UNIVERSITY OF AMSTERDAM

MASTER THESIS

THIS THESIS IS A WORK IN PROGRESS

Modelling Meta-Agreement through a DeGroot 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

July 1, 2025

Declaration of Authorship

I, Amir Sahrani, declare that this thesis, entitled 'Modelling Meta-Agreement through a DeGroot 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:

Amir Sahrani

Date: July 1, 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

I would like to thank Ulle Endriss for helping me walk the tight rope between scientific disciplines, a path I find myself walking all too often. His guidance has been of tremendous help whenever I found myself lost, confused or when I was simply getting ahead of myself.

I thank my friends, and family for keeping me sane. Though stress and tension sometimes ran high; the long talks, cards, and plastic rock have kept my mind from fraying.

Thanks, but I'll take some more fish.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements iv

Contents v

List of Figures vii

List of Tables ix

Abbreviations x

Symbols xi

1 Introduction 1

2 Preliminaries 4
- 2.1 The Basic Model . . . 4
- 2.2 Social Choice Functions . . . 5
- 2.3 Negative Results . . . 6
- 2.4 Domain Restrictions . . . 7

3 Literature Review 10
- 3.1 Condorcet Domain . . . 10
- 3.2 The History of Deliberation and Meta-Agreement . . . 12
  - 3.2.1 Deliberation . . . 12
  - 3.2.2 Meta-Agreement . . . 14
- 3.3 Models of Deliberation . . . 15
- 3.4 Deliberative experiments . . . 18
  - 3.4.1 Deliberative Policymaking . . . 19

3.4.2 America in One Room . . . 19

4 Theory 20
4.1 Our Model . . . 22
4.1.1 Consensus . . . 26
4.1.2 Voter Mapping . . . 27

5 Methods 30
5.1 Replicating Rad and Roy . . . 30
5.2 Experiments . . . 31
5.2.1 Modelling Trust . . . 31
5.2.2 Outcome measures . . . 35

6 Results 37
6.1 Replication . . . 37
6.2 DeGroot Model . . . 38
6.2.1 Policy-Based Ideology Scores . . . 39
6.2.2 Convergence . . . 42
6.3 Sensitivity Analysis . . . 43
6.4 Adding Meta-Agreement . . . 43

7 Discussion 46
7.1 Conclusion . . . 46
7.2 Discussion . . . 47
7.3 Future work . . . 49

A Ethics and Data Management 50

B Extended Proofs 51

C Nominal Values and Supplementary Figures 52
C.1 Additional Figures . . . 52

Bibliography 54

List of Figures

2.1 An election with three voters and five candidates. Each voter has a unique peak, and the profile is single-peaked with respect to a shared axis. 9

3.1 The graph of judgement sets for all preferences over three candidates, brackets indicate ties. For readability the corresponding preferences are uses as node labels 17

6.1 The proportion of cyclic profiles remaining, 0 indicating that no cyclic profiles were present after deliberation. 38

6.2 Number of unique preferences at the final step of deliberation. 38

6.3 The proportion of Condorcet winners left after deliberation, value above one indicate Condorcet winners emerging during deliberation 38

6.4 Proximity to single-peakedness after deliberation. Proximity to single-peakedness as defined in Section 3.3. 38

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. 40

6.6 Change in PBS, relative to the original, pre deliberation, measurement. The control is omitted as there was no significant change. 41

6.7 Prediction error of the model as a function of time, binned relative to the original PBS. 41

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. 42

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. 42

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. 43

6.11 The proportion of cyclic profiles remaining, 0 indicating that no cyclic profiles were present after deliberation. 44

6.12 Proximity to single-peakedness after deliberation via candidate deletion (left) and voter deletion (right). The black line is a fitted sigmoid curve. . 44

C.1 The distribution of knowledge scores for different ranges of policy-based ideology scores. 53

List of Tables

5.1 The parameters of the DeGroot learning based model, as well as their descriptions 36

Abbreviations

PBS	**P**olicy-**B**ased (ideology) **S**core
PtS	**P**roximity **t**o **S**ingle-peakedness
PtS-V	**P**roximity **t**o **S**ingle-peakedness (through) **V**oter (deletion)
PtS-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
BD	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 $\|A\| \times \|\Psi\|$, estimating support of policies for each alternative

CHAPTER 1

INTRODUCTION

“Vaccines are deadly” and “nuclear energy is dangerous” are two examples of claims that run counter to expert consensus, despite experts overwhelmingly vouching for both their safety and efficacy. Many democracies suffer this kind of misinformation, leading to a general dissatisfaction among the electorate. Misinformation not only pushes voters to more extreme opinions, but skews their views of fellow citizens. Elections face a dual challenge, not only carrying out the task of electing broadly appealing candidates, but doing this in the context where people have drastically differing opinions on the nature of the problems, their possibly solutions and the roles of the candidates.

For democracy to function effectively, voters need a shared foundation of understanding — a “shared reality.” This consists of commonly accepted facts and causal relationships allowing meaningful debate about values and priorities. For example, while nuclear energy is considered safe by experts, it comes at high initial cost, and long construction times. Renewable sources such as solar and wind, by contrast, can be scaled up quickly but provide less consistent energy output. When voters share this understanding, they can engage in productive disagreement about whether the time and money investments for nuclear are worth the consistent energy production. However, when some voters believe nuclear to be unsafe, an election seemingly about the trade-off between nuclear and solar becomes a referendum on the perceived safety of nuclear energy.

Traditionally, people’s understanding of the world was shaped by family, friends, and legacy media. These sources tend to reinforce shared viewpoints, friends and family often consumed similar media and held similar beliefs, while newspapers and broadcasters curated a common public narrative — even if this narrative is not entirely factually accurate. Increasingly, however, algorithmic curation shapes individual worldviews creating a fundamental problem: because algorithms tailor content to individual preferences, each person may encounter a unique set of claims about the world, leading to fragmented understandings of reality.

This fragmentation creates a problem for collective decision-making. Voters might be supporting the same candidate for fundamentally different, and possibly opposing, reasons. In this work we formalize this notion of a "reason" using the concept of the *Issue Dimension* introduce by List [23], when people have a common *Issue Dimension*, their disagreement over outcomes can be explained through different trade-off along these dimensions.

From the perspective of social choice, shared issue dimensions can be beneficial. In particular if the problem can be reduced to a singular shared issue dimension, we might get "single-peaked" preferences, a special structure in the preferences of voters. In Chapter 3 we define this formally. Informally however, single-peaked preferences allow for election mechanisms that encourage voters to report their preferences honestly. We elaborate on what we mean by an election mechanism in Chapter 2, but intuitively, it is a procedure for aggregating individual preferences into a collective choice.

To promote the single-peakedness of public preferences, List et al. [24] propose deliberation as a potential strategy, building on List's earlier concept of *meta-agreement* [23], being the idea that voters agree on which issue dimensions matter and where candidates stand on these dimensions. List et al. [24] argue that deliberation can help voters develop more coherent preference structures. Deliberation, then, helps restructure voters' opinions in a more coherent way, particularly on low-salience issues that receive little media coverage.

Given deliberation's potential to generate meta-agreement and more structured preferences, we aim to understand deliberation more rigoriously. With the rise of in-silico experiments in computational social science, we take a computational approach to understanding deliberation. Specifically, we adapt the classic DeGroot model of opinion dynamics [10] to the context of political deliberation, both on voter opinions and perceived candidate positions.

To this end, in Chapter 3 we review work on single-peakedness, deliberation, and experiments, and present a deliberation model by Rad and Roy [30]. In Chapter 4 we formally define some properties of deliberation, and prove negative results regarding "Honesty" during deliberation, showing deliberation is not strategyproof under a variety of circumstances. We also define an adaptation to the DeGroot model, as a mechanistic explanation of deliberation through a computational model. In doing so, we find a limitation in the applicability of this model in the form of a negative computational complexity result. Specifically, we show NP-completeness of mapping voter opinions to trust matrices. In Chapter 5 we explain the experimental setup we use to test our model, the result of which we present in Chapter 6. Finally, we reflect on the results, and broader implications of this thesis in Chapter 7.

CHAPTER 2

Preliminaries

We begin with a short introduction to social choice. We outline the basic voting model, closely following the notation and definitions by Brandt et al. [4], and restate well-known results relevant to the following chapters.

2.1 The Basic Model

To model elections, 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, usually called the alternatives. In line with the topic of political elections, we will refer to the outcomes of an election as candidates instead. Each voter;represents their preference on candidates through a preference relation $\succ_i$, for example if voter i prefers outcome a to outcome b, we write $a \succ_i b$. When a voter's preference is antisymmetric, complete and transitive, i.e. it orders all candidates and $a \succ_i b$ and $b \succ_i c$ implies $a \succ_i c$, we call this a linear order, denoted by R_i. We call the set of possible linear orders over the candidates $\mathcal{L}(A)$. For an election, all voters report a linear order. The vector consisting of each voter's preference is called a profile, denoted by $\boldsymbol{R} = (R_1, \dots R_n) \in \mathcal{L}(A)^n$. Finally, a social choice function (SCF) f decides the outcome of the election based on the profile. We discuss the specifics of these functions in Section 2.2.

The last simple definition we will need is the *majority relation* [25]. Given some profile $\boldsymbol{R}$ we can construct a majority relationship as follows: for each pair of candidates x, y, we ask how many voters strictly prefer x to y; if this number of people is greater than $\frac{n}{2}$ we get $x \succ_{\text{maj}} y$. If it is exactly equal to $\frac{n}{2}$ and thus is a tie, we simply write $x \sim \text{maj} y$ (breaking tie arbitrarily), otherwise we write $y \succ_{\text{maj}} x$. 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 $a \succ_{\text{maj}} b \succ_{\text{maj}} c$.

From this it is easy to see that the majority relation is in some sense a summary of the voter's preferences. In Section 2.3 we show how a divided population can lead to an inconsistent majority relation.

Using this majority relationship, we can formulate our first notion of when an candidate is winning. We call an candidate x a Condorcet winner, if for each pairwise comparison between x and $y \in A \setminus \{x\}$ we have $x \succ_{\text{maj}} y$ for example, in Example 1 a would be the Condorcet winner. We can relax the requirement of always winning to never losing, i.e. we never have $y \succ_{\text{maj}} x$ but $x \sim \text{maj} y$ is allowed. A candidate that never loses in any pair wise comparison is called a weak Condorcet winner.

2.2 Social Choice Functions

As mentioned, in order to decide the outcome of an election we need a social choice function f, this function should map all possible profiles to an outcome, thus $f : \mathcal{L}(A)^n \to A$. A famous and simple example of a SCF is the plurality rule, which simply elects the candidate voted into first place most often, i.e. "most first place votes wins". This rule presents on of the first challenges for many SCF, it must deal with ties.

For elections organizers likely will want to ensure the SCF has certain nice properties, such as not favoring a candidate. In social choice these properties are called axioms, and the procedure of designing a SCF based on desired axioms is called the axiomatic approach. The name of the property just described is the axiom of neutrality, stating that the SCF should be neutral with respect to the candidates. In this work six main axioms are of importance.

Axiom of Resoluteness. A SCF f is resolute, if for every profile $\boldsymbol{R}$ we have $|f(\boldsymbol{R})| = 1$.

Axiom of Surjectivity. A SCF f is surjective, if for every candidate x, there exists a profile $\boldsymbol{R}$ such that $f(\boldsymbol{R}) = x$.

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

Axiom of Strategyproofness. A SCF f is strategyproof if, for any voter $i \in N$, i cannot report an untruthful preference $\succ_i'$, such that $\boldsymbol{R}' = (\succ_1, \dots, \succ_i', \dots, \succ_n)$ and $f(\boldsymbol{R}') \succ_i f(\boldsymbol{R})$.

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

Axiom of Neutrality. A SCF f is neutral if, when the labels of the candidates are shuffled, the winning candidate in the shuffled election, is the candidate that has the ranks of the winning candidate in the original election.

There are many more axioms on could reasonably argue for however, these are enough to lead to the main impossibility results this work 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 resulting in $a \succ_{\text{maj}} b$, next voters 1 and 2 prefer b to c, resulting in $b \succ_{\text{maj}} c$. However, voters 2 and 3 prefer c to a, resulting in $c \succ_{\text{maj}} a$. This yields the cycle $a \succ_{\text{maj}} b \succ_{\text{maj}} c \succ_{\text{maj}} a$.

It can be shown that under weak preferences the Condorcet cycle can occur anytime there are 3 or more candidates and voters. While under strict preferences this can occur anytime there is an odd number of preferences at least 3, with the number of voters being a multiple of the number of candidates. 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-Satterthwaite theorem [19, 31].

Theorem 2.1. [Gibbard-Satterthwaite] There exists no resolute social choice function for elections with $|A| \geq 3$ that is surjective, strategyproof, and non-dictatorial.

Unless we accept a dictatorship, it is impossible to have a voting rule that incentivizes voters to report their preferences truthfully, when we want to pick a singular winner from at least 3 candidates.

Though we do not provide a full proof, the Condorcet cycle offers some intuition for why this result holds. Following Example 2, suppose we have a social choice function (SCF) f that elects candidate a. Voter 1 is very happy with this outcome, but voters 2 and 3 would prefer c instead. Voter 2 could then misreport their preferences by swapping c and b, thereby causing c to become the Condorcet winner.

Now, if f is both strategyproof and resolute, it must still elect a despite c being the Condorcet winner. Since f is also surjective, a cannot be the outcome for all preference profiles. Taken together, the only apparent reason a continues to win in this profile is because voter 1 wants it to—suggesting that voter 1 effectively dictates the outcome.

Fortunately, there seem to be ways around these negative results. Mainly through the assumption that there is some structure in the preferences of voters.

2.4 Domain Restrictions

Negative results often are a result of a small set of ill-behaved profiles. If there is reason to conclude these profiles are impossible in the election at hand, 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. [13].

Definition 1: *Domain*

Given a set of voters N, candidates A, and conditions C, the domain $\mathcal{D}$ of an election is the set of all profiles $\boldsymbol{R}$ 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 candidate, holds a majority preference over another candidate. Naturally one might consider if this profile might even come up in practice, though conceivable, it seems generally unlikely for there to exist 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 candidates, though this is hardly a useful solution since it only prevents

Condorcet cycles, it is the first example of a domain restriction, we define a simple domain that prevents these cycles as follows.

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

Let A be the set of candidates and N be the set of voters, of size n such that $n \neq k \cdot |A|$ 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 an $|A|$-way tie between candidates when applied to $\mathcal{D}_{\text{No-tie}}$.

Proof. Assume, for the sake of contradiction, the plurality rule in fact does return an $|A|$-way tie, this means all candidates were ranked first an equal number of times call this k. Necessarily then, we need exactly $k \cdot |A|$ 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 [18] establishes two ways in which a domain can be restricted. Firstly we can restrict the domain to a number of voters or candidates, 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.

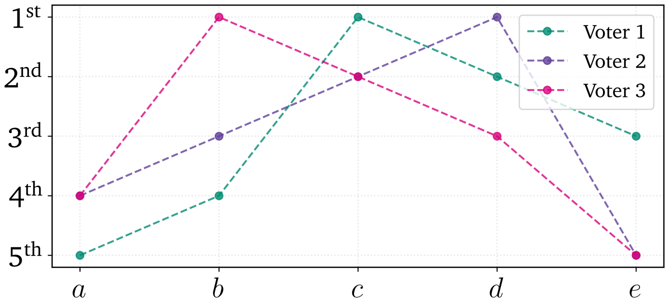
In an election the candidates might represent an axis, such that a voters prefers an candidate more if they are closer to them on the axis. For example, if the candidates represent the minimum wage, where each cent-value constitutes an candidate. Imagine a voter thinks the minimum wage should be some value x and prefers candidates that are closer to this value x. This results in each voter having a "peak" value, and all other values are ranked in terms of their distance to x. Figure 2.1b shows what this might look like for 3 voters. More generally, we call a profile single-peaked if there exists an axis on which we can place the candidates 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 $\boldsymbol{R}$ is single-peaked, if given some ordering $\lhd$ over the candidates, it holds that for all voters i, and all $a, b, c \in A$, if $a \lhd b \lhd c$, then at most $a \succ_i b$ or $c \succ_i b$, but never both.

This thesis will now focus on measures to "increase" single-peakedness of profile.

1	2	3
c	d	b
d	c	c
e	b	d
b	a	a
a	e	e



(A) Preference profile

(B) Single-peaked profile visualization

FIGURE 2.1: An election with three voters and five candidates. Each voter has a unique peak, and the profile is single-peaked with respect to a shared axis.

CHAPTER 3

Literature Review

In this chapter we review the theoretical foundations that inform our computational approach to modeling deliberation. We examine three interconnected areas: domain restrictions in social choice theory (particularly single-peaked preferences and hereditary domains), the literature on deliberation and meta-agreement theory, and computational models of deliberative processes. Together, these establish both the theoretical motivation for understanding how deliberation can produce well-structured preference domains and the methodological foundation for our computational modeling approach.

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 $\boldsymbol{R}$ such that $\boldsymbol{R}$ 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 $\mathcal{D}_{\text{Condorcet}}$ [13].

> *Proof*. (Elkind et al. [13]). Assume, for the sake of a contradiction, we have profiles $\boldsymbol{R} = (\succ_1, \ldots, \succ_i, \ldots, \succ_n)$ and $\boldsymbol{R}' = (\succ_1, \ldots, \succ_{i'}, \ldots, \succ_n)$ such that:
>
> $$f_{Condorcet}(\boldsymbol{R}) = a, \quad f_{Condorcet}(\boldsymbol{R}') = b, \quad \text{and } a \neq b$$
>
> Assume that i has $b \succ_i a$, thus strictly prefers b to a. Then under $\boldsymbol{R}$ there is a strict majority $C \subseteq N$ who have $a \succ b$, but $i \notin C$. Thus, in $\boldsymbol{R}'$, C is still a majority preferring a to b, making a the Condorcet winner in $\boldsymbol{R}'$. This is in contradiction to b winning in $\boldsymbol{R}'$.

This result is strengthened by Campbell and Kelly [5, 7], showing that for an odd number of candidates, $f_{\text{Condorcet}}$ 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-dictatorship to anonymity, $f_{\text{Condorcet}}$ is the only strategyproof voting rule over $\mathcal{D}_{\text{Condorcet}}$ for an odd number of voters [6].

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 candidates or voters drops out, do we keep the nice structure of the domain? If this is true we consider the domain stable and call it *hereditary*.

Definition 4: *Hereditary* (Elkind et al. [13])

A domain $\mathcal{D}$ is called *hereditary* if for every profile $\boldsymbol{R} \in \mathcal{D}$, and every subprofile $\boldsymbol{R}'$ obtained by deleting voters and candidates from $\boldsymbol{R}$, it holds that $\boldsymbol{R}' \in \mathcal{D}$.

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

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

1	2	3	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. If we remove voter 4 however, we return to the original Condorcet cycle.

If a domain fails to be hereditary, designing an election with the domain is mind becomes hard. In the case of $\mathcal{D}_{\text{Condorcet}}$ it might be reasonable to make use of a rule such as Black's rule [2], which uses the Condorcet rule only if there is a Condorcet winner and the Borda rule otherwise. This however is not a strategyproof voting rule in general. Instead, we might want to look at hereditary strategyproof domains. We present the single-peaked domain $\mathcal{D}_{\text{SP}}$, which will be the main focus of this thesis. This is the domain of all single-peaked profiles. We first proceed to show that this domain indeed is hereditary.

Proposition 3.1. (Elkind et al. [13]). $\mathcal{D}_{\text{SP}}$ is hereditary.

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

(Candidate Deletion). Consider any voter i and their single-peaked vote, if we remove

> some candidate x, to this voter all candidates which they preferred to x stay in the same position, while all other candidates move up one rank, thus preserving the order, and thus single-peakedness. ✓

We have demonstrated that $\mathcal{D}_{\text{SP}}$ possesses the desired properties. However, we currently lack a method to ensure that we operate within $\mathcal{D}_{\text{SP}}$. Deliberation may provide a mechanism to ensure that preference profiles move toward single-peakedness. We will now provide a concise overview of the literature on 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. Bochsler [3] argues 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

Deliberation, though intuitively familiar as 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.

Instead of defining deliberation in full generality, we instead focus on deliberation in a political sense. Freeman [17] gives an overview of deliberative democracy. He notes that there is no settled definition of deliberative democracy, however, one account is that of public discussions before voting. Furthermore, he shares the intuitive idea that a deliberative democracy contains open legislative deliberation and a pursuit of the common good. 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 for 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

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 [8] 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, as shown through the Condorcet cycles, it is simply a reflection of the particular institutional constrains at the time, which end up dictating where the cycle breaks. He further goes on to describe the *ideal deliberative procedure* as follows:

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*. Participants must state their reasons for supporting 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 rationally defensible *consensus*.

From both Cohen's and Freeman's account there is clear overlap, with Freeman formulating the necessary preconditions for participants to engage in ideal deliberation. Both Cohen and Freeman require freedom in a broad sense. Freedom to have a personal conception of the good, and to acknowledge and act in accordance to a decision that was made through deliberation.

3.2.2 Meta-Agreement

Consensus, sometimes referred to as substantive agreement, then seems like a natural goal for deliberation. Elster [14] argues that this is not only the goal, but through unanimous agreement this process completely replaces voting, thereby circumventing social choice's classic impossibility theorems: "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 these negative results, in practice people, even after deliberation, might not and indeed often do not come to full substantive agreement. List [23] 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 [23] 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 candidates in this *issue*-dimension

D3 After deliberation, people update their preferences picking a preferred outcome, and ranking all other candidates based on the distance to this outcome in the *issue*-dimension

These are necessary conditions for *meta-agreement*. From this is it also clear to see that, given that there is exactly one *issue*-dimension, single-peaked profiles are, by definition, a direct consequence. This property of inducing single-peakedness makes meta-agreement particularly desirable, as it enables circumvention of the Gibbard–Satterthwaite theorem [19, 31] through domain restriction to $\mathcal{D}_{\text{SP}}$.

List et al. [24] 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 PtS increases following deliberation.

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.

Furthermore, Ottonelli and Porello [28] show single-peakedness from 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 two semantic issues, with only one 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 implicit 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 candidate. Finally, D3 requires D1 and D2 to have happened before in order, indeed this is the weakest of the three requirements.

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 agreement might still play a crucial part in a deliberative process. In the next section we will look into a specific computational model of deliberation.

3.3 Models of Deliberation

Rad and Roy [30] model deliberation and its effect on single-peakedness. 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 have a bias towards their own preference, as such they try to update their preference by minimizing the distance between their current preference

[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.

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) [22], Duddy-Piggins (DP) [11], and Cook-Seiford (CS) [9]. Both KS and DP depend on the judgement set resulting from the voters preferences, which is contains, for each pair of candidates 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 define their betweenness as follows.

Definition 5: *J-Betweenness*

A judgement set J_i is between preferences J_j and J_k if for every $x, y \in A$, the proposition over x and y in J_i either agrees with the proposition over x and y in 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 $\neg(a \succ b)$.

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 candidates, for example if two voters agree on the relative ranking of all but one candidate, 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 case of many disagreements, 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, Figure 3.1 shows a graph used for the DP distance in the case of 3 candidates. The graph shows the benefit of using the DP distance, as the edges in graphs automatically include logical consequences that the KS distance might not account for. In doing capturing logical consequences, the DP distances becomes much harder to compute, mainly through the cost of constructing the full graph of judgement sets, which grows in $f_m = 1 + \sum_{j=1}^{m-1} \binom{m}{j} f_{n-j}$ in the number of vertices, where m is the number of candidates [21]. This can easily be verified by noting that the number of judgements sets over m corresponds to the number of weak preference rankings over m candidates, which is defined as candidates, and a binary choice on each proposition.

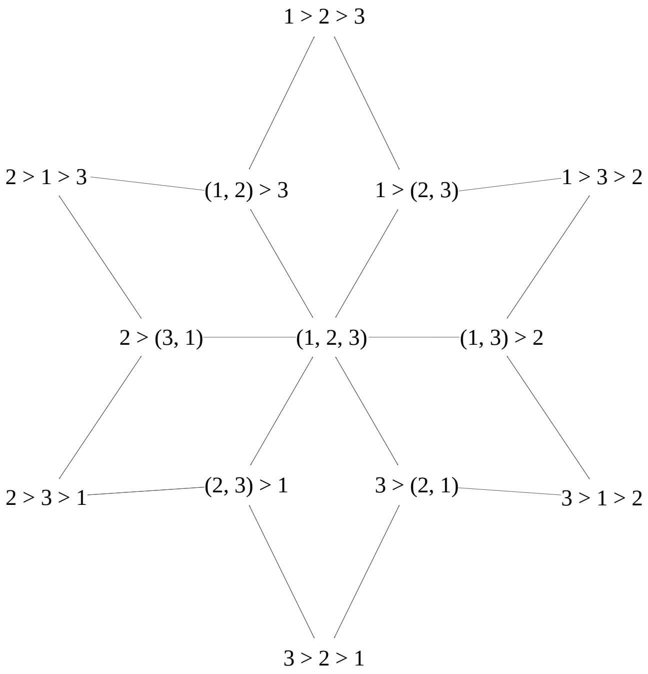


FIGURE 3.1: The graph of judgement sets for all preferences over three candidates, brackets indicate ties. For readability the corresponding preferences are uses as node labels

Apart from these distances, Rad and Roy define a voter as a tuple of a linear order[2] and a bias $v = \langle r, b \rangle$, with $b \in \mathbb{R}_{[0,1]}$. Finally, a deliberation step $D_s : V^n \rightarrow V^n$, where V is the set of all possible voters $(\mathcal{L}(A) \times \mathbb{R}_{[0,1]})^n$ and s being one of the spaces (KS, DP, CS). The deliberation step $D_s(V, v.r)$ returns a fully updated voter set, where each voter has updated their opinion in response to the announced opinion $v.r$. We formulate this procedure in the following program:

[2] We exclude their analysis of preferences containing ties.

```
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
```

Here, we use $v.r$ to denote the preference component of voter $v = \langle r, b \rangle$. The deliberation step $D_s(V, v.r)$ returns a new set of voters, where each voter updates their opinion based on v's preference r, under the influence of the deliberation space s. Each voter updates their preference to a new profile r' that minimizes the weighted distance between their original preference r_i and the announced preference.

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

Here b is this voter's bias, and d_s is the distance between two profiles under distance space s.

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

While this model effectively captures preference communication, it falls short as a model of meta-agreement in at least two important respects. 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.

3.4 Deliberative experiments

We now present some empirical studies showcasing the effects of deliberation in voting populations, focusing on deliberative policymaking, and the America In One Room experiment.

3.4.1 Deliberative Policymaking

3.4.2 America in One Room

Fishkin et al. [16] 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, The 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 [30], 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 2.1.

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

DEFINITION 6: *Biased Deliberation*

A deliberative procedure with biases BD : $\mathcal{L}(A)^n \times \mathbb{R}^n_{[0,1]} \to \mathcal{L}(A)^n$ is an extension on a standard deliberative procedure. BD 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 BD to general deliberation D.

> *Proof.* Take any election consisting of biased deliberation BD 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 BD, resulting in $D'(\boldsymbol{R}) = \text{BD}(\boldsymbol{R}, \boldsymbol{b})$. 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, as defined by Rad and Roy [30], 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 B.

> *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 candidates 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 candidates. 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 candidate x.

This model does not explicitly model $D1$, the discovery of a common issue dimension, on the one hand, if the candidates 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 candidates, but not on the voters, this model allows for the voters to recognize this by properly estimating the candidates' 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 candidate can only support a specific number of policies, in principle there can be candidates that represent the status quo, and thus do not support any policies, and there can be candidates 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 candidates such that the most preferred candidate has the smallest distance between the estimated support matrix for that candidate 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 candidates 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 candidate, 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 candidate a is 0.5, while for candidate b it is 1.5. For voter 2 we see that the distance to a is 0.9, while for candidate 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:

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

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 candidate, as they both agree that candidates 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 [20]. Firstly we focus on the strongly connected graphs.

Proposition 4.4. (Golub and Jackson [20]). For a strongly connected matrix T, the following properties are equivalent:

- T is Convergent
- T is Aperiodic
- 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

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

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 $C = \{1, \dots, n\}$ 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 [20]). 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. ($\implies$ 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 [27]. Thus solving δ-DBVM(S), on the binary domain, is equivalent to solving 0-1 Max-QAP, and thus NP-Hard. ✓

($\implies$ NP-Membership) Given any f, we can evaluate the cost of the allocation 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 [27].

Despite these negative results, we attempted to enlist the help of a QAP-solver [32] to find (approximate) solutions, using the Fast Approximate QAP Algorithm [33]. 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 [30], as well as the experimental setup of our own model. Links to the data used for these experiments can be found in Appendix A. The programs are implemented using `OCaml`, and `Python`.

5.1 Replicating Rad and Roy

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 candidates, and 51 voters. The number of voters is chosen to be an odd number, as this prevents ties between candidates. 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), all as reported by Rad and Roy [30]. In addition to those we also measure the frequency of Condorcet winners. We do this to see if cyclic profiles might be hiding some agreement, where voters might be able to agree on a winner, but not on the full ranking. We do not measure the PtS-C, as any profile with 2 candidates is single-peaked. Since the simulation will have 3 candidates, all values would be either 1 or $\frac{2}{3}$, therefore this metric would be of little added value. Due to the computational complexity of the DP-metric, as well as the calculation of PtS-V, a larger number of candidates is not feasible. Specifically PtS-V is NP-complete [15], though it allows for a 2-approximation. We use the method based on an ILP solver, as implemented in `PrefTools` [1]

Using these results we aim to show the limitations of preference based deliberation.

5.2 Experiments

We aim to replicate the findings by the America in One room experiments [16] 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. 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.

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 Outcome measures

We split the experiments on the adapted DeGroot model into two parts. Firstly, we aim to assess the validity of the model. For this we use data from the America in One room experiment. This data does not provide full preference rankings over the candidates, instead provides data on voters' opinions on 6 different topics of political discussion, such as climate change and immigration. Using these opinions, we assess the validity of the model in so far as it is able to accurately predict the change in opinion. We run 5000 simulations, randomizing the independent variables laid out in Section 5.2.2, excluding `Candidates`, and `Candidate Generator`. We then use an ANOVA to test for the configuration of trust matrices that minimizes the errors in predicted PBS. 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.

Using these results we expand the model to incorporate meta-agreement on alternatives, for this we need to generate potential candidates. This is done either by selecting a voter and creating a candidate with identical opinions, or by selecting 10 voters with replacement and creating an average of their opinions. Using these simulated candidates we are able to create preference rankings using the ℓ_1-norm. We measure whether the same outcomes as in the Rad and Roy replication section, namely whether the final profiles are cyclic, whether they have a Condorcet winner, home many unique profiles there are, and their PtS-V, and finally we also measure PtS-C. The PtS-C can be calculated in $O(|V| \cdot |C|^3)$[29], though the implementation we use is that of the `PrefTools` library [1], which implements a slower $O(|V| \cdot |C|^5)$ algorithm [15].

Finally, we use sensitivity analysis to investigate which parameters have the strongest effect on the variance of model, using Sobol indices to get the first, second and total order effects. The first order indices refer to their direct effects on the variance of the model, when all other parameters are randomized. The second and total order capture this for 2-way interactions of variable and all interactions of a variable respectively.

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

CHAPTER 6

RESULTS

We first present a full replication and extension of the work by Rad and Roy [30]. 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 C 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 [30], 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.

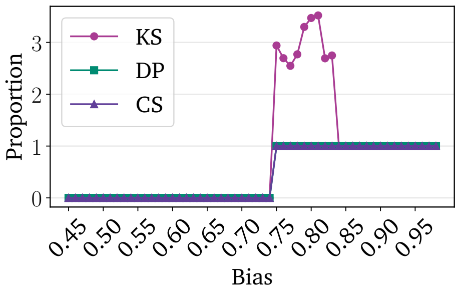


FIGURE 6.1: The proportion of cyclic profiles remaining, 0 indicating that no cyclic profiles were present after deliberation.

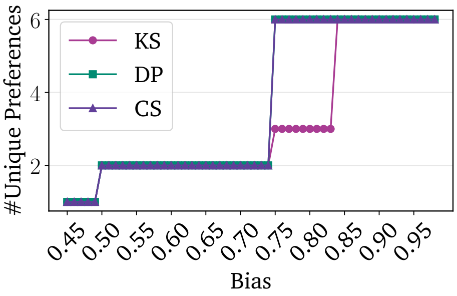


FIGURE 6.2: Number of unique preferences at the final step of deliberation.

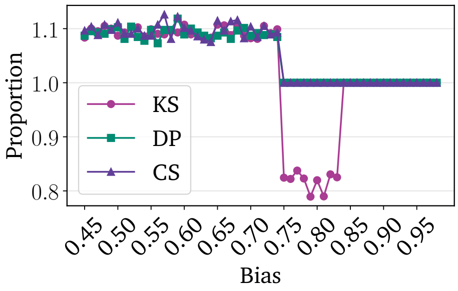


FIGURE 6.3: The proportion of Condorcet winners left after deliberation, value above one indicate Condorcet winners emerging during deliberation

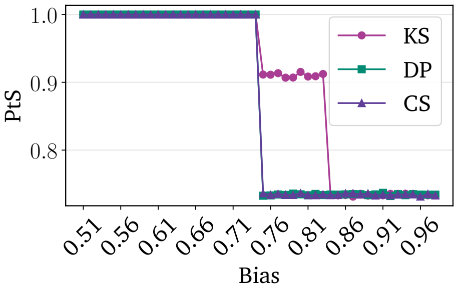


FIGURE 6.4: Proximity to single-peakedness after deliberation. Proximity to single-peakedness as defined in Section 3.3.

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 3.4.2, 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 Elster [14], 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.

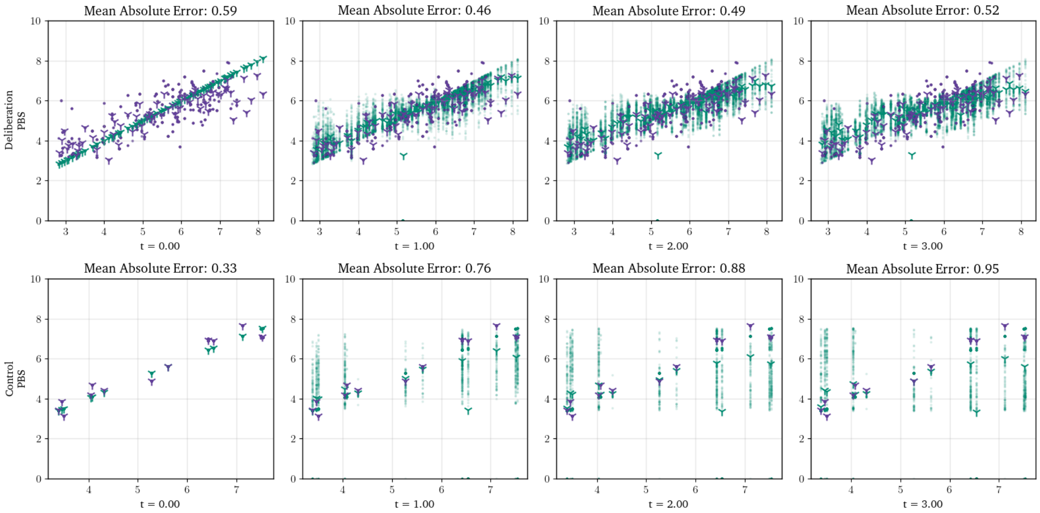


FIGURE 6.5: PBS, purple indicating the PBS after deliberation in the original data, green indicates the results of the simulation in that time step. Large dots indicate the binned data, smaller dots indicate individual voters.

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. [16], 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 C.1 in Appendix C 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 the economy.

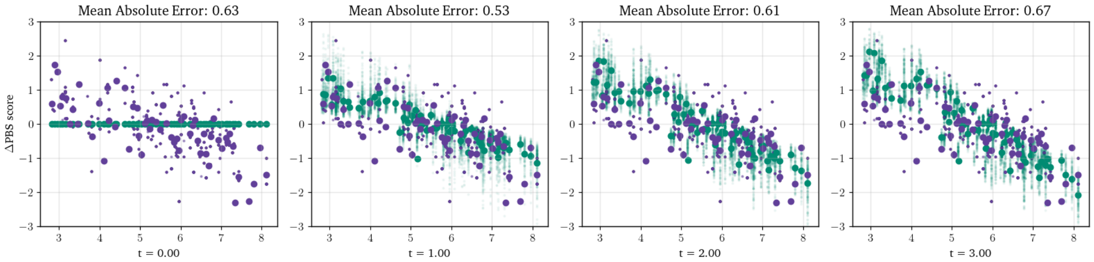


FIGURE 6.6: Change in PBS, relative to the original, pre deliberation, measurement. The control is omitted as there was no significant change.

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.

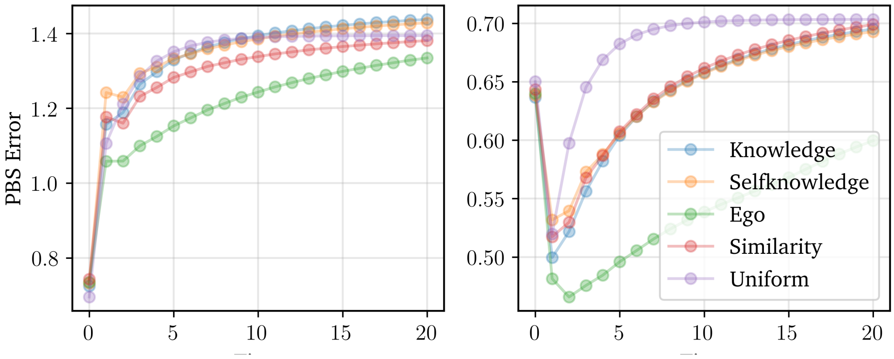


FIGURE 6.7: Prediction error of the model as a function of time, binned relative to the original PBS.

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.

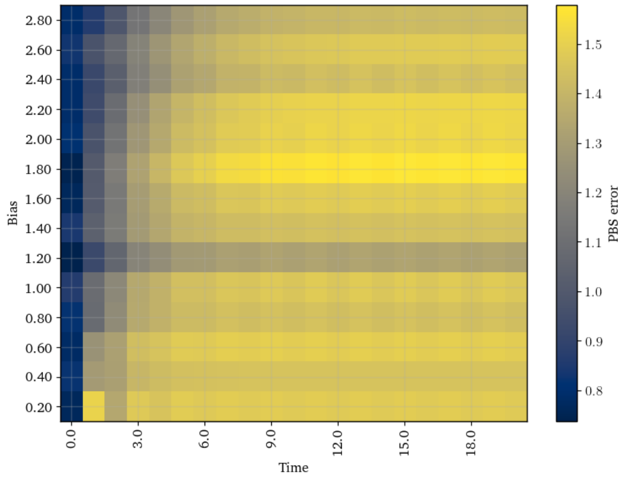


FIGURE 6.8: PBS Errors as a function of bias and time. Bias acts as a damper: when bias is higher the model take longer to over-estimate the change in opinion.

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.

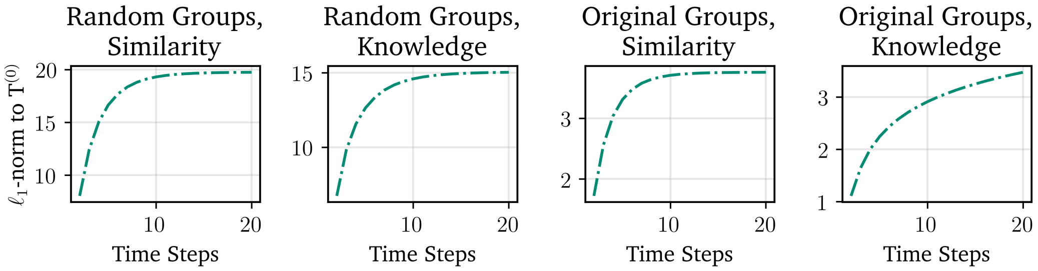


FIGURE 6.9: Convergence of trust matrices, as measured by the ℓ_1-norm between the trust matrix at the start and trust matrix at the current time step.

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.

We argue the non-significant first order indices are a result of these parameters not directly incorporating new information into the model, and thus on average they do not affect on the outcome. When these parameters are used in combination parameters that do introduce new information into the model they start to significantly alter the outcome of the model. As partly supported by the second order sensitivity indices

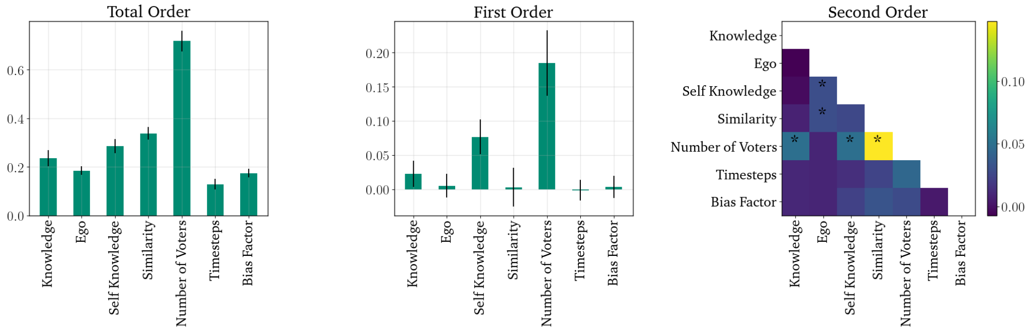


FIGURE 6.10: First, Second and Total sensitivity indices on the PBS prediction error. The stars in the heat map for the Second order sensitivity indices indicate significant interactions.

6.4 Adding Meta-Agreement

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.

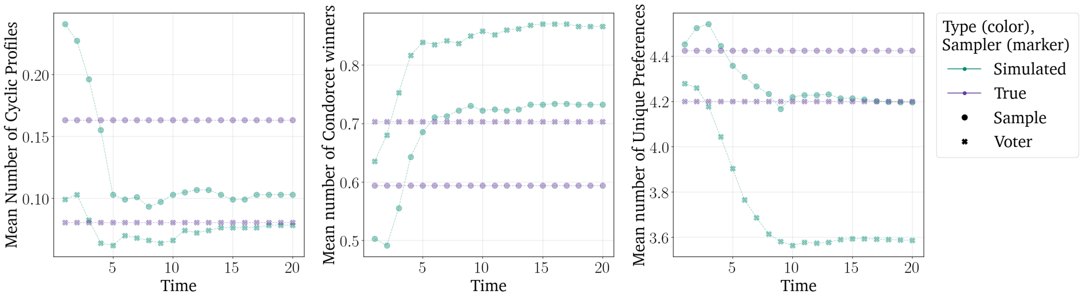


FIGURE 6.11: The proportion of cyclic profiles remaining, 0 indicating that no cyclic profiles were present after deliberation.

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.

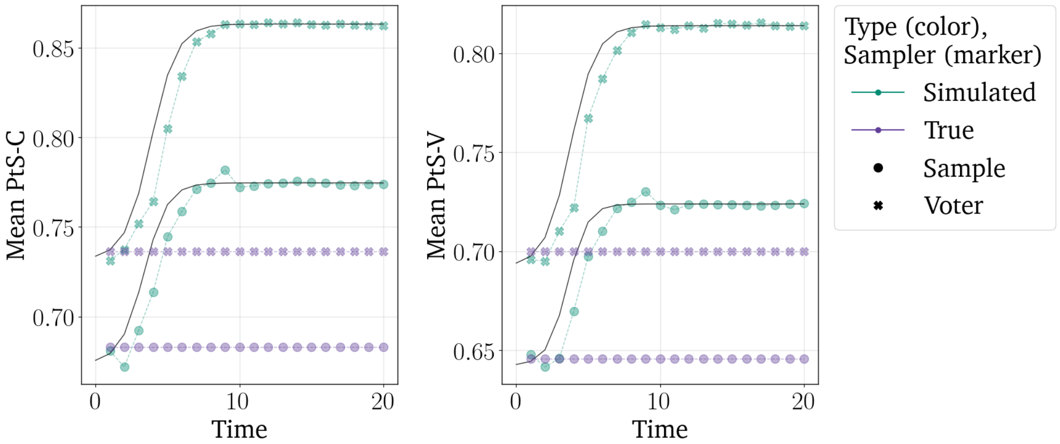


FIGURE 6.12: Proximity to single-peakedness after deliberation via candidate deletion (left) and voter deletion (right). The black line is a fitted sigmoid curve.

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

The main goal of the thesis was to get a deeper understanding of deliberation and its effect on preference profiles. To this end we consulted the literature (Chapter 3) laying out various points of view on the goal of deliberation. From this we follow Cohen's [8] four tenants of deliberation; deliberation should be *free, reasoned, equal*, and it should aim to reach *consensus*. In Chapter 4 we show that the deliberative procedure posited by Rad and Roy [30] cannot be strategyproof under classic notions of strategyproofness as well as novel notion of strategyproofness we define. We use this add one more tenant to Cohen's four, namely *honesty*.

We then set out to mechanically understand deliberation. For this we introduced the DeGroot learning model, and adapted it to deliberation over opinions. We showed NP-hardness on the δ-DBVM(S) problem, and concluded that using de DeGroot model to model sparse graphs is computationally difficult, if one wants to assign voters to nodes based on some distances metrics.

In Chapter 6 replicated the results by Rad and Roy [30], and we use our adapted DeGroot model to test its predictive power on opinions using the America in one Room dataset [16]. We conclude that though in the first time step the model can do well on the population level, the prediction on the change in opinion for individuals was poor. We also show that this is at least partly explain by the fact that the DeGroot model treats all policies equally. The data showed that some topics had large shifts in opinions, while others showed less. The DeGroot model was unable to capture this.

Using sensitivity analysis we showed that all parameters effected the final predictions, but interestingly some parameters had non-significant first-order effects. We argue that this is a result of these parameters not introducing new information. As a result they can only affect the variance of the model by modulating the dynamics induces by the parameters with significant first-order effects.

Finally, we looked at the preference profiles which we simulated based on the opinions from both the data and the simulations. We show, that similar to the population level predictions for the PBS, the profiles based on the simulated and true opinions start looking more similar during the first steps in the simulation. However, after this the model converge too strongly and the profiles of the simulated opinions become too “nice”, in the sense that they get closer to being single-peaked and are acyclic more frequently.

These results led us to conclude that the DeGroot learning model was overly simplistic and therefore was unable to adequately explain individual opinion change. As a result it is a bad approximation of what happens during human deliberation. These patterns are also in contradiction to known results in social psychology, where small extreme groups tend to become more extreme [26].

7.2 Discussion

We first present some limitations of these results. We can broadly put these into three categories.

Firstly, given the lack of a complete data source combining pre- and post-deliberation opinions and preference rankings as well as the opinions of these alternatives, we have had to make many assumptions on both the positions represented by the candidates, and well as the method by which voters generate their preferred rankings over them. In terms of generating candidates, our approach is simple, and only assumes that candidates represent the opinions held by the voters. This is however clearly a less rich process than that by which real-world candidates are selected, where these might bring in new opinions or have traits that are desirable, such as being good leaders or well-spoken. In terms of voters creating a ranking over alternatives, we have gone with the assumptions that this is done strictly through distance in opinions, similar to what a political compass test might do. In reality however, voters might be using different and multiple heuristics to order the candidates. Indeed if there are numerous candidates, the ranking might not even be complete. Therefore, distance-based measures will likely diverge from heuristics, such as pre-selecting some list of candidates deemed acceptable.

Secondly, there are some methodological assumptions we made. These mainly relate to the generation of the trust matrices. For all Knowledge, Self-Knowledge, and Similarity the scores were normalized to be between 0 and 1, while the Ego score was not normalized. This results in an asymmetry that allows Ego to increase the values in the trust matrix, where the other parameters could not. This decision was made as we found no clear ceiling with respect to which we could normalize the Ego score. As mentioned in Chapter 6, this might explain why Ego resulted in the lowest error on the Population level.

The same trust matrix was used for substantive and meta deliberation. Though from a modeling perspective this is a pragmatic solution. In reality this assumption seems too strong. This assumption forces someone to be equally willing to change their opinion as they are to change their perception of a candidate's opinion, where, at least intuitively, one might expect more willingness on the latter towards people with dissimilar opinions.

Apart from these limitations in generating the trust matrices, we also note the noise added to the estimates of candidates' opinions is normally distributed. Though this was done to introduce voter uncertainty, over which they could then deliberate, normally distributed noise seems unlikely, especially for voters that hold more extreme positions. Here we might expect that the noise is dependent on the candidates opinions, where candidates that are more similar in opinion to the voters, will be more accurately estimated than dissimilar candidates. For these dissimilar candidates, it might then also be true that this noise is skewed towards the opposite extreme w.r.t. the voter's opinion.

While we opted for the DeGroot model as a more accurate representation of human belief updating than full Bayesian updating, the DeGroot model does have some inherent limitations. Firstly, it does not take into account why people hold certain beliefs, nor does it constrain 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 [12], as this is able to model the arguments with the deliberative groups. Though, this is theoretically nice, as it allows for formal description on why opinions and preferences are held, not just their descriptions. From a simulation perspective, such a framework introduces major validity questions. 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 update their believes using this framework is unclear. Secondly, it limits voter's belief updates to linear transformations.

Finally, we address some limitations on the real-world implications of these results. The negative results surrounding strategyproofness in Chapter 4 might be less of an issue in human deliberation, as the dishonest participant could be less convincing defending their dishonest opinion than their true opinion. As a result they might have less total influence than if they had defended their true opinion.

In terms of modeling deliberation, we have now focussed on variables that can clearly be measured. While this might paint a good picture of the quantitative aspects of deliberation, in practice deliberation in humans come with rich interactions affecting their judgement and willingness to listen among other things. If we hope to get an accurate mechanistic model of deliberation, these qualitative aspects of deliberation need to be studied.

7.3 Future work

Based on the limitations of this study, and the literature it was based one, we present some areas for future work.

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. Most importantly, it needs to be able to show non-linear affects, and be informed by qualitative descriptions of deliberation. One main improvement of the DeGroot model specifically could be to introduce dynamic trust matrices. When humans deliberate, the amount of trust placed on each person is likely not fixed over time. This can be addressed dynamic trust matrices that update according to voter's familiarity with other voters, and possibly other factors.

Another way in which the trust matrices can be further refined is through introducing topic-dependent trust. As some topics might be more hotly debated, for example as a result of some recent event. These voters could generally be more informed on these topics, and less willing to talk about other topics. This is related to the notion of *Salience* as described by List et al. [24], stating that topics with high salience benefit less from deliberation, as participants have likely received more information on this topic.

Furthermore, any good model will need proper data, as such a study similar to that of Fishkin et al. [16] is needed, where voters are asked not only for their opinion but also their preference order. This could also be a great opportunity to gather qualitative insights into deliberation and the social dynamics thereof. This would also allow for testing participant's knowledge on topics directly, hopefully giving stronger indications of voter's ability to persuade and defend on specific topics.

APPENDIX A

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 B

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 C

NOMINAL VALUES AND SUPPLEMENTARY FIGURES

C.1 Additional Figures

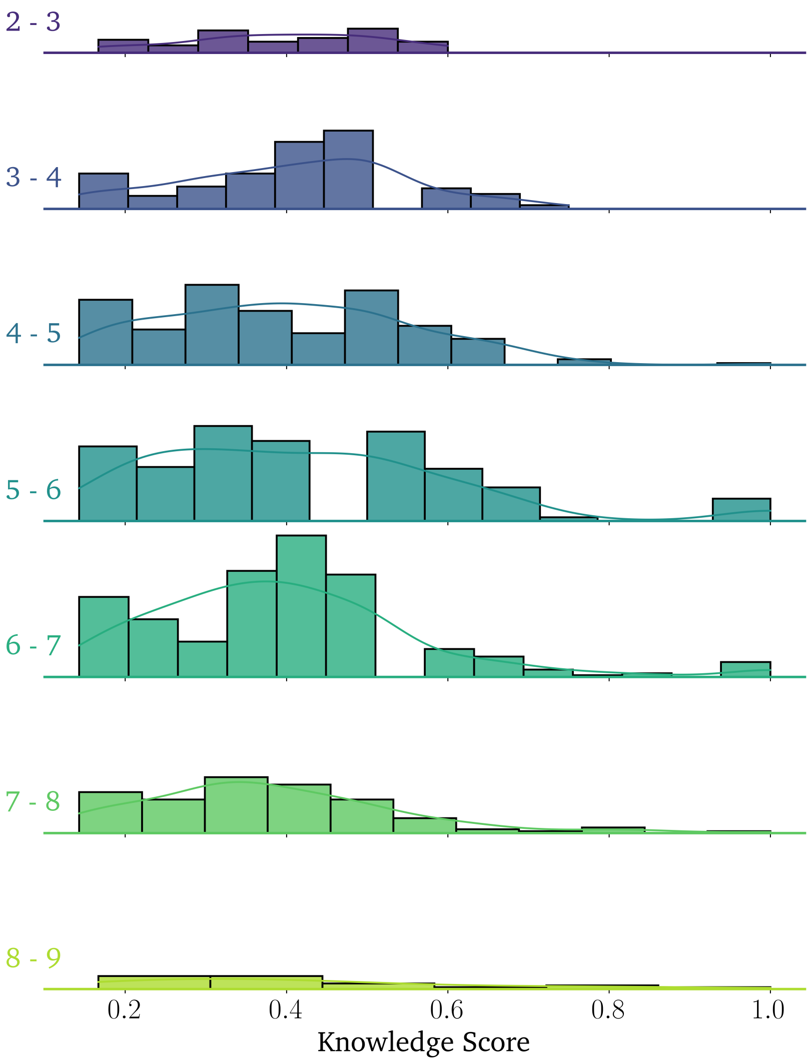


FIGURE C.1: The distribution of knowledge scores for different ranges of policy-based ideology scores.

Bibliography

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

[2] Duncan Black. On the Rationale of Group Decision-making. *Journal of Political Economy*, 56(1):23–34, February 1948. ISSN 0022-3808. doi: 10.1086/256633.

[3] Daniel Bochsler. The Marquis de Condorcet goes to Bern. *Public Choice*, 144(1): 119–131, July 2010. ISSN 1573-7101. doi: 10.1007/s11127-009-9507-y.

[4] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jerome Lang, and Ariel D Procaccia. Handbook of Computational Social Choice. *Handbook of Computational Social Choice*, 2016.

[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. 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.

[7] 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.

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

[9] 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.

[10] Morris H. DeGroot. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974. ISSN 0162-1459. doi: 10.2307/2285509.

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

[12] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, 77(2):321–357, September 1995. ISSN 0004-3702. doi: 10.1016/0004-3702(94)00041-X.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] O. A. Gross. Preferential Arrangements. *The American Mathematical Monthly*, 69 (1):4–8, 1962. ISSN 0002-9890. doi: 10.2307/2312725.

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

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

[24] 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.

[25] Marquis de Marie Jean Antoine Nicolas Caritat Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. In *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Royale, Paris, 1785.

[26] D. G. Myers and H. Lamm. The polarizing effect of group discussion. *American Scientist*, 63(3):297–303, 1975. ISSN 0003-0996.

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

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.