.

[13] Valeria Ottonelli and Daniele Porello. On the elusive notion of meta-agreement. *Politics, Philosophy & Economics*, 12(1):68–92, February 2013. ISSN 1470-594X. doi: 10.1177/1470594X11433742.

[14] Soroush Rafiee Rad and Olivier Roy. Deliberation, Single-Peakedness, and Coherent Aggregation. *American Political Science Review*, 115(2):629–648, May 2021. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055420001045.

[15] John G Kemeny and James L Snell. Preference ranking: An axiomatic approach. *Mathematical models in the social sciences*, pages 9–23, 1962.

[16] Conal Duddy and Ashley Piggins. A measure of distance between judgment sets. *Social Choice and Welfare*, 39(4):855–867, 2012. ISSN 0176-1714.

[17] Wade D. Cook and Lawrence M. Seiford. Priority Ranking and Consensus Formation. *Management Science*, 24(16):1721–1732, December 1978. ISSN 0025-1909. doi: 10.1287/mnsc.24.16.1721.

[18] James Fishkin, Valentin Bolotnyy, Joshua Lerner, Alice Siu, and Norman Bradburn. Can Deliberation Have Lasting Effects? *American Political Science Review*, 118 (4):2000–2020, November 2024. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055423001363.

[19] Benjamin Golub and Matthew O. Jackson. Naïve Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, February 2010. ISSN 1945-7669. doi: 10.1257/mic.2.1.112.

[20] Viswanath Nagarajan and Maxim Sviridenko. On the Maximum Quadratic Assignment Problem. *Mathematics of Operations Research*.

[21] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2.

[22] Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Louis J. Podrazik, Steven G. Kratzer, Eric T. Harley, Donniell E. Fishkind, R. Jacob Vogelstein, and Carey E. Priebe. Fast Approximate Quadratic Programming for Graph Matching. *PLOS ONE*, 10(4):e0121002, April 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0121002.

[23] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.

[24] Gábor Erdélyi, Martin Lackner, and Andreas Pfandler. Computational Aspects of Nearly Single-Peaked Electorates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):283–289, June 2013. ISSN 2374-3468. doi: 10.1609/aaai.v27i1.8608.

[25] PrefLib/preflibtools. PrefLib: A Library for Preferences, February 2025.

[26] Tomasz Przedmojski. *Algorithms and Experiments for (Nearly) Restricted Domains in Elections*. PhD thesis.