Expanding our Participatory Democracy Toolkit using Algorithms, Social Choice, and Social Science

Bailey Flanigan

CMU-CS-24-130

May 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Ariel Procaccia (Harvard), Chair
Nihar Shah
Anupam Gupta (New York University)
Nika Haghtalab (UC Berkeley)
Ashish Goel (Stanford)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © 2024 **Bailey Flanigan**

This thesis was supported by an NSF GRFP Fellowship, a Siebel Scholarship, and a Fannie and John Hertz Foundation Fellowship.

Keywords: Social Choice, Algorithms, Voting, Deliberative Democracy, Sortition

For the storytellers: those who generously share their experiences in order to educate, inspire, and advocate for the interests of others.

Abstract

In most of the world's democracies, policy decisions are primarily made by elected political officials. However, under mounting dissatisfaction with representative government due to issues ranging from social inequality to public distrust, a new proposal is taking off: to augment representative democracy with mechanisms by which the public can *directly participate in policymaking*.

The guiding application of this thesis will be one particular model of participation, *deliberative minipublics* (DMs), though we will argue that our contributions may apply to many models of direct participation. In a DM, a panel of citizens is selected by lottery from the population; then, this panel convenes around a particular policy issue to study background information, deliberate amongst themselves, and then weigh in on the issue. DMs have been gaining momentum over the past decade, and they are now being used at national and supranational levels, and are even being integrated into representative governments.

Motivated by this application domain, we make the following main contributions: In **Part I**, we design algorithms for performing the random selection of DM participants, a process known as *sortition*. Our sortition algorithms permit users to make optimal trade-offs between descriptive representation and other desirable properties conferred by randomness, and we characterize these tradeoffs using game theory, optimization, and empirics. In **Part II**, we use a novel social choice theory framework to investigate a notion of representation that departs from descriptive representation in a key way: it accounts for the political reality that people may be affected to *widely varying degrees* by any given policy decision. In **Part III**, we study an important hypothesized impact of deliberation: increasing the extent to which participants consider how *others in their society* may be affected by different policies. In Part **IV**, we highlight how the enclosed research illustrates new ways to combine tools from political science and computer science.

Acknowledgments

The person who I foremost want to thank is Ariel Procaccia, who has been truly been the best advisor and collaborator I could have asked for. During PhD application season, working with Ariel was my ultimate goal. When I got to CMU, I found that the reasons for my excitement held true: Ariel is indeed a master of asking the right question, defining the right model, and writing papers that cut to the heart of the issue. What I have appreciated by even more over the past five years, though, are our discussions: we have debated research directions, motivating arguments, political scientific theories, potential faculty jobs, paper titles, and even books. I feel so lucky to have spent five years being challenged in my thinking by someone so astute; to debate and usually lose, but always be welcome to come back and try again (usually after writing 40 pages of Overleaf notes). Thank you Ariel for always believing in me, pushing me to sharpen my perspective, and making everyone else look bad by responding to my emails in an average of 3 minutes.

This thesis has also been made possible in large part by the other incredible teachers and mentors I have had along the way. Foremost among them is Paul Gölz, who during the first few years of my PhD spent countless hours after each research meeting, patiently answering all of my questions about what on earth we were talking about. I am continuously grateful for your willingness to let me pick your brain, and you remain one of the technical experts I trust the most. In a similar vein, I want to thank Anupam Gupta as another key teacher in my path, having taught me much of the algorithmic thinking that I still often use today. I want to thank Mor Harchol-Balter for being a wonderful friend, role model, and unconditional supporter for me throughout the many ups and downs of my PhD; Zico Kolter, for taking on CS-JEDI and always being on my team; Vasilis Gkatzelis, for introducing me to and supporting my initial interest in theoretical computer science; Michael Neblo and Kevin Esterling, for supporting my work in political science from the beginning; Steven Wright, for believing in my potential long before I did (and always being up for an evening of pep talks and way too many appetizers); John Puccinelli, for having endless time to support me through my undergraduate degree; and Matt Bowman, for being one of my most enduring teaching role models. Finally, I want to thank my committee members, Anupam, Nika, Ariel, Nihar, and Ashish, for the multitude of ways in which they have each supported me throughout my PhD.

I also want to acknowledge the wonderful community by which I was surrounded throughout my PhD. Chief among them are my extremely talented collaborators, who not only made this research possible, but also made it incredibly fun. I look forward to many more years of enjoying the friendships we forged through fixing last-minute bugs in proofs, pulling together last-minute EC submissions, and going for beers to celebrate wins. I am also forever grateful to the Hertz Foundation for introducing me to Alex Atanasov, Dolev Bluvstein, Robbee Kosak, Philip Welkhoff, Ben Eysenbach, and Katherine Van Kirk, who I could always count during graduate school for equal parts support and fun. In the past year on the job market, I have been so grateful to the friends from our community who have shown up for me, helped prepared me, and commiserated with me, including Lily Xu, Kira Goldner, Jess Sorrell, Kate Donahue, Serina Chang, Serena Wang, Sean Sinclair, Alex Jacquillat, Jessie Finocchiaro, Elisabeth Paulson, Thodoris Lykouris, Ellen Vitercik, Victoria Dean, and many others. Finally and most importantly, I have to thank my core CMU PhD squad, Sara McAllister, Catalina Vajiac, and Ananya

Joshi: you have been my closest friends, strongest supporters, and partners in crime throughout my time at CMU.

Outside of the academic community, I am so fortunate to be supported by an endlessly supportive network of lifelong friends. In the past years, I have been drawn home to the midwest by the promise of morning coffee walks with Susan Miller, and I always look forward an evening of guacamole and beers with our family friends Lisa, Rodney, Elizabeth, and Kurt. Other key people in my support system have been my DC contingent of Jacob, Polly, Aaron, and Cristina; my Philadelphia contingent of Maryann, Matt, Zach, and Freddi; and my lifelong best friend, Rose Brown. Finally, to Grace, Anne, Donita, and Charlie: thank you for being my second family :).

Perhaps the most important thank you is to my family—my parents, Rob and Mona, and my brother Will. Thank you for supporting my loves of learning, logic puzzles, and books long before it was clear that they would turn into anything useful. You have always been there to support my education, listen to me complain, celebrate the wins, and give me great advice. I am so grateful to be able to count my family among my best friends!

Finally, I want to thank my partner, Chara, who also belongs among my collaborators, job market supporters, and best friends. In addition to supporting me through the ups and downs of the work in this thesis, thank you for being my biggest fan, best buddy, greek food chef, running partner, travel companion, and so much more. Every day with you and Terra is an adventure;
είμαι πολύ ενθουσιασμένη που ζω τη ζωή μαζί σας!

Contents

0.1 Introduction 1
0.2 Thesis Overview 5
0.3 A Final Comment on Motivation 5

I Sortition : Representation by Random Representatives 7

1 Background 8
1.1 Our Task: Sortition Subject to Quotas, Under Selection Bias 11
1.2 Overview of Chapters 13

2 A Selection Algorithm for Explicitly Reversing Selection Bias 22
2.1 Introduction 22
2.2 Model 26
2.3 Sampling Algorithm 27
2.4 Learning Participation Probabilities 29
2.5 Experiments 30
2.6 Discussion 33

3 A Framework of Sortition Algorithms 35
3.1 Introduction 35
3.2 Contribution I: Algorithmic Framework 37
3.3 Contribution II: Deployable Selection Algorithm 40
3.4 E ect of Adopting LexiMin over Legacy 41
3.5 Discussion 42
3.6 Additional Methods and Empirical Analysis 44

4 Fairness & Transparency 52
4.1 Introduction 52
4.2 Model 55
4.3 Theoretical Bounds on Marginal Discrepancy 57
4.4 Theoretical Bounds on Fairness Loss 60
4.5 Practical Algorithms for Computing Fair Uniform Lotteries 62
4.6 Discussion 64

5 Manipulation-Robustness 66
5.1 Introduction . . . 66
5.2 Model . . . 69
5.3 Leximin and Nash are Highly Manipulable . . . 72
5.4 $_{?}$-Norms Approach Optimal Manipulability as $? \to 1$ . . . 73
5.5 Manipulability of Real-World Instances . . . 76
5.6 Discussion . . . 78

6 Fairness, Manipulation-Robustness, & Transparency 80
6.1 Introduction . . . 80
6.2 Model . . . 85
6.3 Impossibilities for Maximin / Leximin, Nash, Minimax, and Linear $_{W}$ . . . 89
6.4 Analysis of Goldilocks . . . 92
6.5 Analysis of Transparent Goldilocks . . . 97
6.6 Empirical Evaluation . . . 98
6.7 Discussion . . . 101

7 Ongoing & Future Work 103
7.1 A Second Look At Representation/Randomness Trade-o s . . . 103
7.2 Holistically Designing the Participant Recruitment Process . . . 109

II Beyond Descriptive Representation 113

8 Background 114
8.1 Overview of Chapters . . . 116

9 A Voting Framework for Stakes-Based Representation 118
9.1 Introduction . . . 118
9.2 Model . . . 123
9.3 What if we have perfect stakes information? . . . 126
9.4 What if we have approximate stakes information? . . . 132
9.5 What if we have no stakes information? . . . 133
9.6 Discussion . . . 138

10 Ongoing and Future Work 141
10.1 A Stakes-Based Analysis of Quadratic Voting Ballots . . . 141
10.2 Implementing Stakes-Based Representation in Deliberative Town Halls . . . 144

III Deliberation, Public Spirit , and the Quality of Democratic Outcomes 147

11 Background 148

12 Public Spirit: Voting Beyond Self Interest 151
12.1 Introduction . . . 151
12.2 Model . . . 155
12.3 Distortion Bounds for Voting Rules . . . 158
12.4 PS-monotonicity . . . 164
12.5 Robustness of Distortion Bounds . . . 168
12.6 Discussion . . . 172

13 Extensions to Participatory Budgeting 174
13.1 Introduction . . . 174
13.2 Model . . . 179
13.3 Single-Winner Voting . . . 181
13.4 Rankings by Value . . . 183
13.5 Approval-Based Ballots . . . 187
13.6 A Thrifty Ordinal Ballot Gets Sublinear Distortion . . . 193
13.7 Discussion . . . 196

14 Ongoing and Future Work 197
14.1 Public Spirit in the Wild . . . 197
14.2 Beyond a Deterministic Highest-Welfare Alternative . . . 198

IV Discussion 200
14.3 Translating normative ideals into computational tools . . . 201
14.4 Part II: Grounding computational social choice models in political scientific research 201
14.5 Part III: Distilling where theoretical models have a comparative advantage and where they should give way to support empirical research and measurement . . . 203

V Appendices 225

A Chapter 2 Appendix 226
A.1 Notation Glossary . . . 226
A.2 Supplementary Material for section 2.3 . . . 227
A.3 Supplementary Material for section 2.4 . . . 237
A.4 Supplementary Material for section 2.5 . . . 239

B Chapter 3 Appendix 256
B.1 Illustration of Definitions with Examples . . . 258
B.2 Model . . . 260
B.3 Stratified Sampling . . . 261
B.4 Desiderata for Sortition in the Political Science Literature . . . 262
B.5 Related Work on Panel Selection . . . 265

B.6 Computational Hardness . . . 267
B.7 Small Optimal Portfolios Exist . . . 268
B.8 Algorithmic Framework . . . 268
B.9 Fairness Measures . . . 276
B.10 Description of Leximin . . . 282
B.11 Description of Legacy . . . 286
B.12 Description of Other Existing Algorithms . . . 287
B.13 Instances where Legacy is Unfair . . . 287
B.14 Comparing Legacy and Leximin on Intersectional Representation . . . 290
B.15 Axiomatic Analysis . . . 293

C Chapter 4 Appendix 297
C.1 Panel Selection Datasets . . . 298
C.2 Omitted Proofs and Additional Beyond-Worst-Case Upper Bounds from Section 4.3 298
C.3 Omitted Proofs from Section 4.4 . . . 312
C.4 Omitted Materials from Section 4.5 . . . 316

D Chapter 5 Appendix 325
D.1 Supplemental materials from Section 5.2 . . . 326
D.2 Supplemental materials from Section 5.3 . . . 329
D.3 Supplemental materials from Section 5.4 . . . 334
D.4 Supplemental materials from Section 5.5 . . . 336

E Chapter 6 Appendix 342
E.1 Supplemental Materials for Section 6.2 . . . 343
E.2 Supplemental Materials for Section 6.3 . . . 346
E.3 Supplemental Materials for Section 6.3 . . . 348
E.4 Supplemental Materials for Section 6.4 . . . 352
E.5 Supplemental Materials for Section 6.6 . . . 362

F Chapter 9 Appendix 371
F.1 Supplemental Materials from Section 9.1.2 . . . 372
F.2 Supplemental Materials from Section 9.3 . . . 374

G Chapter 12 Appendix 387
G.1 Supplemental Materials from Section 12.3 . . . 387
G.2 Supplemental Material for Section 12.4 . . . 402

H Chapter 13 Appendix 422
H.1 Rankings by Value for Money . . . 422
H.2 Threshold Approval Votes . . . 425
H.3 Proofs from Section 13.2 (Preliminaries) . . . 429
H.4 Proofs from Section 13.3 (Single Winner) . . . 432

H.5 Proofs from Section 13.4 (Rankings by Value) 437
H.6 Proofs from Section 13.5.1 (: -Approvals) 439
H.7 Proofs from Section 13.5.2 (Knapsack). 441

Listing of figures

1 Our hypothetical process for facilitating direct citizen participation in governance . . . 3

1.1 The two-stage panel selection process commonly used to select citizens' assemblies in practice. The dashed lines through the pool represent the fact that, the while the panel is designed to resemble the population, the pool may be very skewed demographically and ideologically due to selection bias. 10

1.2 Selection bias in the UK Climate Assembly across values of two features, education level and climate concern level. Percentages are omitted above panel bars because by design, they are essentially the same as those for the population. 12

2.1 Expected and realized numbers of panel seats our algorithm gives each feature-value pair in the Climate Assembly pool. 31

3.1 The steps of the algorithm optimizing the fairness measure. The left-hand panel shows the implementation of step (1): constructing a maximally fair output distribution over panels (denoted by white boxes), which is done by iteratively building an optimal portfolio of panels and computing the fairest distribution over that portfolio. The right-hand panel shows step (2): sampling the distribution to select a final panel. 38

3.2 Selection probabilities given by Legacy and LexiMin to the bottom 60% of pool members on six representative instances, where pool members are ordered in order of increasing probability given by the respective algorithms. Shaded boxes denote the range of pool members whose selection probability given by Legacy is lower than the minimum probability given by LexiMin. Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% confidence intervals (see methods section Statistics of the full version). For corresponding graphs for all other instances and up to the 100th percentile, see Figures 3.4 and 3.5 respectively in Section 3.6 42

3.3 HowLexiMin 's output was used to select a panel via a live uniform lottery. (a) First, the output distribution was transformed into a uniform distribution over 1,000 panels, numbered 000 999. (b) The three digits determining the nal panel were drawn from lottery machines, making each panel observably selected with equal probability. (c) The personalized interface (screen-captured with (b)) shows each pool member the number of panels out of 1,000 they are on, allowing them to verify their own and others' selection probabilities. Screenshots credit to for *. 43
3.4 Selection probabilities given by Legacy and LexiMin to the bottom 60% of pool members on the 4 instances that are not shown in Figure 3.2. Pool members are ordered across the x axis in order of increasing probability given by the respective algorithms. Shaded boxes denote the range of pool members with a selection probability given by Legacy that is lower than the minimum probability given by LexiMin . Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% con dence intervals (as described in Statistics in the Methods). Green dotted lines show the equalized probability 1=. 44
3.5 Selection probabilities given by Legacy and LexiMin on all ten instances. Pool members are ordered across the x axis in order of increasing probability given by the respective algorithms. In contrast to Figure 3.2 and Figure 3.4, this graph shows the full range of selection probabilities (up to the 100th percentile). Shaded boxes denote the range of pool members with a selection probability given by Legacy that is lower than the minimum probability given by LexiMin . Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% con dence intervals (as described in Statistics in the Methods). Green dotted lines show the equalized probability 1=. 45
3.6 Gini coe cient and geometric mean of probability allocations of both algorithms, for each instance. On every instance, Legacy has a lower Gini coe cient and a larger geometric mean. For computing the geometric mean, we slightly correct upward empirical selection probabilities of Legacy that are close to zero (as described in Statistics in the Methods). 46
3.7 For each instance, the share of pool members selected with lower probability by Legacy than the minimum selection probability of LexiMin is shown. This corresponds to the width of the shaded boxes in Figures 3.2, 3.4 and 3.5 46
3.8 Relationship between how overrepresented the features of an agent are and how likely they are to be chosen by the Legacy algorithm. The level of overrepresentation is quanti ed as the ratio product (as described in Individuals rarely selected by Legacy in the Methods); agents further to the right are more overrepresented. Across instances, pool members with high ratio product are consistently selected with very low probabilities. 47

3.9 For all intersections of two features on the instance sf(e), how far the expected number of group members selected by Legacy or LexiMin di ers from the proportional share in the population is shown. Although many intersectional groups are represented close to accurately, some groups are over- and underrepresented by more than 15 percentage points by either algorithm. Which groups get over- and underrepresented is highly correlated between both algorithms. Panel shares are computed for a pool of size 1,727, and population shares are based on a survey with 1,915 respondents after cleaning. 49

4.1 The quantization task takes as input a maximally fair panel distribution ? (implying marginals c), and outputs a 1• < -quantized panel distribution ? (implying marginals c). 57

4.2 < = 1000 Shaded regions extend from Maximin¹? º, the fairness of the optimal unconstrained distribution, down to the minimum fairness implied by the tightest theoretical upper bound in that instance (in all instances but obf section 4.3.2 is tightest). Each algorithm or bound's loss relative to Maximin¹? º is written above in the corresponding color. We show a representative run of Pipage, a randomized algorithm. 63

4.3 Instance = sf(a), < = 1000. Line plot shows the Leximin-optimal marginals (implied by panel distribution ?), along with marginals given by all algorithms sorted according to c . Note that each G coordinate then corresponds to an individual. The zoomed box shows the magnitude of marginal discrepancy around c . The surrounding shaded region shows the tightest theoretical bound on the marginal discrepancy, in this case from section 4.3.2, around the optimal marginals. We show a representative run of Pipage, a randomized algorithm. 64

5.1 Rounding-based algorithms Leximin , Nash , $_2$, and $_1$ versus each manipulation strategy in instances sf(a) and hd. 77

5.2 The impact of self-selection bias on the manipulability of Leximin , Nash , $_2$ and $_1$ by an agent playing OPT-1 strategy. 77

6.1 The solid, dashed lines represent maximum, minimum probabilities per algorithm, respectively. The shaded region lies between the optimal maximum probability and optimal minimum probability, establishing the region where no algorithm's extremal probabilities can exist. 100

6.2 Gini coe cient across algorithms and instances. Lower Gini Coe cient means greater fairness. 100

6.3 The maximum amount of probability any single MU manipulator can gain, for 1 and 2 pool copies. 101

6.4 Deviations from Goldilocks $_1$-optimal selection probability assignments by Pipage and ILP. The values for Pipage correspond to averages of minimum, maximum probability per run over 1000 runs. Error bars are plotted to indicate standard deviation, but they are so small that they are not visible. Gray boxes extend vertically from the minimum (resp. maximum) probability given by Goldilocks $_1$ to the theoretical bound , as given by Theorem 6.5.1 Optimal minimum, maximum probabilities per instance are shown for reference. 102

7.1 Caption. 107
7.2 Caption. 108

9.1 Constructions for reducing between the unit stakes assumption (existing model) and B-proportionality (our model). 131

C.1 $<$ = 1000 Shaded regions extend from NW^1? ^{o}, the fairness of the optimal unconstrained distribution, down to the minimum fairness implied by the tightest theoretical upper bound in that instance (in all instances but obf Section 4.3.2 is tightest). Each algorithm or bound's loss relative to NW^1? ^{o} is written above in the corresponding color. We show a representative run of Pipage, a randomized algorithm. 320
C.2 sf(b). 321
C.3 sf(c). 321
C.4 sf(d). 321
C.5 sf(e). 322
C.6 cca 322
C.7 hd 322
C.8 mass 323
C.9 nexus. 323
C.10 obf 323
C.11 ndem. 324

D.1 Figures for remaining instances from analysis in Figure 5.1 338
D.2 Figures for remaining instances from analysis in Figure 5.2(b) 340
D.3 Figures for remaining instances from analysis in Figure 5.2(c) 341

E.1 In instances 7-9, we use Maximin instead of Leximin to indicate the optimal minimum marginal probability because of computational costs due to the size of these instances. We additionally drop only 3 features instead of 4 because instance 9 only has 4 features. 368

G.1 contains all alternatives other than a^0 • 0 , cycled symmetrically over rankings, and all n are used for tie-breaking only. 398

List of Tables

3.1 List of instances used in our experiments. For the instances we study, panels were recruited by the following organisations: f(a-e) : Sortition Foundation; cca: Center for Climate Assemblies; hd: Healthy Democracy; mass: MASS LBP; nexus: Nexus; obf : of by for * (At the request of practitioners, topics, dates, and locations of the panels are not identified.) = is the pool size; is the panel size, and consequently, : • = is the mean selection probability. The # of features is j j, where each 5 2 has between 2 and 49 possible values (with the typical range being 2-5). 41

6.1 Approximations to the optimal minimum, maximum probabilities across algorithms and instances. 99

12.1 Bounds on the distortion of voting rules. Upper bounds hold for all lower bounds hold for all uniform $ = W1. As shorthand, we let W = ¹1 w • w Gray-text results are inherited from more general results. 158

13.1 Asymptotic (in < • W min) distortion bounds for rankings-by-value, comparing results for Single-winner (SW) and Participatory Budgeting (PB) ballots. The unit-sum results are derived in Benadè et al. [45] and are included for comparison. 178

13.2 Asymptotic (in < • W min) deterministic distortion bounds across ballot formats other than ranking-by-value. The colored rows indicate new ballots introduced in this paper. The unit-sum results are derived in Benadè et al. [45] and are included for comparison. 178

A.1 Climate Assembly UK features and values. 240

C.1 Instance parameters and resulting theoretical bounds 298
C.2 Run-times for Pipage, Beck-Fiala, and IP-NW 319

D.1 Overview of real-world instances. is a measure of the self-selection bias in the instance, as defined as Section 5.5.1 336
D.2 Minimum selection probability given to any agent by • 1 across instances. . . 341

E.1 We compare the performance of the two instance-specific gamma values described above against minimax, leximin, and goldilocks with a gamma value of 1. . . . 363

E.2 : , =, and $jW_{\#}j$ values across all 9 instances we analyze. 364
E.3 Times (seconds) of a representative run of all of the various objective-optimizing algorithms on Instances 1 and 8. 364

0.1 Introduction

Democracy as a form of governance, despite its long-debated and evolving imperfections, remains prized for its foundational principle: citizens should have the right to participate actively in governing their society. In democratic countries around the world, this philosophy tends to be implemented via *representative government*, in which citizens elect representatives to make decisions on their behalf and, in theory, in their interests. In recent decades, many have begun sounding the alarm about a "crisis in democracy", spurred by mounting evidence that many feel decreasingly represented by, trusting in, and able to influence their governments [74, 168, 217, 282, 283]. Others might argue that the more enduring crisis is that, over the past centuries, many of the world's democratic powers have consistently under-served the interests of entire subsets of their constituents, including indigenous populations, racial and ethnic minorities, people with disabilities, and the unhoused (e.g., [20, 171, 202, 228, 236, 262]). Either perspective leads to the same basic conclusion: there is a need to create more effective and inclusive access to governing power.

In this thesis, we focus on a proposed solution that is now gaining widespread momentum: augmenting representative democracy with processes that permit *direct* citizen participation in policymaking.[1] Our aim will be to contribute tools that support the principled design, implementation, and proliferation of these processes.

To start, in Section 0.1.1 we will design from the ground up a *hypothetical* process for facilitating direct participation in policymaking. Our goal here is not to build a perfect process, or even survey all possible design choices. Rather, we aim to illustrate some of the major challenges associated with involving citizens directly in policymaking, and to motivate the tools and ideas we will study in response.

0.1.1 Facilitating Direct Participation, Hypothetically

We start from what might seem to many like the biggest hurdle: Do everyday people have the expertise to make high-quality policy recommendations? If we imagine an average member of the public being asked to recommend a policy under everyday conditions, a reasonable answer might be "No": there is significant data supporting that people are susceptible to misinformation and propaganda [271], polarized [265], and politically disengaged [73, 108].

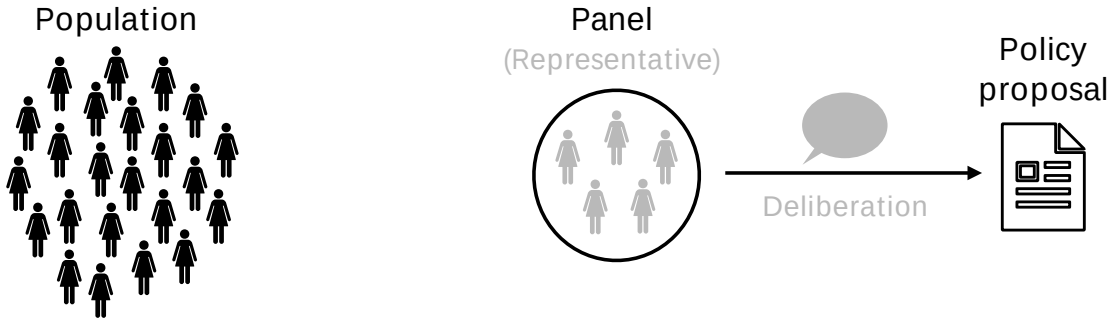
Fortunately, we are not bound to everyday conditions. This the proposal of *democratic deliberation*: to have people reason about politics through a structured discussion that is grounded in evidence and reasoning.[2] A growing community experts have high hopes for deliberation, seeing it as a way to facilitate high-quality political reasoning among everyday people even under conditions of polarization, misinformation, and political distrust [3, 100, 123, 222]. We add this tool to our hypothetical process so that prior to weighing in on what should be done, the citizens weighing in will engage in informed deliberation.

[1] We will use "citizen" to refer to *any* constituent of a democracy, implying nothing about legal citizenship status.

[2] Underlying this simple definition, there is significant research dedicated to the precise definition of deliberation in a democratic context (e.g., [200]).

For all its potential benefits, democratic deliberation has the downside that it is time- and resource-intensive: it often takes place over several days, in-person, and participants are necessarily compensated and reimbursed. As a result, not everyone in the population can participate. In practice, the solution is usually to choose a smaller group of citizens, which we will call a *Panel*, that ultimately participates in the discussion.

As soon as we need to select a smaller panel from within the population, we encounter a second major challenge: whose should be given a seat at the table? This challenge engages the concept of representation, on which there is a rich scholarship investigating who can, and who should, represent who? (e.g., [199, 229]). In many real-world decision-making contexts, it is popular to aim for *descriptive* representation of the population — that is, proportional representation of population subgroups — at least with respect to a predefined set of identities. We will adopt this goal for now, though we will revisit it later.



Finally, regardless of the notion of representation we want to ensure, we must decide: how should we select our representative panel? Here, we take inspiration from a centuries-old example of direct citizen participation: in ancient Athenian democracy, political representatives were selected directly from the population by lottery — a concept known as *sortition* [272]. Randomly selection may seem unprincipled compared to, e.g., choosing citizens who are especially qualified by some criteria. However, it is precisely this *absence of reason* for which many advocate sortition, arguing that a uniform lottery gives everyone a fair chance to participate; mitigates perverse and filtration mechanisms produced by elections; and as an added bonus, produces descriptive representation [71].

Unfortunately, adding sortition to our process will be a bit less straightforward than running

a uniform lottery: when participation is voluntary (as it typically is in models of direct democratic participation), those who are willing to partake tend to be highly skewed demographically and ideologically compared to the underlying population. Under this circumstance, known as selection bias, a uniform lottery will faithfully replicate this skew, failing to ensure the descriptive representation we desire. We therefore must implement a random selection procedure that strikes a desirable tradeoff between the representation we want, and the randomness that originally conferred sortition's defining benefits. We punt, for now, on how to do this, as it will be a central focus of this thesis.

Figure 1 depicts our resulting hypothetical process for facilitating direct citizen participation in governance.

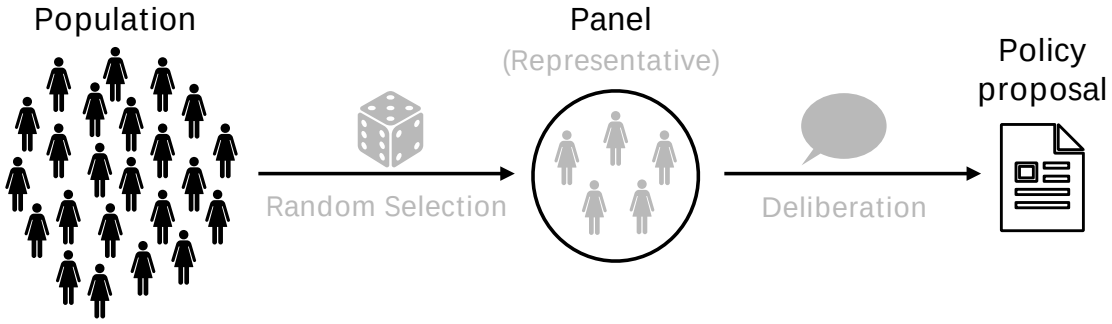


Figure 1: Our hypothetical process for facilitating direct citizen participation in governance.

0.1.2 From a Hypothetical Process to Real Participation Models

Although this process is hypothetical, in developing it we had to posit solutions for challenges that are somewhat fundamental to the task of facilitating direct participation in policymaking: to ensure participants can make well-reasoned decisions, processes often need to be somewhat intensive. When processes are intensive and the underlying population is large, it is often necessary to choose only a subset of the public to participate. When we must choose a subset of the public, we must make judgements about whose representation should be ensured, and how to choose participants in a way that best serves the goals of direct participation.

Given the fundamental nature of these challenges, it is perhaps unsurprising that the process in Figure 1 has many hallmarks in common with real participatory models being adopted around the world:

In participatory budgeting, a subset of citizens are appointed as budget delegates, and they convene to review evidence and collectively decide on how to divide a public budget over candidate public-interest projects. Although budget delegates are not typically selected randomly, many participatory budgeting resources reference the importance of representation among participants, especially from communities that are marginalized in standard politics [223, 244, 255]. In many cases, delegates engage in deliberation or other similar modes of learning about and

collaboratively weighing project alternatives [89, 192]. One guide even proposes deliberative participatory budgeting, suggesting the use of sortition to select a representative deliberative body [36]. Since 2013, participatory budgeting has been used to allocate billions of dollars (or other units of currency) [16]. In 2019 alone, there were upwards of 11,000 participatory budgeting events worldwide spanning 71 countries [89].

In deliberative town halls, citizens have moderated many-on-one discussions with their elected officials [211]. Deliberative town halls are built around goal of facilitating deliberative interaction between constituents and officials [212]. While they are sometimes open to all members of the community, in high-stakes applications with large populations, participants have been randomly selected with measures to ensure descriptive representation [178]. Deliberative town halls, often run through OSU's Connecting to Congress initiative [13], were recently used as part of Chile's national effort to amend their constitution [3, 9], and this participation model is now being scaled up through the online platform Prytaneum [15].

In independent redistricting commissions, a group of voters is convened to draw the boundaries of voting districts in a way that is hopefully more impartial than those produced through partisan gerrymandering. It is often paramount that the members of these commissions are representative on dimensions like political leaning; in some cases, the selection of participants involves randomizing; and many times, the process of collaboratively drawing new maps can involve substantial discussion [66, 245]. Independent redistricting commissions have been increasing in uptake over the past few years, having been used recently to draw congressional districts in Michigan, New York, Virginia, and Colorado [83].

Finally, the participation model that perhaps most closely resembles Figure 1 is the deliberative minipublic (DM). A DM proceeds much like our hypothetical model: a representative sample of the public is chosen by lottery to serve on a panel. Then, this panel convenes around a policy issue for several days, learning from experts, deliberating, and then finally weighing in on what should be done. DMs are actually an entire category of democratic paradigms, encompassing citizens' assemblies, citizens' panels, citizens' juries, deliberative polls, and more[1]. DMs constitute one of the most rapidly-growing models of citizen participation globally, with hundreds having been run around the world in just the past decade [4, 225]. DMs have been used at the national level in many countries including Mongolia [1], South Korea [7], Ireland [169], France [2, 65], and Germany [5]; they have even been used at the supranational scale (e.g., in the COP 26 Global Climate Assembly [11]). In the past few years, several instances have begun charting a path toward formal integration of deliberative minipublics into representative government. Citizens' assemblies are being integrated as permanent arms of governing bodies in major regions and cities, including Ostbelgian [215], Madrid [29], and Brussels [6], and there is now a law in Mongolia requiring deliberative polls before making certain kinds of constitutional amendments [1].

[1] The main distinction deliberative polls and many other DMs is how opinions are elicited post-deliberation: deliberative polls end with an anonymous poll of participants [125], while many other DMs end with participants collaboratively forming a policy proposal.

0.2 Thesis Overview

In this thesis, we will consider each of the three key components depicted in Figure 1.1. In Part I, we study procedures for randomly sampling a representative panel when there is selection bias; in Part II, we consider the implications of descriptive representation and explore an alternative; and in Part III, we examine potential impacts of deliberation on participants' political reasoning. Our work on these tools will be guided primarily by the application of deliberative minipublics, and sometimes citizens' assemblies in particular. However, given these tools' relevance to even just the participation models that already exist and the fact that new ones are emerging all the time [12, 14, 91, 289] we hope the ideas in this thesis can be applied to many other participatory processes as the landscape of citizen participation evolves.

Across Parts I - III, this thesis contains eight completed original papers [82, 43, 128, 130, 131, 134, 135], each in their own chapter.[1] Each of these Parts is laid out as follows:

> A background chapter outlines the problems we will solve, their context and motivation, modeling notions that apply across chapters, and an overview of the chapters themselves.
>
> Several research chapters each enclose the body of a single original paper.
>
> An ongoing and future work chapter discusses limitations of the existing work and some of the follow-up questions needed to address them.

Part IV is the Discussion, and Part V contains the appendices of all enclosed papers.

0.3 A Final Comment on Motivation

The primary motivation for the enclosed research, as discussed above, is advancing and supporting deliberative minipublics and other emerging models of direct citizen engagement. However, it is important to acknowledge that the political potential of these processes is yet unclear: multiple of these paradigms have so far shown mixed impacts [167, 220, 244, 267], and in many cases there remains the question of how to balance giving adequate authority to citizens' input while also maintaining robustness against bad actors [268, 284].

I tend to interpret these concerns as an indication that direct participation is a necessarily complex solution to a complex problem, and while it may have great potential to work, a lot of research is required to get there. The uncertainty of the present, however, requires us to address the question: what does this research contribute, in the event that direct citizen participation does not ultimately find a path to integration into representative government?

[1] It omits four additional published papers [83, 132, 133, 279], two of which relate to topics in this thesis but were deemed insufficiently relevant to democratic participation to be included. The first of these studies voting axiom satisfaction under a smoothed model of preferences, delineating classes of axioms and voting rules by whether semi-random noise is enough to escape axiomatic impossibilities [131]. The second covers a student-designed discussion-based course on diversity and inclusion in computer science [283], the discussion portion of which was influenced by and has influenced my thinking around democratic deliberation.

Fortunately, the ultimate implementation of these democratic innovations is not required for their study to be worthwhile. As we will illustrate throughout this thesis, the process of trying to understand these innovations o ers new angles from which to study concepts that are fundamental to how citizens engage with democracy in general. Moreover, direct participation models o er a uniquelygood settings in which to study these concepts: rst, they tend to bring people together in a location for long periods, allowing more in-depth inquiry of people's political knowledge, opinions, and patterns of political reasoning. Second, these processes are not yet entrenched in institutions, and the resulting uidity of their design makes them fertile ground for experimentation. Several ideas in this thesis will make use of or propose new ways for others to make use of these features for future research.

Part I

Sortition : Representation by Random Representatives

1 Background

1.0.1 Why Sortition (In Theory)?

There is a large body of political science scholarship arguing for the use of sortition over other methods for selecting political representatives. Here, we overview three of the arguments considered must centrally in this literature. We then distill these arguments into four technical ideals, numbered (i)-(iv), which we will pursue algorithmically throughout Part I. The arguments we discuss here were originally laid out in 1989 by Fredrik Engelstad [112] and have since been expanded upon by several scholars [38, 124, 259, ?]. Importantly, this body of literature conceives of sortition as a *uniform lottery* over the population, so for the purpose of interpreting the following arguments, we will adopt this conception for now.

The first argument in favor of sortition, articulated here by Carson and Martin, is that "those chosen [by sortition] are far more likely to be a typical cross section of the population, with the same sort of distribution according to sex, age, ethnicity, income, occupation, and so forth" [71]. Here, Carson and Martin refer to the fact that a uniform lottery will in expectation (and ex post, with high probability) choose a panel that is proportionally representative of all population cross-sections. Fishkin makes a similar argument, specifically with regards to how sortition can support the legitimacy of deliberative democracy: "We can only know [what the people would think] if we start the deliberations with a good microcosm, as representative as possible in both demographics and attitudes." [124] We distill these points into the following ideal:

(i) Descriptive Representation: The panel should be (at least nearly) proportionally representative of all population cross-sections.

A second popular argument for sortition has to do with its equal treatment of potential partici-

pants. Carson and Martin talk about the importance of equality of opportunity: [Sortition] gives everyone an equal chance of being chosen, whereas in elections, factors such as funding, appearance, speaking ability, threats, and promises play a big role [?]. [Peter Stone offers a] allocative justice as a normative argument in favor of this type of equality: "Allocative justice...is to treat public office as a type of good to which citizens might have various claims...Random selection is...appropriate...when all citizens have equal claims to that office" [259]. We summarize these these arguments as the following ideal:

(ii) Fairness: All people should have an equal opportunity to participate.

A third benefit of sortition is its ability to protected against subversion. As put by Oliver Dowlen, "[Sortition's] primary political potential is its ability to protect the public process of selection from subversion by those who might...use it for their own private or partisan ends" [98]. To distill technically well-defined ideals from this argument, we first ask: who might want to subvert the sortition process to their own ends? Peter Stone identifies two such parties: the organizers who select the panel, and the potential panel participants themselves [258].

According to Stone, sortition avoids potential subversion by the former group because "If [the agent who must select officials] selects randomly, then she must act on the basis of no reasons, and therefore cannot be influenced by corrupting or dominating interests even if she would like to be" [259]. However, we note a caveat to this argument: it relies on the fact that the public can confirm the selection was random. Otherwise, what is to stop the organizers from hand-picking a panel behind the scenes, and then claiming the selection was random? This can be avoided if there is transparency:

(iii) Transparency: Observers of the panel selection process should be able to confirm their probability of selection using only simple intuition about probability.

Fortunately, transparency is not hard to achieve when participants are selected by uniform lottery: one can just run a public lottery using simple physical randomness (e.g., drawing balls from bins).

The second entity who may want to subvert the sortition process are participants themselves: they might engage dishonestly in the selection process in order to stack the panel with people supporting their interests. For reasons that will soon become clear, we will be concerned with one particular method of dishonesty: misrepresenting one's attributes during the selection process (and/or convincing others to do so). Of course, because a uniform lottery treats everyone equally and independently, it ensures that this issue avoided. To capture this potential issue, we define the ideal of manipulation-robustness:

(iv) Manipulation-Robustness: Potential participants should not be able to affect their own or others' probabilities of selection by misrepresenting their identities in the selection process.

1.0.2 Sortition in Practice

Today, one of the most prominent (and growing!) use cases of sortition is to choose participants of deliberative minipublics. The practical reality of this use case, however, di ers in a crucial way from the ideal. In idealized sortition, it is assumed that any member of the population selected by uniform lottery will participate. In contrast, organizers of modern deliberative minipublics cannot compel participants to participate, and must rely on people to opt in. The typical rate of opting in among the general population is around 2-5% [28], and as we will illustrate with data shortly, those who agree to participate are usually highly demographically skewed. A simple uniform lottery would replicate this skew, producing a panel that is far from descriptively representative. Seeing this as an important issue for public and normative legitimacy, practitioners at least of citizens' assemblies, which will be the primary application of Part I perform random selection via the following two-stage process (depicted in Figure 1.1).

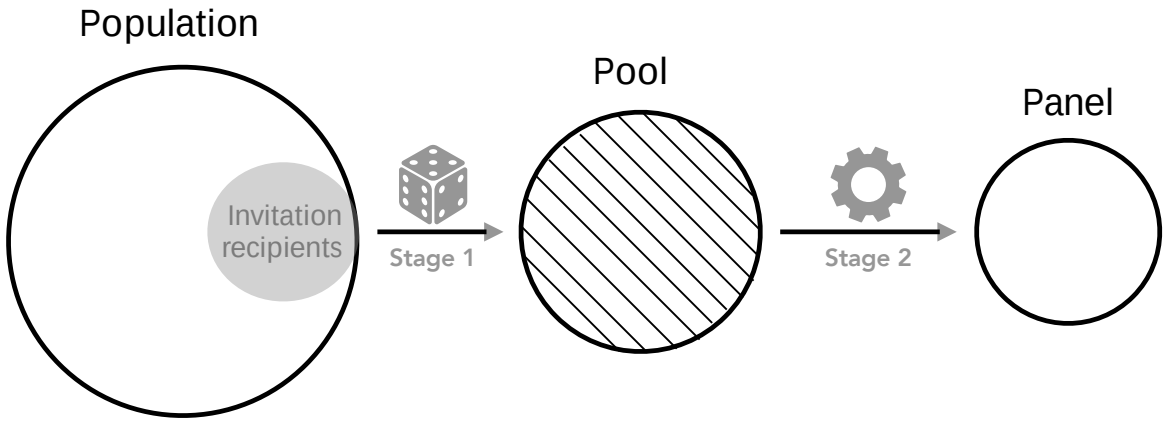


Figure 1.1: The two-stage panel selection process commonly used to select citizens' assemblies in practice. The dashed lines through the pool represent the fact that, the while the panel is designed to resemble the population, the pool may be very skewed demographically and ideologically due to selection bias.

Stage 1 (Uniform Lottery Invitations). Practitioners invite participants uniformly randomly from the underlying population (usually via either letters or phone calls). Those who receive an invitation and respond a rmatively form the Pool of volunteers. Upon volunteering (i.e., joining the pool), all pool members must ll out a survey about their demographic and ideological attributes, which will be used in Stage 2 to ensure that the panel is representative.

Stage 2 (Panel Selection). The Panel is selected from within the pool. In practice, this panel must satisfy two main requirements deterministically:

- Panel size. The panel must contain exactly members, where is chosen by the panel organizers. This requirement arises from budgetary constraints, as practitioners must cover some per-participant cost.
- Representative quotas. The panel must satisfy upper and lower quotas on a set of pre-

de ned attributes, again chosen by practitioners. More precisely, these quotas are structured in the following way:

Let be a set of attribute categories, which we will call features, for example, in a UK-wide assembly on climate change in 2020 (our running example for the remainder of this section), included the features education level, gender, age, climate concern level, race/ethnicity, geography 1, and geography 2 [2232]. For each feature 5 2 , practitioners de ne a set of mutually exclusive and exhaustive values, which we call feature-values. For example, in the UK climate assembly, climate concern level contained the values Not concerned, not very concerned, fairly concerned, very concerned.

Upper and lower quotas are imposed at the feature-value level, denoted respectively as 5 • E 5 • E for feature-value 5 • E. Take the example of the feature-value gender, female. Lower and upper quotas of 64=34A • 5 4<0; = 8 and 64=34A • 5 4<0; = 12 would mean that the panel would be required to contain between 8 and 12 women. Typically, upper and lower quotas are imposed on all feature-values in 52 +5, and enforce that each feature-value-de ned group receives a number of panel seats near-proportional to their share of the population. For example, if women comprise 49% of the population, then quotas on a panel of size : = 100 might require between 5 • E = 48 and 5 • E = 50 women.[1]

Let # denote the pool, let 5 1 8 º 2 +5 denote a pool member's value for feature 5, and let + := 52 +5 denote the set of feature-values on which quotas are imposed. will be the indicator function. Then, an instance of the panel problem is de ned by four quantities: #, • 1 5 • E 2 + º • 1 5 • E 2 + º • :. In any given instance, the set of valid panels is

$$: \quad \# \;\hat{}\; j \quad j = : \;\hat{}\; \tilde{O}_{8 2} \; \mathrm{I}\,1\,5\,1\,8^{\circ} = E^{\circ} 2 \;\; \ll_{5 \bullet E} \bullet D_{5 \bullet E} \text{ for all } 5 \bullet E 2 \; + \;\; ”$$

An instance of the panel selection problem is solved by a selection algorithm, which is any procedure (mapping) that intakes an instance of the panel selection task and outputs a valid panel, provided at least one valid panel exists.

1.1 Our Task: Sortition Subject to Quotas, Under Selection Bias

Notably absent from this discussion so far has been any discussion of randomness, the hallmark of sortition. As the previous section suggests, in our version of sortition, we must randomize within the quotas. This restriction already precludes one aspect of idealized sortition: we can no longer independently sample the panel members. The question is then whether we can still retain the

[1] Occasionally, practitioners impose quotas on combinations of attributes as well, though of course the extent to which this is possible is limited by the combinatorial explosion of attribute combinations relative to the small panel. To implement quotas on arbitrary combinations of attributes within this model, one would just make each combination they care about a feature-value. For example, if you wanted to enforce representation on all intersections of age and height, you would de ne the feature 5 = age height with values + age height = f young & short, young & tall, old & short, old & tall g.

other, perhaps more de ning property of idealized sortition: giving peopleequal probability of selection. Unfortunately, in the practical case, the answer is No. This is because there is selection bias in who opts into the pool from the population, meaning that the pool is far from representative of the population on the dimensions on which we impose quotas. To illustrate this point, Figure 1.2 shows the compositions of the Population, Pool, and Panel in the UK Climate Assembly for two features: ℓ_1 = education level, whose values were $\{0/1, 2/3, 4+\}$, where higher levels correspond to more education, and ℓ_2 = climate concern level, whose values were {Not concerned, not very concerned, fairly concerned, very concerned}. In the gure below, we group the two lowest levels of climate concern.

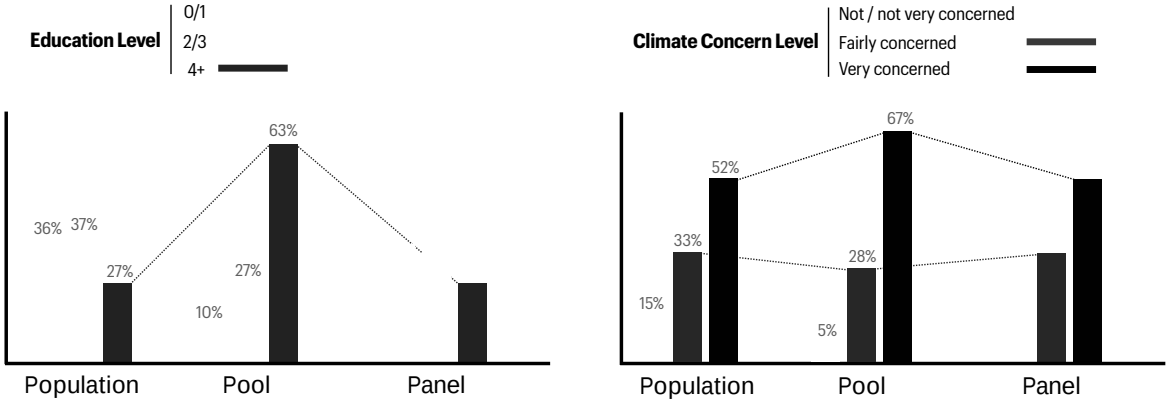


Figure 1.2: Selection bias in the UK Climate Assembly across values of two features, education level and climate concern level. Percentages are omitted above Panel bars because by design, they are essentially the same as those for the Population.

To see the selection bias in Figure 1.2, we must compare the composition of the population versus the pool. In making this comparison, note that because invitations to participate were sent out to members of the population uniformly at random, the pool is e ectively a uniform random sample of people who would opt if invited. Examining the features of education and climate concern level as in Figure 1.2, we see that less educated groups and those who are less concerned about climate change are dramatically underrepresented among those who opted into the pool.

To understand the implications of this selection bias for the panel selection problem, we now compare the pool versus the panel. Comparing pool members with education level 0/1 versus those with level 4+, notice that there are way more panel seats per person reserved for the former group: there are about 1/12 as many pool members in the former group, and they are entitled to about 1.5x as many panel seats. They key consequence is that as a result, those with education level 0/1 must be selected for the panel with higher probability, on average, than those with education level 4+. This example illustrates a very general impossibility that grounds our subsequent work:

Key Impossibility: When we must satisfy representative quotas under the condition of selection bias, we cannot select all pool members with equal probability.

Things now seem a bit bleak: in practical sortition, we cannot give people equal probabilities of

selection, and thus cannot achieve the defining property of a uniform lottery. Our last hope is that despite these impossibilities, we can still achieve ideals (i)-(iv) i.e., the reasons for using sortition at all to at least a reasonable degree. This will be the goal of Part I, whose chapters we now overview. For most of this part, we will consider the first ideal (I), Descriptive Representation, to be automatically satisfied by the requirement of quotas. We examine the extent to which this an oversimplification in our ongoing and future work (Chapter 7).

Remark 1.1.1 (The Case of Mandatory Participation). One might argue that these algorithms designed around the challenge of selection bias would become obsolete if participation in direct democratic processes became required. The assumption implicit in this argument is that requiring participation would eliminate selection bias and restore the viability of a uniform lottery; we now examine this assumption via the case study of U.S. jury selection, in which participation is legally required upon being summoned, except if the person meets one of the many criteria for excused absence, commonly including being above a certain age, being a student, or having a dependent child (precise regulations vary by jurisdiction, e.g., [138, 235]). These exemptions are necessary to avoid placing undue burdens on citizens, and similar exemptions would undoubtedly be needed if mandating participation in deliberative minipublics were mandated. Unsurprisingly, these exemptions along other issues with people simply failing to appear result in documented issues with certain populations being underrepresented among those who report for jury duty [270], posing a concern for the equity of our justice system's verdicts [26]. Based on this case study, it seems unlikely that a participation mandate or any other intervention could completely eliminate selection bias in deliberative minipublics, and moreover, failing to account for what bias remains would likely lead to systematic exclusion of certain groups. Our algorithms can ensure representation in the presence of selection bias, while retaining key properties of uniform lotteries to the greatest extent possible. Of course, as we will discuss in this thesis, efforts to decrease selection bias are extremely important; our algorithms align with this goal, improving in performance as selection bias decreases in severity.

1.2 Overview of Chapters

We begin by posing the question of whether our Key Impossibility above is really so fundamentally problematic. That impossibility says that we cannot give all pool members an equal chance of selection; however, it seems like to imitate idealized sortition, we should care about giving all population members an equal chance of participating, rather than all pool members. We now consider the implications of this distinction for the most directly-related ideal (ii), Fairness. We make the distinction explicit by delineating two possible interpretations of fairness:

Fairness of outcome. All population members should have the same probability of participating.

Fairness of opportunity. All population members should have the same probability of receiving the opportunity to participate.

To unpack these notions, we first define the central object of Part I: a pool member's selection probability is their probability of being chosen for the panel. These two notions of fairness, then,

differ fundamentally in what they mean for the ideal selection probabilities. The first notion amounts to giving all population members an equal probability of ending up on the panel, which would be achieved by choosing each pool member with a probability inversely proportional to their chance of opting into the pool. The second notion amounts giving equal selection probability to all pool members, because the first stage is a uniform lottery, which gives everyone the same probability of being invited; then, to maintain equal opportunity to participate, we need to give all those who opted in an equal chance of being chosen for the panel in the second stage.

For reasons that will become apparent throughout Part I, we will focus primarily on the second notion of Fairness. However, the first notion *a priori* perhaps seems more theoretically natural, so in Chapter 2, we explore what it would take to design a selection algorithm that achieves it.

> Chapter 2: A Selection Algorithm for Explicitly Reversing Self-Selection Bias
>
> Based on "Neutralizing Self-Selection Bias in Sampling for Sortition" [126].
>
> In this chapter, we design a selection algorithm that achieves the first notion (ii) of Fairness via the intuition above: if each population member i opts into the pool (conditional on being invited) with probability π_i, we want to choose them with probability proportional to $1/\pi_i$ in the second stage. This will give all population members an equal probability of ending up on the panel end-to-end, regardless of their opt-in probability.
>
> The key technical challenges here are (1) knowing individuals' unobservable opt-in probabilities π_i, so that our algorithm can set probabilities correctly; (2) determining conditions under which we can set all pool members' selection probabilities proportional to $1/\pi_i$ such that they remain in $[0, 1]$; and finally, (3) designing a procedure for turning these probabilities into a final panel that a. preserves these probabilities and b. is deterministically guaranteed to satisfy descriptively representative quotas.
>
> Our approach to challenge (1) begins with the observation that, by comparing the composition of the pool to that of the population, one can make inferences about which types of people tend to participate. For example, if the population is 50% women but the pool is only 20% women, you might infer that being a woman decreases one's chance of participating. We formalize this intuition by using these data to fit a model predicting the π_is via maximum likelihood estimation. Addressing challenge (2) requires an assumption that no participation probability is too low, and our guarantees depend on the extent to which this assumption holds. Finally, we address challenge (3) by designing a dependent rounding procedure based on a celebrated discrepancy theorem by Beck and Fiala [40]. This dependent rounding procedure preserves the $1/\pi_i$-proportional selection probabilities, while also guaranteeing that the final panel does not deviate from perfect proportional representation by more than 1.

While the selection algorithm presented in Chapter 2 is theoretically appealing, we argue that a main takeaway of this chapter is that trying to explicitly reverse self-selection bias in practice is fraught with risks. First, the levels of Descriptive Representation and Fairness achieved by this method hinge on one's ability to accurately estimate pool members' individual opt-in probabilities. In practice, this estimation must be done by comparing the pool versus the population

composition, which makes estimations fundamentally susceptible to errors for several reasons: (1) the population data serves as the *control group*, but it truly contains many people who might have opted in if invited, so we have no data on exclusively people who decline to participate; (2) the available population data may be limited, especially at the level of combinations of attributes,[1] and (3) practitioners typically know very few feature-values about pool members, so predictions would likely be based on only some of the features that are crucial to the decision to opt-in.

Even if we could perfectly estimate pool members' opt-in probabilities, our rounding-based approach does not give strong enough guarantees on ex-post representation: in contrast to the custom quotas practitioners prefer to use, this rounding method may relax representation in ways by which in practice typically ranges from 4-8; for groups that are small (which is not uncommon in real instances), this relaxation can dramatically weaken or even eliminate any guarantee of their inclusion. While it might be possible to drop this bound to based on a conjecture by Beck and Fiala [40], it seems unlikely that it could be improved further due to a lower bound previously proven by Olson and Spencer [221]. In principle, one could altogether abandon the approach of first determining selection probabilities and then rounding them, though there is not a clear alternative approach that would reverse the selection bias while meaningfully circumventing this issue.

For the remainder of Part I, we will for now abandon the goal of *Fairness of outcome*, and instead pursue *Fairness of opportunity*, which we will henceforth refer to as *Fairness*. Our goal will now be to design selection algorithms that achieve the ideals outlined above (ii) *Fairness*, (iii) *Transparency*, and (iv) *Manipulation Robustness*, to the greatest degree possible subject to custom practitioner-defined quotas (standing in for (i) *Descriptive Representation*).

Our approach begins from the observation that these ideals were originally conferred by giving people *equal* selection probabilities. Given that this is impossible by our Key Impossibility, we pursue the next best goal: to make pool members' selection probabilities as equal as possible, subject to the quotas. Solving this technical task is the purpose of Chapter 3, whose primary contribution is an algorithmic framework that will serve as the basis of all later chapters.

> Chapter 3. An Algorithmic Framework for Maximal Equality.
>
> Based on *Fair Algorithms for Selecting Citizens' Assemblies* [136].
>
> Before designing any algorithms of its own, this paper's first contribution was to evaluate the selection algorithms that were being used in practice at the time. These pre-existing algorithms were greedy heuristics whose main goal was to find any valid panel, randomizing in ad-hoc fashion where possible along the way (this was a reasonable first goal, given that finding any single valid panel is NP-hard [130]). In our empirical evaluation, we found that a popular such algorithm whose probabilistic properties had never been characterized was prone to giving a large portion of the pool near-zero chance of selection, posing a significant

[1] For example, the European Social Survey data (the public population-level data corresponding to the UK climate assembly) is missing 335 of the 762 unique feature-value combinations that appear in the pool.

issue from the perspective of Fairness.

In response, our goal was to design a selection algorithm that made pool members' selection probabilities *maximally equal*, subject to practitioner-defined quotas. A natural question, of course, is how one should measure "maximally equal". The algorithms we present in this paper will permit the use of *any* convex function that intakes a vector of selection probabilities and outputs a real-numbered measurement of their level of equality a class which encompasses well-known notions like *Maximin* (no one receives too little selection probability), Nash Welfare (the geometric product of selection probabilities), or the Gini Coefficient (a popular measure of inequality).

The algorithmic framework we present, at a high level, first computes a distribution over valid panels, which we call a *panel distribution*, and then samples the final panel from that distribution. Note that by taking this approach, we ensure that the resulting panel is valid. Then, since any panel distribution implies selection probabilities[a], our task boils down to finding an optimal panel distribution i.e., one that makes pool members' selection probabilities maximally equal.

The major technical challenge in computing an optimal panel distribution is that *a priori*, any optimal distribution might need to place selection probability on all valid panels, of which in practice there are astronomically many. Fortunately, we show that due to our equality objective being a function of just the selection probabilities (a very low-dimensional object relative to the space of all valid panels), by Caratheodory's theorem there must exist an optimal solution over only very few panels. Unfortunately, this theorem does not tell us how to find such a small-support distribution.

Our main contribution is an algorithmic framework for finding an optimal panel distribution over a practicable (but not theoretically bounded, in the worst case) number of panels. To understand this algorithm, it is useful to envision the primal program: we have one variable per every possible valid panel, corresponding to the probability we will place on that panel. Our goal is to find values of these (astronomically many) variables that optimize our convex objective function, which captures the equality of the selection probabilities implied by our panel distribution. The framework solves this massive program by solving its dual, using column generation to iteratively add panels to the support (corresponding to adding constraints to the dual) until a stopping condition is reached. As we prove, this stopping condition is sufficient for the existence of a KKT-condition-satisfying solution to the full dual program, corresponding to the current randomization being optimal among all randomizations over all valid panels

[a] Given a panel distribution, the selection probability of any pool member is simply the probability of drawing a panel containing them.

With this algorithmic framework in hand, the question is then, *How* can we best use it to serve our ideals of Fairness, Manipulation-Robustness, and Transparency? This is the main question that will occupy us for the rest of Part I.

In the paper from Chapter 3, we instantiated our framework with an equality objective chosen to promote Fairness. To translate this ideal into a mathematical objective, we began from the allocative justice perspective [259] that the chance to hold public office is a good to which people are entitled to their fair share. Accordingly, we defined maximal fairness according to the objective Maximin, which when optimized maximizes the minimum selection probability received by any pool member, corresponding to ensuring that no one receives too much less than their share. Our implementation (of a slight refinement of this objective called Leximin) is publicly available on Panelot.org and on the widely-used selection tool of the Sortition Foundation, a major organizer of citizens' assemblies [163]. Since 2020, this algorithm has been used to select high-profile assemblies, including the Global Climate Assembly in 2020, Michigan's statewide assembly on COVID-19 in 2021, Scotland's national climate assembly in 2022, Germany's National Assembly on Nutrition in 2023, and Ostbelgien's permanent assembly in 2023.

An additional contribution of Chapter 3 is a proposal for an algorithmic add-on targeting the goal of Transparency. In its standard implementation, our algorithmic framework from Chapter 3 is not very transparent; it computes a complicated distribution over a few thousand panels, and then samples this distribution. In principle, one could publish this distribution, and then sample it by publicly running code that generates a random number. Needless to say (but we will say it anyway), this would be virtually impossible for the average person to understand. To remediate this, we proposed the following approach:

1. Round the optimal panel distribution produced by our algorithmic framework so that all probabilities in the panel distribution are multiples of 1/1000 (or some other integer denominator of the user's choice), while ensuring it remains a valid distribution.
2. Number these probability blocs from 000...999. Note that each bloc corresponds to a panel (with multiple blocs potentially corresponding to the same panel). Now, one has a list of 1000 panels (with duplicates).
3. Uniformly sample this list of panels to choose the final panel. By construction, this corresponds to sampling the rounded distribution from step 1.

This method is transparent in an important sense: if the list of panels and their (anonymized) members can be made public which in practice, it has been [130] this uniform lottery over panels allows the public to observe their own and other pool members' selection probabilities by the same simple reasoning required to understand that by buying more lottery tickets, you increase your chance of winning the lottery.[1]

Although Chapter 3 proposed this method, it left out an important open question: Does the rounding of the optimal panel distribution significantly compromise its optimality? In Chapter 4, we prove that it does not.

[1] If someone sees that they are on 20 out of 1000 panels, they immediately see that their chance of selection is 20/1000 = 2%.

Chapter 4. Transparency.

Based on "Fair Sortition Made Transparent" [131].

In this paper, we propose several algorithms for rounding the panel distribution output by the algorithm in Chapter 3. For each rounding algorithm, we upper-bound the maximum extent that it can change any individual selection probability in order to quantize any panel distribution. Though we extend these bounds to bound the optimality loss for only two equality objectives, Maximin and Nash Welfare, our bounds on changes to individual selection probabilities are general enough to bound optimality loss for most reasonable equality objectives. Finally, we empirically evaluate these rounding algorithms in real citizens' assembly datasets. This analysis identifies a simple and fast rounding procedure that almost exactly retains the optimality of the original panel distribution across instances. We conclude that transparency comes at essentially no cost to maximal equality in practice, and at a practically bounded cost in theory.

Since its proposal, this rounding method has been used in conjunction with our algorithmic framework to select multiple citizens' assemblies, including aforementioned assemblies in Michigan and Germany.

We now turn our attention to the ideal of "Manipulation Robustness". In an unfortunate turn of events, our study of this ideal will reveal some bad news about the equality objectives we have studied so far.

Chapter 5. Manipulation Robustness

Based on "Manipulation-Robust Citizens' Assembly Selection" [135].

In Chapter 3, translating the conceptual ideal of "Fairness" to a mathematical equality objective to plug into our algorithmic framework was relatively straightforward. However, the analogous transformation for "Manipulation Robustness" is less straightforward: to understand what equality objective minimizes incentives for pool members to misreport their features, we first need to define a game theoretic model. Instead of defining just one such model, we define three, each corresponding to different potential motive for misreporting one's features: to increase one's own selection probability, decrease someone else's, or steal a seat; that is, you might impersonate another group so that if you are selected, you will have taken a panel seat reserved for that group.

Our first finding is quite troubling: we show that Maximin and Nash Welfare, the two objectives we've studied for their prioritization of "Fairness", are arbitrarily manipulable; that is, they permit a pool member, by misreporting their features, to gain selection probability 1. This is a worst-case result, but we show that this finding also holds in real datasets, even for very rudimentary manipulations requiring no knowledge of the algorithm.

The most striking aspect of this impossibility is that it persists even as the pool grows arbitrarily large relative to the panel. To see why this is surprising, let n be the pool size; as n grows relative to k, the average selection probability k/n should go down. Then, it seems that there

is less probability available per person, so should everyone's selection probability decrease, both pre- and post-manipulation, thereby decreasing the amount of probability that can be gained by manipulating? The key to understanding this impossibility is the intuition people can misreport combinations of features that do not exist in the pool, and which can make them "unicorns" to the objectives Maximin and Nash Welfare[b], both which almost exclusively care about making sure the lowest probability is not too low. Here, someone is a "unicorn" when giving them more selection probability makes it feasible to raise the lowest selection probabilities. When someone misreports a combination of features that makes them a unicorn, both Maximin and Nash Welfare may pile probability onto them to the greatest extent possible, bringing their selection probability up to 1.

Based on the intuition that high selection probabilities are a problem for manipulation robustness, it should not come as a surprise that the equality objective Minimax, which minimizes the maximum selection probability, provably minimizes manipulation incentives[c]. We show that the manipulation incentives induced by Minimax decline at a rate of $\Theta(1/n)$ as n grows relative to k, which is the optimal possible rate for any selection algorithm.

[a] Growing the pool just means sending out more letters in the first stage, so the composition of the pool (and thus the level of selection bias) remains relatively constant.

[b] Maximin does this by definition. Nash Welfare, by being the product of selection probabilities, is relatively unaffected by probabilities near 1, but is extremely affected by even a single probability near 0.

[c] In the paper, we do not strictly study the objective Minimax, but rather the p norms of selection probabilities, which effectively converge to Minimax as $p \to \infty$ (a regime we characterize). In subsequent work, we will consider Minimax in place of the ℓ_1 norm.

Remark 1.2.1 (Revisiting Our Approach From Chapter 2). With Chapters 4 and 5 under our belt, we can now identify another reason why the approach taken in Chapter 2 to make pool members' probabilities inversely proportional to their chance of opting in is practically dicey. The key reason is that this method will make pool members' probabilities widely different far more different, potentially, than algorithms maximizing their equality. This is a problem for Transparency: if we make selection probabilities visible to the public and they are extremely disparate, this may create the sentiment that the process is very unfair (and the estimates upon which these probabilities are based are hard to soundly and transparently justify). Second, these disparate problems are an even bigger problem for Manipulation Robustness: the higher the probability someone can receive based on their features, the stronger the incentives for manipulation.

We have now studied Fairness, Transparency, and Manipulation Robustness; in our final research chapter in Part I, we will study the extent to which we can achieve these ideals simultaneously. If we can achieve this, we will complete our original goal: to design a sortition algorithm that achieves ideals (i)-(iv) to the greatest extent possible given the non-ideal conditions of real-world sortition.

Chapter 6. Fairness, Manipulation-Robustness, and Transparency

Based on "Fair, Manipulation-Robust, and Transparent Sortition" [31]

Setting aside Transparency for a moment, the previous chapters reveal a potential tradeo

between Fairness and Manipulation Robustness: low probabilities are a problem for the former (essentially by de nition), and high probabilities are a problem for the latter. No equality objective we have studied achieves anywhere close to both ideals: Maximin/Leximin and Nash Welfare control only low probabilities, making them very fair but arbitrarily manipulable; by controlling only high probabilities, Minimax is optimally manipulation-robust but arbitrarily unfair, giving many pool members zero chance of selection.

In this chapter, we propose a new equality objective, called Goldilocks, that aims to achieve these ideals simultaneously by controlling both high and low selection probabilities, and which can be optimized via the framework in Chapter 3. The fundamental challenge in controlling high and low probabilities simultaneously is that manipulating coalitions, by misreporting, can a ect the quality of available lotteries by reporting features between which there must be fundamental gaps between the maximum and minimum probability. Thus, in order to analyze our algorithm, we must rst characterize the extent to manipulation can damage the space of feasible solutions, and then we can analyze the ability of our algorithm to recover good solutions despite this.

After circumventing these challenges, we give theoretical bounds (many of them tight) on the extent to which Goldilocks achieves Fairness and Manipulation Robustness, nding that in a very important sense, Goldilocks recovers among the best available solutions in a given instance. We then extend these theoretical bounds to the case where the output of Goldilocks is transformed to achieve a third goal, Transparency. Our empirical analysis of Goldilocks in real data is even more promising: we nd that this objective achieves nearly instance-optimal minimum and maximum selection probabilities simultaneously in most real instances an outcome not even guaranteed to be possible for any algorithm.

Although there is always room for future work here, in many respects, Goldilocks closes the question of whether we can simultaneously achieve three key ideals of sortition: Fairness, Manipulation Robustness, and Transparency, and contributes a practicable algorithm for doing so. Now that we better understand what is possible in lottery design regarding ideals (ii)–(iv), in ongoing and future work, we can circle back to the very rst ideal (i) Descriptive Representation and consider the implications of how quotas are used to enforce it.

Chapter 7: Ongoing and Future Work. In our rst ongoing project, we are trying to help practitioners set more principled quotas. Currently, practitioners hand-design quotas without insight into how small changes in the quotas may change the lottery changes that are di cult to predict, given the combinatorial relationship between the quotas and the space of possible lotteries. To close this gap, we are designing a deployable tool for ne-tuning quotas that are optimized to permit better lotteries. Empirically, our preliminary results show that small changes in quotas can permit signi cantly more uniform lotteries.

Sometimes our quota-tuning method relaxes a quota, e.g., reserving 22 instead of 24 seats for women, in favor of a far more uniform lottery. Often, we think of loosening quotas as de nitively harming representation; however, expanding our conception of representation beyond

just the few features protected by quotas, it is not actually clear: a more uniform lottery can also support representation by ensuring that no group unprotected by quotas is systematically given low selection probability, plus it can decrease incentives for representation-corrupting manipulation. We are investigating this ambiguity by studying how different trade-offs between quota tightness and lottery uniformity affect representation of unprotected groups and the diversity of the resulting panel.

Additional ongoing and future work aims to alleviate other bottlenecks in the selection process, including the need to select alternate panel members to handle dropout.

2

A Selection Algorithm for Explicitly Reversing Selection Bias

Neutralizing Self-Selection Bias in Sampling for Sortition [126].
Bailey Flanigan, Paul Gölz, Anupam Gupta, and Ariel D. Procaccia.
NeurIPS 2020.

2.1 Introduction

What if political decisions were made not by elected politicians but by a randomly selected panel of citizens? This is the core idea behind sortition, a political system originating in the Athenian democracy of the 5th century BCE [72]. A sortition panel is a randomly selected set of individuals who are appointed to make a decision on behalf of population from which they were drawn. Ideally, sortition panels are selected via uniform sampling without replacement — that is, if a panel of size k is selected from a population of size n, then each member of the population has a $\frac{k}{n}$ probability of being selected. This system offers appealing fairness properties for both individuals and subgroups of the population: First, each individual knows that she has the same probability of being selected as anyone else, which assures her an equal say in decision making. The resulting panel is also, in expectation, proportionally representative to all groups in the population: if a group comprises 33% of the population, they will in expectation comprise 33% of the panel as well. In fact, if k is large enough, concentration of measure makes it likely that even a group's post-share of the panel will be close to 33%. Both properties stand in contrast to the status quo of electoral democracy, in which the equal influence of individuals and the fair participation of minority groups are often questioned.

Due to the evident fairness properties of selecting decision makers randomly, sortition has seen

a recent surge in popularity around the world. Over the past year, we have spoken with several nonpro t organizations whose role it is to sample and facilitate sortition panels [75]. One of these nonpro ts, the Sortition Foundation[1], has organized more than 20 panels in about the past year. Recent high-pro le examples of sortition include the Irish Citizens' Assembly[2], which led to Ireland's legalization of abortion in 2018, and the founding of the rst permanent sortition chamber of government[3], which occurred in a regional parliament in the German-speaking community of Belgium in 2019.

The fairness properties of sortition are often presented as we have described them in the setting where panels are selected from the whole population via uniform sampling without replacement. As we have learned from practitioners, however, this sampling approach is not applicable in practice due to limited participation: typically, only between 2 and 5% of citizens are willing to participate in the panel when contacted. Moreover, those who do participate exhibit self-selection bias, i.e., they are not representative of the population, but rather skew toward certain groups with certain features.

To address these issues, sortition practitioners introduce additional steps into the sampling process. Initially, they send a large number of invitation letters to a random subset of the population. If the recipients are willing to participate in a panel, they can opt into a pool of volunteers. Ultimately, the panel of size is sampled from the pool. Naturally, the pool is unlikely to be representative of the population, which means that uniformly sampling from the pool would yield panels whose demographic composition is unrepresentative of that of the population. To prevent grossly unrepresentative panels, many practitioners impose quotas on groups based on orthogonal demographic features such as gender, age, or residence inside the country. These quotas ensure that the ex-post number of panel members belonging to such a group lies within a narrow interval around the proportional share. Since it is hard to construct panels satisfying a set of quotas, practitioners typically sample using greedy heuristics. While these heuristics tend to be successful at nding valid panels, the probability with which an individual is selected is not controlled in a principled way.

Since individual selection probabilities are not deliberately chosen, the current panel selection procedure gives up most of the fairness guarantees associated with sortition via sampling from the whole population. Where uniform sampling selects each person with equal probability, currently-used greedy algorithms do not even guarantee a minimum selection probability for members of the pool, let alone fair end-to-end probabilities with which members of the population will end up on the panel. As a further downside, the greedy algorithms we have seen being applied may need many attempts to produce a valid panel and might take exponential time to produce a valid panel even if one exists.

[1] https://www.youtube.com/watch?v=hz2d_8eBEKg at 8:53.
[2] https://2016-2018.citizensassembly.ie/en/
[3] https://www.politico.eu/article/belgium-democratic-experiment-citizens-assembly/

2.1.1 Our Techni es and Results

The main contribution of this paper is a more principled sampling algorithm that, even in the setting of limited participation, retains the individual fairness of sampling without replacement while allowing the deterministic satisfaction of quotas. In particular, our algorithm satis es the following desiderata:

> End-to-End Fairness: The algorithm selects the panel via a process such that all members of the population appear on the panel with probability asymptotically close to $\frac{k}{n}$. This also implies that all groups in the population have near-proportional expected representation.
>
> Deterministic Quota Satisfaction: The selected panel satis es certain upper and lower quotas enforcing approximate representation for a set of speci ed features.
>
> Computational E ciency: The algorithm returns a valid panel (or fails) in polynomial time.

Deterministic quota satisfaction is a guarantee of group fairness, while end-to-end fairness, which recovers most of the ex ante guarantees of sampling without replacement, can be seen primarily as a guarantee of individual fairness. The phrase "end-to-end" refers to the fact that we are fair to individuals with respect to their probabilities of going from "population" to "panel", across the intermediate steps of being invited, opting into the pool, and being selected for the panel.

The key challenge in satisfying these desiderata is self-selection bias, which can result in the pool being totally unrepresentative of the population. In the worst case, the pool can be so skewed that it contains no representative panel — in fact, the pool might not even contain k members. As a result, no algorithm can produce a valid panel from every possible pool. However, we are able to give an algorithm that succeeds with high probability, under weak assumptions mainly relating the number of invitation letters sent out to n and the minimum participation probability over all agents.

Crucially, any sampling algorithm that gives (near-)equal selection probability to all members of the population must reverse the self-selection bias occurring in the formation of the pool. We formalize this self-selection bias by assuming that each agent i in the population agrees to join the pool with some positive participation probability π_i when invited. If these π_i values are known for all members of the pool, our sampling algorithm can use them to neutralize self-selection bias. To do so, our algorithm selects agent i for the panel with a probability (close to) proportional to $1/\pi_i$, conditioned on i being in the pool. This compensates for agents' di ering likelihoods of entering the pool, thereby giving all agents an equal end-to-end probability. On a given pool, the algorithm assigns marginal selection probabilities to every agent in the pool. Then, to nd a distribution over valid panels that implements these marginals, the algorithm randomly rounds a linear program using techniques based on discrepancy theory. Since our approach aims for a fair distribution of valid panels rather than just a single panel, we can give probabilistic fairness guarantees.

As we mentioned, our theoretical and algorithmic results take the probabilities π_i of all pool members i as given in the input. While these values are not observed in practice, we then show that

they can be estimated from available data. We cannot directly train a classifier predicting participation, however, because practitioners collect data only on those who join the pool, yielding only positively labeled data. In place of a negatively labeled control group, we use publicly available survey data, which is unlabeled (i.e., includes no information on whether its members would have joined the pool). To learn in this more challenging setting, we use techniques from learning from contaminated controls, which combine the pool data with the unlabeled sample of the population to learn a predictive model for agents' participation probabilities. Finally, we use data from a real-world sortition panel to show that plausible participation probabilities can be learned and that the algorithm produces panels that are close to proportional across features. For a synthetic population produced by extrapolating the real data, we show that our algorithm obtains fair end-to-end probabilities.

2.1.2 Related Work

Our work is broadly related to existing literature on fairness in the areas of machine learning, statistics, and social choice. Through the lens of fair machine learning, our quotas can be seen as enforcing approximate statistical fairness for protected groups, and our near-equal selection probability as a guarantee on individual fairness. Achieving simultaneous group- and individual-level fairness is a commonly discussed goal in fair machine learning [49, 151, 166], but one that has proven somewhat elusive. To satisfy fairness constraints on orthogonal protected groups, we draw upon techniques from discrepancy theory [34, 40], which we hope to be more widely applicable in this area.

Our paper addresses self-selection bias, which is routinely faced in statistics and usually addressed by sample reweighting. Indeed, our sampling algorithm can be seen as a way of reweighting the pool members under the constraint that weights must correspond to the marginal probabilities of a random distribution. While reweighting is typically done by the simpler methods of post-stratification, calibration [165], and sometimes regression [233], we use the more powerful tool of learning with contaminated controls [185, 277] to determine weights on a more fine-grained level.

Our paper can also be seen as a part of a broader movement towards statistical approaches in social choice [195, 197, 252]. The problem of selecting a representative sortition panel can be seen as a fair division problem, in which n indivisible copies of a scarce resource must be randomly allocated such that an approximate version of the proportionality axiom is imposed. Our group fairness guarantees closely resemble the goal of apportionment, in which seats on a legislature are allocated to districts or parties such that each district is proportionally represented within upper and lower quotas [13, 58, 157].

So far, only few papers in computer science and statistics directly address sortition [44, 246, 274]. Only one of them [44] considers, like us, how to sample a representative sortition panel. Unfortunately, their stratified sampling algorithm assumes that all agents are willing to participate, which, as we address in this paper, does not hold in practice.

2.2 Model

Agents. Let $\mathcal{A}$ be a set of n agents, constituting the underlying population. Let $\mathcal{F}$ be a set of features, where feature $f \in \mathcal{F}$ is a function $f : \mathcal{A} \to V_f$, mapping the agents to a set V_f of possible values of feature f. For example, for the feature gender, we could have $V_{\text{gender}} = \{\text{male}, \text{female}, \text{non-binary}\}$. Let the feature-value pairs be $\mathcal{D} = \bigcup_{f \in \mathcal{F}} \{(f, v) \mid v \in V_f\}$. In our example, the feature-value pairs are (gender, male), (gender, female), and (gender, non-binary). Denote the number of agents with a particular feature-value pair (f, v) by $n_{(f,v)}$.

Each agent $a \in \mathcal{A}$ is described by her feature vector $\phi(a) := \{(f, f(a)) \mid f \in \mathcal{F}\}$, the set of all feature-value pairs pertaining to this agent. Building on the example instance, suppose we add the feature education-level, so $\mathcal{F} = \{\text{gender}, \text{education level}\}$. If education level can take on the values college and no college, a college-educated woman would have the feature-vector $\{(\text{gender}, \text{female}), (\text{education level}, \text{college})\}$.

Panel Selection Process. Before starting the selection process, organizers of a sortition panel must commit to the panel's parameters. First, they must choose the number of recipients R who will be invited to potentially join the panel, and the required panel size k. Moreover, they must choose a set of features and values $\{V_f\}_{f \in \mathcal{F}}$ over which quotas will be imposed. Finally, for all feature-value pairs (f, v), they must choose a lower quota $\ell_{(f,v)}$ and an upper quota $\mathcal{D}_{(f,v)}$ implying that the eventual panel of k agents must contain at least $\ell_{(f,v)}$ and at most $\mathcal{D}_{(f,v)}$ agents with value v for feature f. Once these parameters are fixed, the panel selection process proceeds in three steps:

$$\text{population} \xrightarrow{\text{STEP 1}} \text{recipients} \xrightarrow{\text{STEP 2}} \text{pool} \xrightarrow{\text{STEP 3}} \text{panel}$$

In STEP 1, the organizer of the panel sends out R letters, inviting a subset of the population (sampled with equal probability and without replacement) to volunteer for serving on the panel. We refer to the random set of agents who receive these letters as the recipients. Only the agents in the recipients will have the opportunity to advance in the process toward being on the panel.

In STEP 2, each letter recipient may respond affirmatively to the invitation, thereby opting into the pool of agents from which the panel will be chosen. These agents form the random set, the pool, defined as the set of agents who received a letter and agreed to serve on the panel if ultimately chosen. We assume that each agent a joins the pool with some participation probability $\pi_a > 0$. Let π be the lowest value of π_a across all agents $a \in \mathcal{A}$. A key parameter of an instance is $\alpha := R/k$, which measures how large the number of recipients is relative to the other parameters. Larger values of α will allow us the flexibility to satisfy stricter quotas.

In STEP 3, the panel organizer runs a sampling algorithm, which selects the panel from the pool. This panel, denoted as the set panel, must be of size k and satisfy the predetermined quotas for all feature-value pairs. The sampling algorithm may also fail without producing a panel.

We consider the first two steps of the process to be fully prescribed. The focus of this paper is to develop a sampling algorithm for the third step that satisfies the three desiderata listed in the

introduction: end-to-end fairness, deterministic quota satisfaction, and computational e ciency.

2.3 Sampling Algorithm

In this section, we give an algorithm which ensures, under natural assumptions, that every agent ends up on the panel with probability at least 1 >¹1º : •= as= goes to in nity.[1] Furthermore, the panels produced by this algorithm satisfy non-trivial quotas, which ensure that the ex-post representation of each feature-value pair cannot be too far from being proportional.

Our algorithm proceeds in two phases: assignment of marginals, during which the algorithm assigns a marginal selection probability to every agent in the pool, and rounding of marginals, in which the marginals are dependently rounded to 0-1 values, the agents' indicators of being chosen for the panel. As we discussed previously, our algorithm succeeds only with high probability, rather than deterministically; it may fail in phase I if the desired marginals do not satisfy certain conditions. We refer to pools on which our algorithm succeeds as good pools. A good pool, to be de ned precisely later, is one that is highly representative of the population — that is, its size and the prevalence of all feature values within it are close to their respective expected values. We leave the behavior of our algorithm on bad pools unspeci ed: while the algorithm may try its utmost on these pools, we give no guarantees in these cases, so the probability of representation guaranteed to each agent must come only from good pools and valid panels. Fortunately, under reasonable conditions, we show that the pool will be good with high probability. When the pool is good, our algorithm always succeeds, meaning that our algorithm is successful overall with high probability.

Our algorithm satis es the following theorem, guaranteeing close-to-equal end-to-end selection probabilities for all members of the population as well as the satisfaction of quotas.

Theorem 2.3.1. Suppose that U ! 1 and =5•E =•: for all feature-value pairs 5 • E. Consider a sampling algorithm that, on a good pool, selects a random panel Panel, via the randomized version of lemma 2.3.3, and else does not return a panel. This process satis es, for all 8 population, that

$$\mathbb{P}[8 \in \text{Panel}] = 1 - (1 - \cdot)^{\circ} : \cdot =$$

All panels produced by this process satisfy the quotas B ¹1 U ”49º : = 5•E= j j and D5•E B ¹1 ¸ U ”49º : = 5•E= ¸ j j for all feature-value pairs 5 • E.

The guarantees of the theorem grow stronger as the parameter @A•: tends toward in nity, i.e., as the number A of invitations grows. Note that, since A =, this assumption requires that @ : •=. We defer all proofs to appendix A.2 and discuss the preconditions in appendix A.2.1.

2.3.1 Algorithm Part I: Assignment of Marginals

To a ord equal probability of panel membership to each agent, we would like to select agent 8 with probability inversely proportional to her probability @8 of being in the pool. For ease of

[1] We allow : 1 and A 1 to vary arbitrarily in = and assume that the feature-value pairs are xed.

notation, let $w_i = 1/n$ for all i. Specifically, for agent i, we want $\Pr[i \in \text{Panel} \mid i \in \text{Pool}]$ to be proportional to w_i. Achieving this exactly is tricky, however, because each agent's selection probability from pool P, call it $c_{i,P}$, must depend on those of all other agents in the pool, since their marginals must add to the panel size. Thus, instead of reasoning about an agent's probability across all possible pools at once, we take the simpler route of setting agents' selection probabilities for each pool separately, guaranteeing that $\Pr[i \in \text{Panel} \mid i \in P]$ is proportional to w_i across all members i of a good pool P. For any good pool P, we select each agent $i \in P$ for the panel with probability

$$c_{i,P} := w_i \cdot \frac{1}{\sum_{j \in P} w_j}.$$

Note that this choice ensures that the marginals always sum up to k.

Definition of Good Pools. For this choice of marginals to be reasonable and useful for giving end-to-end guarantees, the pool P must satisfy three conditions, whose satisfaction defines a good pool P. First, the marginals do not make much sense unless all $c_{i,P}$ lie in $[0, 1]$:

$$0 \le c_{i,P} \le 1 \quad \forall i \in P. \tag{2.1}$$

Second, the marginals summed up over all pool members of a feature-value pair $f = (\ell, v)$ should not deviate too far from the proportional share of the pair:

$$(1 - \gamma)\frac{|\{i \in P : x_{i\ell} = v\}|}{n} \le \sum_{i \in P : x_{i\ell} = v} c_{i,P} \le (1 + \gamma)\frac{|\{i \in P : x_{i\ell} = v\}|}{n} \quad \forall f = (\ell, v). \tag{2.2}$$

Third, we also require that the term $\sum_{i \in P} w_i$ is not much larger than $\mathbb{E}_{i \in \text{Pool}}[w_i] = A$, which ensures that the $c_{i,P}$ do not become too small:

$$\sum_{i \in P} w_i \le A \cdot (1 + \gamma). \tag{2.3}$$

Under the assumptions of our theorem, pools are good with high probability, even if we condition on any agent i being in the pool:

Lemma 2.3.2. Suppose that $\alpha \le 1$ and $|\{i \in P : x_{i\ell} = v\}| \ge \tau$ for all (ℓ, v). Then, for all agents $i \in$ Population, $\Pr[P \text{ is good} \mid i \in \text{Pool}] \ge 1 - \alpha$.

Note that only constraint (2.1) prevents Phase II of the algorithm from running; the other two constraints just make the resulting distribution less useful for our proofs. In practice, if it is possible to rescale the $c_{i,P}$ and cap them at 1 such that their sum is k, running phase II on these marginals seems reasonable.

2.3.2 Algorithm Part II: Rounding of Marginals

The proof of Theorem 2.3.1 now hinges on our ability to implement the chosen $c_{i,P}$ for a good pool P as marginals of a distribution over panels. This phase can be expressed in the language of randomized dependent rounding: we need to define random variables $X_i = \mathbb{1}[i \in \text{Panel}]$ for each $i \in \text{Pool}$ such that $\mathbb{E}[X_i] = c_{i,P}$. This difficulty of this task stems from the ex-post requirements on the pool, which require that $\sum_i X_i = k$ and that $\sum_{i : x_{i\ell} = v} X_i$ is close to $\alpha \cdot |\{i : x_{i\ell} = v\}|$ for all feature-value

pairs $5 \bullet E$While o -the-shelf dependent rounding [78] can guarantee the marginals and the sum-to-: constraint, it cannot simultaneously ensure small deviations in terms of the representation of all $5 \bullet E$

Our algorithm uses an iterative rounding procedure based on a celebrated theorem by Beck and Fiala [40]. We sketch here how to obtain a deterministic rounding satisfying the ex-post constraints; the argument can be randomized using results by Bansal [34] [or via column generation (Appendix A.2.4).[1] The iterated rounding procedure manages a variable $G_8 2 \gg 0 \bullet 1 \frac{1}{4}$ for each $8 2$ Pool, which is initialized as $c_{8 \bullet \%}$ As the G_8 are repeatedly updated, more of them are xed as either 0 or 1 until the G_8 ultimately correspond to indicator variables of a panel. Throughout the rounding procedure, it is preserved that $\sum_8 G_8 = \sum_8 c_{8 \bullet \%} = :$, and the equalities $\sum_{8 5 1 8 ^ p = E} G_8 = \sum_{8 5 1 8 ^ p = E} c_{8 \bullet \%}$ are preserved until at most j variables G_8 in the sum are yet to be xed. As a result, the nal panel has exactly y members, and the number of members from a feature-value pair $5 \bullet E$ is at least $\sum_{8 5 1 8 ^ p = E} c_{8 \bullet \%} j \quad j \quad ^ 1 \; 1 \quad U \; ” 49 ^ o : = 5 \bullet E = \quad j \quad j$ (symmetrically for the upper bound).[2] As we show in appendix A.2.4,

Lemma 2.3.3. There is a polynomial-time sampling algorithm that, given a good pool $\%$ produces a random panel Panel such that (1) $P \gg 8 2$ Panel $\frac{1}{4} = c_{8 \bullet \%}$ for all $8 2 \%$, (2) jPanel$j = :$, and (3) $\sum_{8 5 1 8 ^ p = E} c_{8 \bullet \%} j \quad j \quad j f \; 8 2$ Panel $j \; 5 1 8 ^ p = E g j \quad \sum_{8 5 1 8 ^ p = E} c_{8 \bullet \%} j \quad j.$

Our main theorem follows from a simple argument combining Lemma 2.3.2 and 2.3.3 (Appendix A.2.5).

While the statement of theorem 2.3.1 is asymptotic in the growth of U, the same proof gives bounds on the end-to-end probabilities for nite values of U If one wants bounds for a speci c instance, however, bounds uniquely in terms of U tend to be loose, and one might want to relax Condition (2.2) of a good pool in exchange for more equal end-to-end probabilities. In this case, plugging the speci c values of $f \bullet A \bullet : \bullet @ = 5 \bullet E$ into the proof allows to make better trade-o s and to extract sharper bounds.

2.4 Learning Participation Probabilities

The algorithm presented in the previous section relies on knowing $@_8$ for all agents 8 in the pool. While these $@_8$ are not directly observed, we can estimate them from data available to practitioners.

First, we assume that an agent 8’s participation probability $@_8$ is a function of her feature vector $^ 1 8 ^ p$. Furthermore, we assume that 8 makes her decision to participate through a speci c generative model known as simple independent action [119, as cited in [280]]. First, she ips a coin with

[1] Bansal [34] gives a black-box polynomial-time method for randomizing our rounding procedure. We found column-generation-based algorithms to be faster in practice, with guarantees that are at least as tight.

[2] Observe that our Beck-Fiala-based rounding procedure only increases the looseness of the quotas by a constant additive term beyond the losses to concentration. The concentration properties of standard dependent randomized rounding do not guarantee such a small gap with high probability. Moreover, our bound does not directly depend on the number of quotas (i.e., twice the number of feature-value pairs) but only depends on the number of features, which are often much fewer.

probability V_0 of landing on heads. Then, she flips a coin for each feature $f \in F$, where her coin pertaining to f lands on heads with probability $V_{f,\text{flip}}$. She participates in the pool if and only if all coins she flips land on heads, leading to the following functional dependency:

$$\text{Pool}(x) = V_0 \prod_{f \in F} V_{f,\text{flip}}^{x_f}$$

We think of $1 - V_{f,\text{flip}}$ as the probability that a reason specific to the feature-value pair $f = x_f$ prevents the agent from participating, and of $1 - V_0$ as the baseline probability of her not participating for reasons independent of her features. The simple independent action model assumes that these reasons occur independently between features, and that the agent participates if none of the reasons occur.

If we had a representative sample of agents — say, the recipients of the invitation letters — labeled according to whether they decided participate (positive) or not (negative), learning the parameters V would be straightforward. However, sortition practitioners only have access to the features of those who enter the pool, and not of those who never respond. Without a control group, it is impossible to distinguish a feature that is prevalent in the population and associated with low participation rate from a rare feature associated with a high participation rate. Thankfully, we can use additional information: in place of a negatively-labeled control group, we use a *background sample*—a dataset containing the features for a uniform sample of agents, but without labels indicating whether they would participate. Since this control group contains both positives and negatives, this setting is known as *contaminated controls*. A final piece of information we use for learning is the fraction $\Pr(\text{Pool}) = A$, which estimates the mean participation probability across the population. In other applications with contaminated controls, including in the estimation increased model identifiability [277].

To learn our model, we apply methods for maximum likelihood estimation (MLE) with contaminated controls introduced by Lancaster and Imbens [185]. By reformulating the simple independent action model in terms of the logarithms of the parameters, their estimation (with a fixed value of θ) reduces to maximizing a concave function.

Theorem 2.4.1. The log-likelihood function for the simple independent action model under contaminated controls is concave in the model parameters.

By this theorem, proven in Appendix A.3, we can directly and efficiently estimate V. Logistic models, by contrast, require more involved techniques for efficient estimation [277].

2.5 Experiments

Data. We validate our θ estimation and sampling algorithm on pool data from Climate Assembly UK,[1] a national-level sortition panel organized by the Sortition Foundation in 2020. The panel consisted of $n = 110$ many UK residents aged 16 and above. The Sortition Foundation invited all members of 30 000 randomly selected households, which reached an estimated $A = 60\,000$ eligible

[1] https://www.climateassembly.uk/

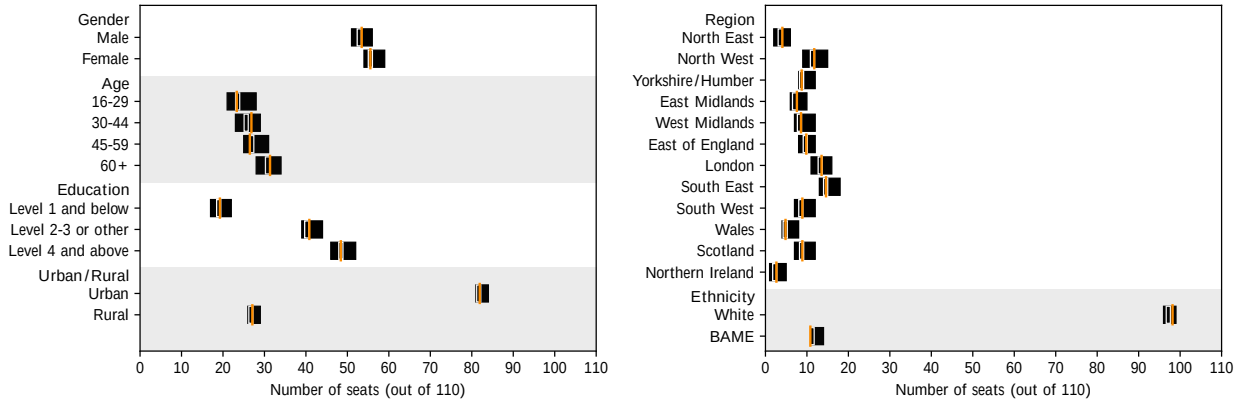


Figure 2.1: Expected and realized numbers of panel seats our algorithm gives each feature-value pair in the Climate Assembly pool.

participants.[1] Of these letter recipients, 1 715 participated in the pool,[2] corresponding to a mean participation probability of @ 2”9% The feature-value pairs used for this panel can be read o the axis of g. 2.1. We omit an additional feature, climate concern level, in our main analysis because only 4 members of the pool have the value not at all concerned, whereas this feature-value pair's proportional number of panel seats is 6.5. To allow for proportional representation of groups with such low participation rates, A should have been chosen to be much larger. We believe that the merits of our algorithm can be better observed in parameter ranges in which proportionality can be achieved. For the background sample, we used the 2016 European Social Survey [2018], which contains 1,915 eligible individuals, all with features and values matching those from the panel. Our implementation is based on PyTorch and Gurobi, runs on consumer hardware, and its code is available on github. Appendix A.4 contains details on Climate Assembly UK, data processing, the implementation, and further experiments (including the climate concern feature).

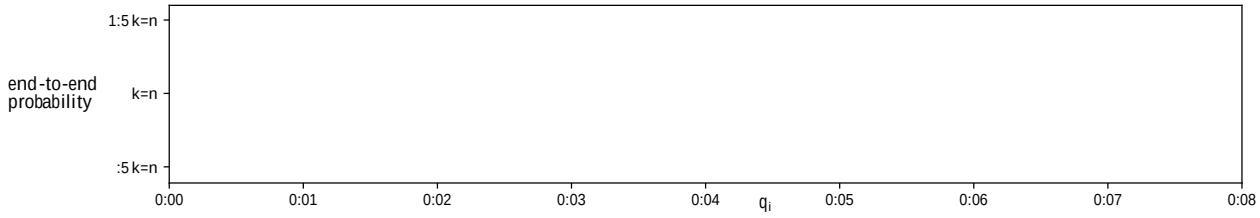
Estimation of mV Parameters. We nd that the baseline probability of participation is $\hat{V}_0$ = 8”8% Our $\hat{V}_{5.E}$ estimates suggest that (from strongest to weakest e ect) highly educated, older, urban, male, and non-white agents participate at higher rates. These trends re ect these groups' respective levels of representation in the pool compared to the underlying population, suggesting that our estimated V values t our data well. Di erent values of the remaining feature, region of residence, seem to have heterogeneous e ects on participation, where being a resident of the South West gives substantially increased likelihood of participation compared to other areas. The lowest participation probability of any agent in the pool, according to these estimates, is @ = 0”78% implying that U 4”25. See Appendix A.4.4 for detailed estimation results and validation.

[1] Note that every person in the population has equal probability of being invited. We ignore correlations between members of the same household.

[2] Excluding 12 participants with gender other as no equivalent value is present in the background data.

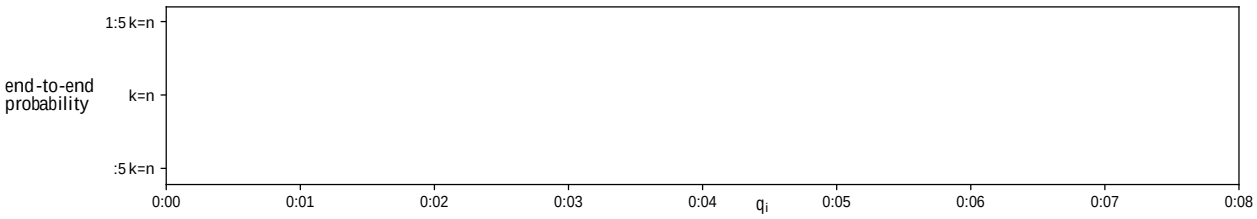
Running the Sampling Algorithm on the Pool. The estimated q_i allow us to run our algorithm on the Climate Assembly pool and thereby study its fairness properties for non-asymptotic input sizes. We find that the Climate Assembly pool is good relative to our q_i estimates, i.e., that it satisfies eqs. (2.1) to (2.3). As displayed in fig. 2.1, the marginals produced by Phase I of our algorithm give each feature-value pair f an expected number of seats $k \cdot p_f$ within one seat of its proportional share of the panel, $k \cdot p_f$. By lemma 2.3.3, Phase II of our algorithm then may produce panels from these marginals in which f receives up to j $j = 6$ fewer or more seats than its expected number. However, as the black bars in fig. 2.1 show, the actual number of seats received by any f across any panel produced by our algorithm on this input never deviates from its expectation by more than 4 seats. As a result, while theorem 2.3.1 only implies lower quotas of $\lfloor 51 \cdot k/n \rfloor = \lfloor 5 \cdot k/n \rfloor - j$ j and upper quotas of $\lceil 49 \cdot k/n \rceil = \lceil 5 \cdot k/n \rceil + j$, j for this instance, the shares of seats our algorithm produces lie in the much narrower range $\lfloor 5 \cdot k/n \rfloor - 5$ (and even $\lfloor 5 \cdot k/n \rfloor - 3$ for 18 out of 25 feature-value pairs). This suggests that, while the quotas guaranteed by our theoretical results are looser than the quotas typically set by practitioners, our algorithm will often produce substantially better ex-post representation than required by the quotas.

End-to-End Probabilities. In the previous experiments, we were only able to argue about the algorithm's behavior on a single pool. To validate our guarantees on individual end-to-end probabilities, we construct a synthetic population of size 60 million by duplicating the ESS participants, assuming our estimated q_i as their true participation probabilities. Then, for various values of A, we sample a large number of pools. By computing $\pi_{i\%}$ values for all agents i in each pool, we can estimate each agent's end-to-end probability of ending up on the panel. Crucially, we assume that our algorithm does not produce any panel for bad pools, analogously to theorem 2.3.1. As shown in the following graph, for $A = 60\,000$ (as was used in Climate Assembly UK), all agents in our synthetic population, across the full range of q_i, receive probability within $\pm 1 \cdot k/n$ of k/n (averaged over 100 000 random pools):



That these end-to-end probabilities are so close to k/n also implies that bad pools are exceedingly rare for this value of A. As we show in appendix A.4.6, we see essentially the same behavior for values of A down to roughly 15 000, when $U \approx 1$. For even lower A, most pools are bad, so end-to-end probabilities are close to zero under our premise that no panels are produced from bad pools.

To demonstrate that our algorithm's theoretical guarantees lead to realized improvements in individual fairness over the state-of-the-art, we re-run the experiment above, this time using the Sortition Foundation's greedy algorithm to select a panel from each generated pool. Since their algorithm requires explicit quotas as input, we set the lower and upper quotas for each feature-value group to be the floor and ceiling of that group's proportional share of seats. This is a popular way of setting quotas in current practice.



The results of this experiment show that the individual end-to-end probabilities generated by the currently-used greedy algorithm range from below 0.5 times the target up to 1.3 times the target. In comparison to the end-to-end probabilities generated by our algorithm, those generated by the greedy algorithm are substantially skewed, and tend to disadvantage individuals with either low or high participation probabilities. One might argue that the comparison between our algorithm and the greedy is not quite fair, since the greedy algorithm is required to satisfy stronger quotas. However, looser quotas do not improve the behavior of the greedy algorithm; they simply make it behave more similarly to uniform sampling from the pool, which further disadvantages agents with low participation probability (for details, see appendix A.4.5).

Taken together, these results illustrate that, although greedy algorithms like the one we examined achieve proportional representation of a few pre-specified groups via quotas, they do not achieve fairness to individuals or to groups unprotected by quotas. Compared to the naive solution of uniform sampling from the pool, greedily striving for quota satisfaction does lead to more equal end-to-end probabilities, as pool members with underrepresented features are more likely to be selected for the panel than pool members with overrepresented features. However, this effect does not neutralize self-selection bias when there are multiple features, even when selection bias acts through the independent-action model as in our simulated population. Indeed, in this experiment, the greedy algorithm insufficiently boosts the probabilities of agents in the intersection of multiple low-participation groups (the agents with lowest π_i), while also too heavily dampening the selection probability of those in the intersection of multiple high-participation groups (with highest π_i). These observations illustrate the need for panel selection algorithms that explicitly control individual probabilities.

2.6 Discussion

In a model in which agents stochastically decide whether to participate, our algorithm guarantees similar end-to-end probabilities to all members of the population. Arguably, an agent's decision to participate when invited might not be random, but rather deterministically predetermined.

From the point of view of such an agent i, does our algorithm, based on a model that doesn't accurately describe her (and her peers') behavior, still grant her individual fairness? If i deterministically participates, the answer is yes (if not, of course she cannot be guaranteed anything). To see why, first observe that, insofar as it concerns i's chance of ending up on the panel, all other agents might as well participate randomly.[1] Indeed, from agent i's perspective, the process looks

[1] Fix a group of agents who, assuming the stochastic model, will participate if invited with probability π. Then, sampling letter recipients from this set of agents in the stochastic model is practically equivalent to sampling recip-

like the stochastic process where every other agent participates with probability q, where herself always participates, and where the algorithm erroneously assumes that she participates only with some probability q. Therefore, the pool is still good with high probability conditioned on being in it, as argued in lemma 2.3.2. Even if the algorithm knew that $q = 1$, her end-to-end probability would be at least $1 - (1 - \cdot)^2 = \cdot - \cdot^2$, and the fact that the algorithm underestimates her q only increases her probability of being selected from the pool. It follows that her end-to-end probability in this setting still must be at least around $2\cdot$.

Thus, in a deterministic model of participation, our individual guarantees are reminiscent of the axiom of population monotonicity in fair division: If the whole population always participated when invited, every agent would reach the panel with probability $\cdot$. The fact that some agents do not participate cannot (up to lower-order terms) decrease the selection probabilities for those who do.

ients from this group in the deterministic model, if a q fraction of the group deterministically participate.

3

A Framework of Sortition Algorithms

Fair Algorithms for Selecting Citizens Assemblies [136].
Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, & Ariel D. Procaccia.
Nature, 2021.

This exposition is adapted from the slightly modified version of this paper enclosed in [152].

3.1 Introduction

In representative democracies, political representatives are usually selected by election. However, over the last 35 years, an alternative selection method has been gaining traction among political scientists [71, 87, 90] and practitioners [126, 203, 222, 225]: sortition, the random selection of representatives from the population. The chosen representatives form a panel, commonly called a citizens' assembly, which convenes to deliberate on a policy question. Citizens' assemblies are now being administered by around 50 organizations in over 25 countries [88], and just one of these organizations, the Sortition Foundation in the UK, recruited 29 panels in 2020. While many citizens' assemblies are initiated by civil-society organizations [71, 87, 90, 92, 126, 203, 222, 225], they are also increasingly being commissioned by public authorities on municipal, regional, national, and supranational levels [222]. In fact, since 2019, multiple regional parliaments in Belgium and the Council of Paris have internally established permanent sortition bodies [137, 215]. Citizens' assemblies' growing utilization by governments is giving their decisions a more direct path to policy impact. For example, two recent citizens' assemblies commissioned by Ireland's national legislature led to the legalization of same-sex marriage and abortion [169].

Ideally, a citizens' assembly selected via sortition acts as a microcosm of society: its participants are representative of the population, and thus its deliberation simulates the entire population

convening under conditions where it can really consider competing arguments and get its questions answered from different points of view [124]. Whether this goal is realized in practice, however, depends on exactly how assembly members are chosen.

Panel selection is generally done in three stages: first, thousands of randomly chosen constituents are invited to participate. Second, a subset of the invited constituents opt into a pool of volunteers. Third, a panel of pre-specified size is randomly chosen from the pool via some fixed procedure, which we call a *selection algorithm* [82, 170, 211, 218]. As the final and most complex component of the selection process, the selection algorithm has great power in deciding who will be chosen to represent the population. In this chapter, we introduce selection algorithms that preserve the key desirable property of existing algorithms, while also more fairly distributing the sought-after opportunity [82, 170, 211, 218] of being a representative.

To our knowledge, all of the selection algorithms used in practice aim to satisfy one particular property, known as *descriptive representation*, the idea that the panel should reflect the composition of the population [124]. Unfortunately, the pool from which the panel is chosen tends to be far from representative. Specifically, it tends to overrepresent groups whose members are more likely to accept an invitation to participate, such as high educational attainment. To ensure descriptive representation despite the biases of the pool, selection algorithms require that the panels they output satisfy upper and lower *quotas* on a set of specified features, which are roughly proportional to each feature's population rate (e.g. quotas might require that a 40-person panel contain between 20 and 21 women). These quotas are generally imposed on feature categories delineated by gender, age, education level, and other attributes relevant to the policy issue at hand. We note that quota constraints of this form are more general than those achievable via stratified sampling, a common technique for drawing representative samples.

Selection algorithms that pre-date this work focused solely on satisfying quotas, leaving unaddressed a second property that is also central to sortition: that all individuals should have an equal chance of being chosen for the panel. Several political theorists present equality of selection probabilities as a central advantage of sortition, stressing its role in promoting the ideals such as equality of opportunity [71, 224], democratic equality [123, 124, 224, 258], and allocative justice [257, 258]. In fact, Engelstad, who introduced an influential model of sortition's benefits, argues that this form of equality constitutes "The strongest normative argument in favor of sortition" [112]. (See Appendix B.4 for more details on sortition desiderata from political theory.) In addition to political theorists, major practitioner groups have also advocated for equal selection probabilities [21, 201]. However, they face the fundamental hurdle that, in practice, the quotas almost always necessitate selecting people with somewhat unequal probabilities, as individuals from groups that are underrepresented in the pool must be chosen with disproportionately high probabilities to satisfy the quotas.

Though it is generally impossible to achieve *perfectly* equal probabilities, the reasons to strive for equality also motivate a more gradual version of this goal: making probabilities as equal as possible, subject to the quotas. We refer to this goal as *maximal fairness*. We find that our benchmark, a selection algorithm representing the previous state of the art, falls far short of this

goal, giving volunteers drastically unequal probabilities across several real-world instances. This algorithm even consistently selects certain types of volunteers with near-zero probability, thereby excluding them in practice from the chance to serve. We further show that, in these instances, it is possible to give all volunteers probability well above zero while still satisfying the quotas, demonstrating that the level of inequality produced by the benchmark is avoidable.

In this chapter, we close the gaps we have identified, both in theory and in practice. We first introduce not just one selection algorithm that achieves maximal fairness, but a more general (I) algorithmic framework for producing such algorithms. Motivated by the multitude of possible ways to quantify the fairness of an allocation of selection probabilities, our framework gives a maximally fair selection algorithm for any measure of fairness with a certain functional form. Notably, such measures include the most prominent from the literature on fair division [55, 206], and we show that these well-established metrics can be applied to our setting by casting the problem of assigning selection probabilities as one of fair resource allocation. Then, to bring this innovation into practice, we implement a (II) deployable selection algorithm, which is maximally fair according to one specific measure of fairness. We evaluate this algorithm and find that it is substantially fairer than the benchmark on several real-world datasets and by multiple fairness measures. Our algorithm is now in use by a growing number of sortition organizations around the world, making it one of only a few [62, 114, 150, 261] deployed applications of fair division.

3.2 Contribution I: Algorithmic Framework

3.2.1 Definitions

We begin by introducing necessary terminology. We refer to the input to a selection algorithm—a pool of size n, a set of quotas, and the desired panel size k—as an *instance* of the panel selection problem. Given an instance, a selection algorithm randomly selects a *panel*, which is a quota-compliant set of k pool members. We define the algorithm's *output distribution* on an instance as the distribution specifying the probabilities with which the algorithm outputs each possible panel. Then, a pool member's *selection probability* is the probability that they are on a panel randomly drawn from the output distribution. We refer to the mapping from pool members to their selection probabilities as the *probability allocation*, which we aim to make as fair as possible. Finally, a *fairness measure* is a function that maps a probability allocation to a fairness score (e.g. the geometric mean of probabilities, where higher is fairer). An algorithm is called *optimal* with respect to a fairness measure if, on any instance, the fairness of the algorithm's probability allocation is at least as high as that of any other algorithm.

3.2.2 Formulating the Optimization Task

To inform our approach, we first analyze the algorithms pre-dating ours. Those we have seen in use all have the same high-level structure: they select individuals for the panel one-by-one, in each step randomly choosing whom to add next from among those who, according to a myopic heuristic, seem unlikely to produce a quota violation later. Since finding a quota-compliant panel

is an algorithmically hard problem,[1] it is already an achievement that such simple algorithms find any panel in most practical instances. Due to their focus on finding any panel at all, however, these algorithms do not tightly control which panel they output, or more precisely, their output distribution (the probabilities with which they output different panels). Since an algorithm's output distribution directly determines its probability allocation, existing algorithms' probability allocations are also uncontrolled, leaving room for them to be highly unfair. In contrast to these

Figure 3.1: The steps of the algorithm optimizing the fairness measure. The left-hand panel shows the implementation of step (1): constructing a maximally fair output distribution over panels (denoted by white boxes), which is done by iteratively building an optimal portfolio of panels and computing the fairest distribution over that portfolio. The right-hand panel shows step (2): sampling the distribution to select a final panel.

existing algorithms, which have output distributions that arise implicitly from a sequence of myopic steps, the algorithms in our framework (1) explicitly compute their own output distribution, and then (2) sample from that distribution to select the final panel (Figure 3.1). Crucially, the maximal fairness of the output distribution found in the first step makes our algorithms optimal. To see why, note that the behavior of any selection algorithm on a given instance is described by some output distribution; thus, since our algorithm finds the fairest possible output distribution, it is always at least as fair as any other algorithm.

Since step (2) of our selection algorithm is simply a random draw, we have reduced the problem of finding an optimal selection algorithm to the optimization problem in step (1) finding a maximally fair distribution over panels. Now, to fully specify our algorithm, it remains only to solve this optimization problem.

[1] See supplementary information 6

3.2.3 Solving the Optimization Task

A priori, it would seem that computing a maximally fair distribution might require constructing all possible panels, since achieving optimal fairness might necessitate assigning non-zero probability to all of them. Such an approach would be impracticable, however, as the number of panels in most instances is intractably large. Fortunately, since we measure fairness according to only individual selection probabilities, there must exist an optimal portfolio a set of panels over which there exists a maximally fair distribution containing few panels by Carathéodory's theorem:

Proposition 3.2.1. Fix an arbitrary instance and a fairness measure for this instance. If there exists any maximally fair distribution over panels for , there exists a maximally fair output distribution whose support includes at most 1 panels.

Proof. Consider the hypercube $[0,1]^n$, and associate each dimension with one pool member. A panel P can be embedded into this space by its characteristic vector $v_P \in \{0,1\}^n$, whose ith component is one exactly if pool member i is contained in P.

Fix a maximally fair output distribution, let $\mathcal{P}$ denote its support, and let $(\lambda_P)_{P \in \mathcal{P}}$ denote its probability mass function. Note that

$$\vec{x} = \sum_{P \in \mathcal{P}} \lambda_P v_P$$

is a probability allocation maximizing , and that it is a convex combination of the $\{v_P\}_{P \in \mathcal{P}}$. By Carathéodory's theorem, there is a subset $\mathcal{P}' \subseteq \mathcal{P}$ of size at most $n+1$ such that $\vec{x}$ still lies in the convex hull of this smaller set. Thus, there are nonnegative real numbers $(\lambda'_P)_{P \in \mathcal{P}'}$ adding up to one such that

$$\vec{x} = \sum_{P \in \mathcal{P}'} \lambda'_P v_P.$$

These λ'_P form the probability mass function of a distribution over at most $n+1$ panels, which has the same probability allocation $\vec{x}$ as the original maximally fair distribution, which implies that the new distribution is also maximally fair for .

This result brings a practical algorithm within reach, and shapes the goal of our algorithm: to find an optimal portfolio while constructing as few panels as possible.

We accomplish this goal using an algorithmic technique called *column generation*, where, in our case, the columns being generated correspond to panels. A more in-depth discussion and formal description of this algorithm, as well as proofs of correctness, can be found in supplementary information 8. As shown in fig. 3.1, our algorithms find an optimal portfolio by iteratively adding panels to a portfolio $\mathcal{P}$, in each iteration alternating between two subtasks: (i) finding the optimal distribution D over only the panels currently in $\mathcal{P}$ and (ii) adding a panel to $\mathcal{P}$ that, based on the gradient of the fairness measure, will move the portfolio furthest towards optimality. This second subtask makes use of integer linear programming, which we use to generate quota-compliant panels despite the theoretical hardness of the problem. Eventually, the panel with the most promising

gradient will already be in P, in which case P is provably optimal and D must be a maximally fair distribution. In practice, we observe that this procedure terminates after few iterations.

Our techniques extend column generation methods that are typically applied to linear programs, allowing them to be used to solve a large set of convex programs. This extension allows our framework to be used with a wide range of fairness measures — essentially any for which the fairest distribution over a portfolio can be found via convex programming. Supported measures include those most prominent in the fair division literature: egalitarian welfare [11], Nash welfare [206], Gini inequality [110, 187], and the Atkinson indices [110, 247] [247]. Our algorithmic approach also has the bene t of easily extending to organization-speci c constraints beyond quotas; for example, practitioners can prevent multiple members of the same household from appearing on the same panel. Due to its generality, our framework even applies to domains outside of sortition, including the allocation of classrooms to charter schools [182] and kidney exchange [243].

3.3 Contribution II: Deployable Selection Algorithm

To bring fair panel selection into practice, we develop an e cient implementation of one speci c selection algorithm, which we call LexiMin (formally de ned in supplementary information 10). LexiMin optimizes the well-established fairness measure *leximin* [51, 182, 206], a fairness measure that is sensitive to the very lowest selection probabilities. In particular, leximin is optimized by maximizing the lowest selection probability, then breaking ties between solutions in favor of probability allocations with highest second-lowest probability, and so on. This choice of fairness measure is motivated by the fact that, as we show in this section and supplementary information 13, Legacy gives some pool members a near-zero probability when much more equal probabilities are possible. This type of unfairness is especially pressing because, if it consistently impacted pool members with certain combinations of features, these individuals and their distinct perspectives would be systematically excluded from participation [250], which runs counter to a key promise of random selection.

To increase the accessibility of LexiMin, we made its implementation available through an existing open-source panel selection tool [164] and on Panelot [153], a website where anyone can run the algorithm without installation. LexiMin has since been deployed by several organizations, including Cascadia (US), the Danish Board of Technology (Denmark), Nexus (Germany), of by for * (US), Particitiz (Belgium), and the Sortition Foundation (UK). As of July 2021, the Sortition Foundation alone had already used LexiMin to select more than 40 panels.

We measure the impact of adopting LexiMin over pre-existing algorithms by comparing its fairness to that of a benchmark, Legacy (supplementary information 11), the algorithm used by the Sortition Foundation prior to their adoption of LexiMin. We choose Legacy as a benchmark because it was widely used prior to this work, it is similar to several other selection algorithms used in practice (see supplementary information 13), and it is the only existing algorithm we found that was fully speci ed by an o cial implementation. We compare the LexiMin and Legacy on ten datasets from real-world panels, with respect to several fairness measures including the minimum probability (table 3.1), the Gini coe cient, and the geometric mean. In this analysis, we nd that

instance	n	k	# of features	k/n	Legacy min. probability (sampled[1])	LexiMin min. probability (exact)	LexiMin running time
sf(a)	312	35	6	11.2%	0.32%	6.7%	20 sec
sf(b)	250	20	6	8.0%	0.17%	4.0%	9 sec
sf(c)	161	44	7	27.3%	0.15%	8.6%	6 sec
sf(d)	404	40	6	9.9%	0.11%	4.7%	46 sec
sf(e)	1727	110	7	6.4%	0.03%	2.6%	67 min
cca	825	75	4	9.1%	0.03%	2.4%	7 min
hd	239	30	7	12.6%	0.09%	5.1%	37 sec
mass	70	24	5	34.3%	14.9%	20.0%	1 sec
nexus	342	170	5	49.7%	2.24%	32.5%	1 min
obf	321	30	8	9.3%	0.03%	4.7%	3 min

Table 3.1: List of instances used in our experiments. For the instances we study, panels were recruited by the following organisations: sf(a-e): Sortition Foundation; cca: Center for Climate Assemblies; hd: Healthy Democracy; mass: MASS LBP; nexus: Nexus; obf: of by for * (At the request of practitioners, topics, dates, and locations of the panels are not identified.) n is the pool size, k is the panel size, and consequently k/n is the mean selection probability. The # of features is j, where each of the j features has between 2 and 49 possible values (with the typical range being 2-5).

LexiMin is fairer on all instances we examine, and substantially so in nine out of ten.

3.4 Effect of Adopting LexiMin over Legacy

We study datasets from ten sortition panels, organized by six different sortition organizations in Europe and North America. As Table 3.1 shows, our instances are diverse in panel size (range: 20–170, median: 37.5) and number of quota categories (range: 4–8). On consumer hardware, the run-time of our algorithm is well within the time available in practice.

Out of concern about low selection probabilities, we first compare the minimum selection probabilities given by Legacy and LexiMin, summarized in the second and third columns from the right in Table 3.1. Strikingly, in all instances except mass (an outlier in that its quotas only mildly restrict the fraction of panels that are feasible), Legacy chooses some pool members with probability close to zero. In fact, we can identify combinations of features that lead to low selection probabilities across all instances,[2] raising the concern that Legacy may in fact systematically exclude some groups from participation. By contrast, LexiMin selects no one nearly so infrequently, with minimum selection probabilities ranging from 26% to 65% (median: 49%) of the ideal probability individuals would receive in the absence of quotas. One might wonder whether this increased minimum probability achieved by LexiMin affects only a few pool members most disadvantaged by Legacy. This is not the case: As shown in Figure 3.2 by the shaded

[2] See methods section "Individuals rarely selected by Legacy" of the full version.

Figure 3.2: Selection probabilities given by Legacy and LexiMin to the bottom 60% of pool members on six representative instances, where pool members are ordered in order of increasing probability given by the respective algorithms. Shaded boxes denote the range of pool members whose selection probability given by Legacy is lower than the minimum probability given by LexiMin. Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% confidence intervals (see methods section Statistics of the full version). For corresponding graphs for all other instances and up to the 100th percentile, see Figures 3.4 and 3.5 respectively in Section 3.6.

boxes, between 13% and 56% of pool members (median 46%) across instances receive probability from Legacy lower than the minimum given to anyone by LexiMin (Table 3.7). Thus, even just the first stage of LexiMin, i.e., maximizing the minimum probability, provides a sizable section of the pool with more equitable access to the panel.

We have so far compared Legacy and LexiMin over only the lower end of selection probabilities, as this is the range in which LexiMin prioritizes being fair. However, even considering the entire range of selection probabilities, we find that LexiMin is quantifiably fairer than Legacy on all instances by two established metrics of fairness, namely the Gini coefficient and the geometric mean (Table 3.6). For example, across instances excluding mass, LexiMin decreases the Gini coefficient, a standard measure of inequality, by between 5 and 16 percentage points (median: 12; negligible improvement on mass). Strikingly, the 16-point improvement in the Gini coefficient achieved by LexiMin on the instance obf (from 59% to 43%) approximately reflects the gap between relative income inequality in Namibia (59% in 2015) and the United States (42% in 2019) [285].

3.5 Discussion

As the recommendations made by citizens' assemblies increasingly impact public decision-making, so grows the urgency that selection algorithms distribute this power fairly across constituents.

Figure 3.3: How LexiMin 's output was used to select a panel via a live uniform lottery. (a) First, the output distribution was transformed into a uniform distribution over 1,000 panels, numbered 000 999. (b) The three digits determining the nal panel were drawn from lottery machines, making each panel observably selected with equal probability. (c) The personalized interface (screen-captured with (b)) shows each pool member the number of panels out of 1,000 they are on, allowing them to verify their own and others' selection probabilities. Screenshots credit: by for *.

We have made substantial progress on this front: the optimality of our algorithmic framework conclusively resolves the search for fair algorithms for a broad class of fairness measures, and the deployment of LexiMin puts an end to some pool members being virtually never selected in practice.

Beyond these immediate bene ts to fairness, the exchange of ideas we have initiated between practitioners and theorists presents continuing opportunities to improve panel selection in areas such as transparency. For example, for an assembly in Michigan, we assisted for * in selecting their panel via a live lottery in which participants could easily observe the probabilities with which each pool member was selected. This is an advance over the transparency possible with previous selection algorithms. We found that, in this instance, the output distribution of LexiMin could be transformed into a simple lottery without meaningful loss of fairness (3.3). Subsequent work by Flanigan et al. [31] developed general procedures and bounds for this transformation.

The Organisation for Economic Co-operation and Development (OECD) describes citizens' assemblies as part of a broader democratic movement to give citizens a more direct role in [. . .] shaping the public decisions that a ect them [122]. By bringing mathematical structure, increased fairness, and greater transparency to the practice of sortition, research in this area promises to put practical sortition on rmer foundations, and to promote citizens' assemblies' mission to give everyday people a greater voice.

3.6 Additional Methods and Empirical Analysis

3.6.1 Extended results for Figure 3.2, Table 3.1

Figure 3.4: Selection probabilities given by Legacy and LexiMin to the bottom 60% of pool members on the 4 instances that are not shown in Figure 3.2. Pool members are ordered across the x axis in order of increasing probability given by the respective algorithms. Shaded boxes denote the range of pool members with a selection probability given by Legacy that is lower than the minimum probability given by LexiMin. Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% confidence intervals (as described in Statistics in the Methods). Green dotted lines show the equalized probability $\frac{1}{n} = \frac{k}{n}$.

3.6.2 Individuals rarely selected by Legacy

The empirical results in Table 3.1 demonstrate that, in most instances, Legacy selects some pool members with very low probability. However, in any given citizens assembly, this does not automatically imply that these individuals had low probability of serving on the panel. Indeed, if such an individual would have been selected by Legacy with higher probability in most other pools that could have formed (as a result of other sets of agents being randomly invited alongside this individual), then the individual might still have had a substantial overall probability of serving on the citizens assembly.

In this section, we show how our data suggest that this is not the case, and that some people do in fact seem to have very low likelihood overall of ending up on the panel when Legacy is used. We make this case by demonstrating two separate points. First, we show that, across instances, Legacy tends to give very low selection probabilities to agents who have many features that are overrepresented in the observed pool relative to the quotas. Second, we discuss why it is likely that, across possible pools for the same citizens assembly, it is usually the same agents who have many overrepresented features. These two points, taken together, suggest that agents who have many overrepresented features in the pools we observe are rarely selected by Legacy overall.

Figure 3.5: Selection probabilities given by Legacy and LexiMin on all ten instances. Pool members are ordered across the x axis in order of increasing probability given by the respective algorithms. In contrast to Figure 3.2 and Figure 3.4, this graph shows the full range of selection probabilities (up to the 100th percentile). Shaded boxes denote the range of pool members with a selection probability given by Legacy that is lower than the minimum probability given by LexiMin. Legacy probabilities are estimated over 10,000 random panels and are indicated with 99% con dence intervals (as described in Statistics in the Methods). Green dotted lines show the equalized probability $\frac{1}{k}$: $\bullet = {}^{o}$.

Relationship between overrepresentation of features and selection probability. To measure the relationship between the level of overrepresentation of an agents features and that agents selection probability by Legacy, we rst construct a simple indicator called the ratio product, which measures the level of overrepresentation of a given agents set of features in the pool. The ratio product is composed of, for each of the features of an agent, the ratio between the fraction of this feature in the pool and the fraction of the quotas of the feature (speci cally, the mean of lower and upper quota) in the panel. That is, if we denote the set of pool members with a feature 5 by $\#_5$ and if we denote the lower and upper quotas of the feature by ℓ_5 and D_5, respectively,

Instance	Gini coefficient of LEGACY (lower is fairer)	Gini coefficient of LEXIMIN (lower is fairer)	Geometric mean of LEGACY (higher is fairer)	Geometric mean of LEXIMIN (higher is fairer)
sf(a)	51.2%	37.3%	6.5%	8.1%
sf(b)	59.6%	47.4%	3.5%	4.8%
sf(c)	57.0%	52.5%	8.3%	16.3%
sf(d)	59.3%	48.7%	3.5%	6.0%
sf(e)	64.4%	51.2%	2.2%	3.9%
cca	75.3%	67.8%	0.7%	3.5%
hd	64.5%	52.9%	3.1%	7.3%
mass	14.9%	14.8%	32.6%	32.7%
nexus	30.8%	25.4%	40.9%	44.2%
obf	58.9%	42.7%	3.7%	6.2%

Figure 3.6: Gini coe cient and geometric mean of probability allocations of both algorithms, for each instance. On every instance, Legacy has a lower Gini coe cient and a larger geometric mean. For computing the geometric mean, we slightly correct upward empirical selection probabilities of Legacy that are close to zero (as described in Statistics in the Methods).

Instance	Share selected by LEGACY with probability below LEXIMIN minimum selection probability
sf(a)	47.1%
sf(b)	44.8%
sf(c)	40.4%
sf(d)	38.6%
sf(e)	48.4%
cca	55.8%
hd	53.1%
mass	12.9%
nexus	33.6%
obf	51.1%

Figure 3.7: For each instance, the share of pool members selected with lower probability by Legacy than the minimum selection probability of LexiMin is shown. This corresponds to the width of the shaded boxes in Figures 3.2, 3.4 and 3.5.

then the ratio product of an agent i is de ned as:

$$\prod_{\text{features } f \text{ of } i} \frac{\#_f}{\frac{1}{n} \cdot D_f}$$

Given that the quotas are typically set in proportion to the share of the feature in the population, we say that agents with a high ratio product have many overrepresented features. Using this indicator, we nd that there is a clear negative relationship in all instances between the ratio product of an individual and their selection probability by Legacy (Figure 3.8). Most importantly, as this trend would suggest, we nd that the pool members with the largest ratio products consistently have some of the lowest selection probabilities.

Figure 3.8: Relationship between how overrepresented the features of an agent are and how likely they are to be chosen by the Leegacy algorithm. The level of overrepresentation is quanti ed as the ratio product (as described in Individuals rarely selected by Legacy in the Methods); agents further to the right are more overrepresented. Across instances, pool members with high ratio product are consistently selected with very low probabilities.

The same agents probably have many overrepresented features across most possible pools. Recall that we de ne an instance with respect to a single pool. However, this observed pool is only one among several hypothetical pools that could have resulted from the random process of sending out invitation letters. We de ne the ratio product of an agent with respect to a single instance and, therefore, a single observed pool. Then, if a di erent hypothetical pool (including that agent) had instead been drawn during the invitation process, the ratio product of the same agent with respect to that pool would probably be di erent, depending on which constituents were invited to join the pool alongside them. As the quotas and the target panel size k would be the same for all these hypothetical instances, the di erences in ratio product