# Networks, Polarization, and Voting: Models for Information Aggregation in Social Settings

by

Yan Jin

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Social and Engineering Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

© Massachusetts Institute of Technology 2023. All rights reserved.

Author .....
Institute for Data, Systems, and Society
Feb 5, 2023

Certified by .....
Elchanan Mossel
Professor of Mathematics
Thesis Supervisor

Accepted by .....
Fotini Christia
Chair, Social and Engineering Systems Doctoral Programs

# Networks, Polarization, and Voting: Models for Information Aggregation in Social Settings

by

Yan Jin

Submitted to the Institute for Data, Systems, and Society
on Feb 5, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Social and Engineering Systems

## Abstract

Social networks, voting, and polarization all fall into the realm of the instrument, process, and consequence of information aggregation in social settings. These are both classical topics that have motivated studies from various disciplines, and active areas in need of new models as novel phenomenon, demand, and proposals continue to emerge in recent years. In this thesis, we study three models for social information aggregation inspired by these three topics respectively.

In the first chapter, we consider how to detect corruption when each network nodes' true identities are only locally known. In this model, each vertex reports about the types - truthful or corrupt - of its neighbors, where truthful nodes report the true types and corrupt nodes report adversarially. We show that detecting corruption in this model yields linear-time algorithm while the minimal number of nodes the corrupt party needs to control in order to hide all corruption is hard to approximate to any multiplicative factor, assuming the Small Set Expansion Hypothesis.

In the second chapter, we propose a geometric opinion dynamic model where a strong form of polarization in high-dimension emerges: public opinions not only radicalize on each issue, but also correlate across issues. We demonstrate that this type of polarization could arise as an unintended byproduct of influencers' natural effort to promote a product or an idea. We analyze this mechanism with one or more influencers, sending messages strategically, heuristically, or randomly, and examine the computational aspects of optimal influencing strategy and its effect on polarization.

The third chapter considers whether distributed election procedure can aggregate to good social choice outcomes when voters delegate strategically. We model liquid democracy as a game where voters with continuous-valued preference peaks choose between delegation and learning about policies at a cost and voting directly. We derive the pure-strategy coalition-proof Nash equilibrium and show that equilibrium delegation network varies with learning cost. When cost is low, all voters delegate to the median is a cpNE. As learning cost increases, new forms of cpNE emerge, where extreme voters delegate inward and moderate voters delegate outward to the nearest incentized voters.

Thesis Supervisor: Elchanan Mossel
Title: Professor of Mathematics

# Acknowledgments

First of all, I would like to thank Elchanan for being my advisor, without whose patience, support and guidance this thesis would have not been possible. I really appreciate his guidance in my early years of advisor search, all the engaging discussions that we have verbally or on board when we work together, and his understanding and support that helped me through unexpected illness. He always firmly encourages me to work on what I am interested in, raises questions that I missed, which further carves out the problem at hand, and points out arguments that send me steps closer to the proofs that I tried to go after. I have learned and feel there is much more to learn from him, from algorithms, social choice theory, probability, to the dilemma in voting security, and randomization in games.

I am deeply grateful to Prof. Guy Bresler and Prof. Ali Jadbabaie for being on my thesis committee and all the generous help they offer. I thank Guy for teaching me graphical models and enhancing my interests in graphs in his graduate course. I thank Ali, who has also been my academic advisor since day one at MIT, for many valuable advice, help, for many acknowledgment of my work, and connecting me to resources and opportunities throughout my progress at the SES program.

Next, I would like to thank my co-authors, Govind Ramnarayan, Jan Hązła, and Chin-Chia Hsu. These collaborations have made the research process much more enjoyable, widened my abilities as a researcher and expanded the possibility of the projects. I have learned more about complexity theory, stochastic processes and game theory from them. I am especially indebted to Govind for giving me the first bit of recognition and affirmation for my ideas in my graduate research. I thank him for teaching me complexity theory, mentoring me in our first project spontaneously, and continuing to be a mentor, collaborator, and friend. I will miss the many discussions we have at the white board on the sixth floor of STATA and the presentation slides we made starring stick men.

My PhD years have been quite an adventure at times. I am deeply grateful to all the family members and friends that escorted me through these challenges. I would

like to thank my parents Kai Jin and Caiyun Sun for their tireless support. Though they are never convinced, growing up by their side has been my happiest time, which will continue being my greatest treasure. I thank Zihan Xie for always being willing to back me up in every way possible, beyond ways that I could think of, in my most difficult time in 2021. I thank Ruyuan Liu for always having an answer when I have a question, and trying his very best to understand me. I thank my uncle, Dr. Jian Sun, and his family for hosting me in Shanghai and connecting me to medical resources patiently through many trials and errors together with me in 2021. I thank my grandparents to whom I owe many positive sides of my personality, many happy memories, and warm visits.

I thank Qi Yang, Pu Zhao, Penny Chen, Sherry Muyuan He, Qinyun Zhang, Yifei Wang, Amir Tohidi, Liang Chang, Yunzong Xu, Hanwei Li, Tessa Ganser, Xiaofei Ge, Jinghui Liu, Qisheng Li, Qing Zhang, Mengdie Wang, Siyuan Sheng, Miaorong Wang, Qingyu Zhu, Yu Zhao, Yuan Yuan, Paolo Bertolotti, Hongbo Fang, Xinyi Wu, Manon Revel, Yue Zhang, among many others, for their friendship, comradeship, wise advice, and generous help at various points in my graduate years.

I am grateful to Prof. Susan Silbey for many advice and kind guidance in my early years of graduate school, Prof. Iyad Rahwan for mentoring, valuing and supporting my work in part of my graduate study, Prof. Thomas Malone for suggesting the topic of liquid democracy to me, followed by many inspiring conversations, and Prof. Yannis Paschalidis for teaching me optimization, a subject that I deeply enjoyed.

It would be a miss to not thank my undergraduate advisor, Prof. David Shuman who opened the door for me to the world of research and offered all his guidance in my first research project and the decision of pursuing graduate study. I am deeply indebted to David for teaching me signal processing and working along my side through learning about the problem to writing our first paper. This process enabled me to grasp what research is; his generous guidance and encouragement granted me a precious positive first impression of this new acquaintance, and infused me with interest and confidence. I am also indebted to Prof. Andrew Beveridge, Prof. Tom Halverson, Prof. Erik Larson, Prof. Mahnaz Kousha, Prof. Wayne Roberts, and my unrivalled

high school math teacher, Mr. Juguo Sun, for inspiring my interests and helping me build firm foundations in math and social sciences.

Two women that I would like to thank are Dr. Yina Gu at Amazon, under whose mentorship I undertook a remote internship in the summer of 2020, during the pandemic, and Dr. Karene Chu, with whom I helped developed materials for 6.419x, an online MicroMaster course on statistics and computation in 2019. Their enthusiastic attitudes towards the work that we had together and their genuine and compassionate personalities were empowering, and working with them provided an energizing experience in some particular difficult times.

Finally, I would like to dedicate special thanks to Dr. Huijuan Wang, Dr. Shuna Gong, and Dr. Zhengqing Zhao, for helping me recover.

# Contents

**1 Introduction** **15**

1.1 Overview of Chapters . . . . . 16

**2 Being Corrupt Requires Being Clever, But Detecting Corruption Doesn't** **19**

2.1 Introduction . . . . . 19

2.1.1 Corruption Detection and Problem Set-up . . . . . 19

2.1.2 Our Results . . . . . 22

2.1.3 Related Work . . . . . 26

2.1.4 Comparison to Corruption in Practice . . . . . 27

2.2 Preliminaries . . . . . 28

2.2.1 General Preliminaries . . . . . 28

2.2.2 Preliminaries for Corruption Detection on Networks . . . . . 31

2.3 Proofs of Theorems 1, 2, and 3 . . . . . 33

2.3.1 2-Approximation by Vertex Separation . . . . . 33

2.3.2 SSE-Hardness of Approximation for $m(G)$ . . . . . 37

2.3.3 An $O(\log|V|)$ Approximation Algorithm for $m(G)$ . . . . . 41

2.4 Directed Graphs . . . . . 43

2.5 Finding an Arbitrary Fraction of Good Nodes on a Graph . . . . . 46

**3 A Geometric Model for Opinion Polarization** **55**

3.1 Introduction . . . . . 55

3.1.1 Model definition . . . . . 58

3.1.2 Example . . . 59
3.1.3 Outline of our results . . . 60
3.1.4 Design choices . . . 63
3.1.5 Other variants . . . 67
3.2 Related works . . . 68
3.3 Asymptotic scenario: random interventions polarize opinions . . . 71
3.3.1 Notation and main ingredients . . . 73
3.3.2 Proof of Claim 5 for $d = 2$ . . . 75
3.3.3 Proof of Claim 5 for $d \geq 3$ . . . 76
3.3.4 Proof of Claim 6 . . . 77
3.4 Asymptotic scenario: finding densest hemisphere . . . 79
3.4.1 Equivalence of optimal strategy to finding densest hemisphere 79
3.4.2 Computational equivalence to learning halfspaces . . . 83
3.5 Short-term scenario: polarization as externality . . . 86
3.5.1 One intervention, two agents: polarization costs . . . 87
3.5.2 One intervention, many agents: finding the densest spherical cap 89
3.6 Asymptotic effects of two dueling influencers: two randomized interventions polarize . . . 91
3.6.1 Proofs of Theorem 12 and Proposition 4 . . . 93
3.6.2 Proof of Theorem 13 . . . 97

**4 A Median Voter Theory for Liquid Democracy 101**
4.1 Introduction . . . 101
4.1.1 Liquid Democracy and Problem Set-up . . . 101
4.1.2 Main Results . . . 104
4.1.3 Related Work . . . 105
4.2 Model . . . 106
4.3 Preliminary . . . 108
4.3.1 Terminologies for liquid democracy game . . . 110
4.3.2 Discussion of model constructs . . . 112

4.4 Results on Pure-strategy Coalition-proof Nash Equilibria . . . . . 114
4.4.1 Incentive structure . . . . . 114
4.4.2 Results on low learning cost . . . . . 118
4.4.3 Intermediate learning cost . . . . . 122
4.4.4 High learning cost . . . . . 125
4.5 Discussion and future directions . . . . . 127

**5 Conclusion** **129**

**A Supplementary Material for Chapter 2** **131**
A.1 Omitted Results . . . . . 131

**B Supplementary Material for Chapter 3** **135**
B.1 Proof of Claim 7 . . . . . 135
B.2 Example with two advertisers . . . . . 136
B.3 Proof of Proposition 2 . . . . . 137
B.4 Proof of Proposition 3 . . . . . 141

# List of Figures

3-1 Graphical illustration of the example discussed in Section 3.1.2. Since we are working in $d = 4$, we illustrate the first three dimensions as spatial positions and the fourth dimension with a color scale. Initially the opinions are uniformly distributed on the sphere, with the fourth dimension equal to 0 (no opinion) everywhere. Consecutive applications of the intervention $v = (\sqrt{7}/4, 0, 0, 3/4)$ in $\mathbb{R}^4$ result in polarization both in spatial dimensions and in the color scale. . . . . 61

3-2 On the left an illustration of the vectors and angles in the proof of Claim 5. On the right an illustration for the proof of Claim 6. . . . . 75

3-3 Illustration of the polarization cost as a function of the initial correlation $c$. The dashed line is the initial correlation included as a reference point. The red and blue lines are correlations after applying two- and one-agent interventions respectively. The green line shows the polarization cost $c_{\text{two}} - c_{\text{one}}$. . . . . 89

3-4 The after-intervention opinions of both agents $\widetilde{u}_{i,d}$ as functions of initial correlation $c$. The red line represents the opinion of either agent after applying the two-agent intervention. The blue line is the opinion of the second agent after the one-agent intervention. For reference, the dashed line (1/3) shows the opinion of the first agent in the one-agent intervention (which does not depend on $c$). The grey area represents the range of thresholds $T$ where it is preferable for the influencer to apply the one-agent intervention. . . . . 89

3-5 Projection onto the subspace $V = \text{span}\{v, v'\}$. . . . . 95

3-6 The graph of the "pull function" $\alpha - f(\alpha)$ in case $\eta = 1$. . . . . 99

B-1 Illustration of the process described in Appendix B.2. This time we need to visualize five dimensions. This is done with spatial positions for the first three dimensions $j = 1, 2, 3$ and two different color scales for $j = 4, 5$. Accordingly, two figures are displayed for each time step $t = 5, 9, 13$. In each pair of figures the points in the left figure have the same spatial positions as in the right figure and the colors illustrate dimensions $j = 4$ (on the left) and $j = 5$ (on the right). . . . . 138

B-2 The projection of one-agent (left) and two-agent (right) interventions onto the first two dimensions. . . . . 139

# Chapter 1

# Introduction

Social networks, voting, and polarization all fall into the realm of the instrument, process, and consequence of information aggregation in social settings. These are both classical topics that have motivated studies from various disciplines, and also active areas in need of new models as novel phenomenon, demand and proposals continue to emerge in recent years.

Survey on public opinions and data analysis from online social media have shown nontrivial correlation on the public's opinions on *a priori* unrelated topics. Remote elections during the COVID-19 pandemic, including the election for US president in 2020, reignite interest in the trustworthiness and effectiveness of online voting, novel alternatives, and voting methods in general. Large online social networks including long-standing ones such as Facebook and Twitter, and new virtual communities have become not only more prevalent, but also more influential and relied on, catalyzed and demanded by the pandemic and largely remain so in the post-pandemic world where social distancing measures prolong. As one of the main information exchange channel in an increasingly online social world, online social media's reliability becomes crucial more than ever.

In this thesis, we study three models for social information aggregation inspired by these three topics respectively. These include a strong form of polarization where public opinions not only radicalize on each issue, but also correlate on different dimensions, a novel election scheme proposal called liquid democracy that allows transitive

delegation to other voters, and large social networks containing malicious agents with an additional local audit capability. In particular, we will consider how to detect corruption when each network nodes' true identities are only locally known, study whether distributed election procedure can aggregate to good social choice outcomes when agents delegate strategically, and propose a new model where issue alignment, the strong form of polarization can naturally emerge as a consequence of agents responding to persuading information outlet. In all of these models, information is local, but the impact of the procedure is global.

In the next sections, we give a general summary for the main results of each chapter, before proceeding to elaborate on each in independent chapters.

## 1.1 Overview of Chapters

**Chapter 2: Corruption detection on networks: a computational complexity lens.**

In Chapter 2, we consider a variation of the problem of corruption detection on networks posed by Alon, Mossel, and Pemantle '15. In this model, each vertex of a graph can be either truthful or corrupt. Each vertex reports about the types (truthful or corrupt) of all its neighbors to a central agency, where truthful nodes report the true types they see and corrupt nodes report adversarially. The central agency aggregates these reports and attempts to find a single truthful node. Inspired by real auditing networks, we pose our problem for arbitrary graphs and consider corruption through a computational lens. We identify a key combinatorial parameter of the graph $m(G)$, which is the minimal number of corrupted agents needed to prevent the central agency from identifying a single corrupt node. We give an efficient (in fact, linear time) algorithm for the central agency to identify a truthful node that is successful whenever the number of corrupt nodes is less than $m(G)/2$. On the other hand, we prove that for any constant $\alpha > 1$, it is NP-hard to find a subset of nodes S in G such that corrupting $S$ prevents the central agency from finding one truthful node and $|S| \leq \alpha m(G)$, assuming the Small Set Expansion Hypothesis (Raghavendra

and Steurer, STOC '10). We conclude that being corrupt requires being clever, while detecting corruption does not.

Our main technical insight is a relation between the minimum number of corrupt nodes required to hide all truthful nodes and a certain notion of vertex separability for the underlying graph. Additionally, this insight lets us design an efficient algorithm for a corrupt party to decide which graphs require the fewest corrupted nodes, up to a multiplicative factor of $O(\log n)$.

This chapter is based on the paper "Being Corrupt Requires Being Clever, But Detecting Corruption Doesn't", which is joint work with Elchanan Mossel and Govind Ramnarayan. This appeared in ITCS 2019 [52].

**Chapter 3: The curse of dimensionality in opinion dynamics**

The second line of work, contained in Chapter 3, proposes a model that attempts to capture the multi-dimensionality of public opinion in real life. At its core are two key "axioms": (i) population's opinion structure consists of "logically" or "philosophically" relatively independent dimensions , and (ii) these dimensions can influence each other in interactions of with others' opinions, and are used in such ways in interventions such as campaigns.

To capture these features, we introduce a simple, geometric model of opinion polarization. It is a model of political persuasion, as well as marketing and advertising, utilizing social values. It focuses on the interplay between different topics and persuasion efforts. We demonstrate that societal opinion polarization often arises as an unintended byproduct of influencers attempting to promote a product or idea. We discuss and define a strong form of polarization where the end opinion structure demonstrates large correlation across independent topics. We discuss a number of mechanisms for the emergence of this polarization involving one or more influencers, sending messages strategically, heuristically, or randomly. We also examine some computational aspects of choosing the most effective means of influencing agents, and the effects of those strategic considerations on polarization.

This chapter is based on the paper "A Geometric Model of Opinion Polarizatiohn",

which is joint work with Jan Hązła, and Elchanan Mossel, and Govind Ramnarayan. A preprint is available at [48] and will appear in *Mathematics of Operations Research* in 2023.

**Chapter 4: Voting and strategy in a delegative democracy**

In Chapter 4, we propose a game-theoretic model of liquid democracy where a population has heterogeneous and continuous ideological preferences (e.g., a continuous left to right opinion position). In this model, a finite set of voters face unknown policy candidates generated from a common prior. Each voter either invests effort to learn the policies and votes directly or delegates to another voter, leveraging ready information about other voters' preferences. We derive pure-strategy coalition-proof Nash equilibria in this game and show a relation between learning cost and the form of equilibrium delegation networks.

In particular, with low learning cost, echoing the classical median voter theory, all voters delegating to median is a coalition-proof NE. However, as learning cost increases, a region of disincentivised voters forms in the middle of the political spectrum. New structure of coalition-proof NE emerges where extreme voters delegate inward and moderate voters delegate outward to the most moderate voter who is still incentivised to learn. Non-trivial delegation to opposite political spectrum occurs in order to rule out more unfavored coalitions. This exploratory analysis sheds light on how rational agents may decide to delegate in a population with varying political stances under a delegative democratic scheme.

This chapter is based on the paper "A Median Voter Theory for Liquid Democracy", which is joint work with Chin-Chia Hsu and Elchanan Mossel.

# Chapter 2

# Being Corrupt Requires Being Clever, But Detecting Corruption Doesn't

## 2.1 Introduction

### 2.1.1 Corruption Detection and Problem Set-up

We study the problem of identifying truthful nodes in networks, in the model of *corruption detection on networks* posed by Alon, Mossel, and Pemantle [1]. In this model, we have a network represented by a (possibly directed) graph. Nodes can be *truthful* or *corrupt*. Each node audits its outgoing neighbors to see whether they are truthful or corrupt, and sends reports of their identities to a central agency. The central agent, who is not part of the graph, aggregates the reports and uses them to identify truthful and corrupt nodes. Truthful nodes report truthfully (and correctly) on their neighbors, while corrupt nodes have no such restriction: they can assign arbitrary reports to their neighbors, regardless of whether their neighbors are truthful or corrupt, and coordinate their efforts with each other to prevent the central agency from gathering useful information.

In [1], the authors consider the problem of recovering the identities of almost all nodes in a network in the presence of many corrupt nodes; specifically, when the fraction of corrupt nodes can be very close to 1/2. They call this the *corruption*

*detection* problem. They show that the central agency can recover the identity of most nodes correctly even in certain bounded-degree graphs, as long as the underlying graph is a sufficiently good expander. The required expansion properties are known to hold for a random graph or Ramanujan graph of sufficiently large (but constant) degree, which yields undirected graphs that are amenable to corruption detection. Furthermore, they show that some level of expansion is necessary for identifying truthful nodes, by demonstrating that the corrupt nodes can stop the central agency from identifying any truthful node when the graph is a very bad expander (e.g. a cycle), even if the corrupt nodes only make up 0.01 fraction of the network.

This establishes that very good expanders are very good for corruption detection, and very bad expanders can be very bad for corruption detection. We note that this begs the question of how effective graphs that do not fall in either of these categories are for corruption detection. In the setting of [1], we could ask the following: given an *arbitrary* undirected graph, what is the smallest number of corrupt nodes that can prevent the identification of almost all nodes? When there are fewer than this number, can the central agency *efficiently* identify almost all nodes correctly? Alon, Mossel, and Pemantle study these questions for the special cases of highly expanding graphs and poorly expanding graphs, but do not address general graphs.

Additionally, [1] considers corruption detection when the corrupt agencies can choose their locations and collude arbitrarily, with no bound on their computational complexity. This is perhaps overly pessimistic: after all, it is highly unlikely that corrupt agencies can solve NP-hard problems efficiently and if they can, thwarting their covert operations is unlikely to stop their world domination. We suggest a model that takes into account computational considerations, by factoring in the computation time required to select the nodes in a graph that a corrupt party chooses to control. This yields the following question from the viewpoint of a corrupt party: given a graph, can a corrupt party compute the smallest set of nodes it needs to corrupt *in polynomial time*?

In addition to being natural from a mathematical standpoint, these questions are also well-motivated socially. It would be naïve to assert that we can weed out

corruption in the real world by simply designing auditing networks that are expanders. Rather, these networks may already be formed, and infeasible to change in a drastic way. Given this, we are less concerned with finding certain graphs that are good for corruption detection, but rather discerning how good *existing* graphs are; specifically, how many corrupt nodes they can tolerate. In particular, since the network structure could be out of the control of the central agency, algorithms for the central agency to detect corruption on arbitrary graphs seem particularly important.

It is also useful for the *corrupt* agency to have an algorithm with guarantees for any graph. Consider the following example of a corruption detection problem from the viewpoint of a corrupt organization. Country A wants to influence policy in country B, and wants to figure out the most efficient way to place corrupted nodes within country B to make this happen. However, if the central government of B can confidently identify truthful nodes, they can weight those nodes' opinions more highly, and thwart country A's plans. Hence, the question country A wants to solve is the following: given the graph of country B, can country A compute the optimal placement of corrupt nodes to prevent country B from finding truthful nodes? We note that in this question, too, the graph of country B is fixed, and hence, country A would like to have an algorithm that takes as input *any* graph and computes the optimal way to place corrupt nodes in order to hide all the truthful nodes.

We study the questions above for a variant of the corruption detection problem in [1], in which the goal of the central agency is to find a single truthful node. While this goal is less ambitious than the goal of identifying almost all the nodes, we think it is a very natural question in the context of corruption. For one, if the central agency can find a single truthful node, they can use the trusted reports from that node to identify more truthful and corrupt nodes that it might be connected to. The central agency may additionally weight the opinions of the truthful nodes more when making policy decisions (as alluded to in the example above), and can also incentivize truthfulness by rewarding truthful nodes that it finds and giving them more influence in future networks if possible (by increasing their out-degrees). Moreover, our proofs and results extend to finding larger number of truthful nodes as we discuss below.

Our results stem from a tie between the problem of finding a single truthful node in a graph and a measure of vertex separability of the graph. This tie not only yields an efficient and relatively effective algorithm for the central agency to find a truthful node, but also allows us to relate corrupt party's strategy to the problem of finding a good vertex separator for the graph. Hence, by analyzing the purely graph-theoretic problem of finding a good vertex separator, we can characterize the difficulty of finding a good set of nodes to corrupt. Similar notions of vertex separability have been studied previously (e.g. [63, 74, 14]), and we prove NP-hardness for the notion relevant to us assuming the Small Set Expansion Hypothesis (SSEH). The *Small Set Expansion Hypothesis* is a hypothesis posed by Raghavendra and Steurer [83] that is closely related to the famous Unique Games Conjecture of Khot [57]. In fact, [83] shows that the SSEH implies the Unique Games Conjecture. The SSEH yields hardness results that are not known to follow directly from the UGC, especially for graph problems like sparsest cut and treewidth ([84] and [4] respectively), among others.

### 2.1.2 Our Results

We now outline our results more formally. We analyze the variant of corruption detection where the central agency's goal is to find a single truthful node. First, we study how effectively the central agency can identify a truthful node on an arbitrary graph, given a set of reports. Given an undirected graph[1] $G$, we let $m(G)$ denote the minimal number of corrupted nodes required to stop the central agency from finding a truthful node, where the minimum is taken over all strategies of the corrupt party (not just computationally bounded ones). We informally call $m(G)$ the "critical" number of corrupt nodes for a graph $G$. Then, we show the following:

**Theorem 1.** *Fix a graph $G$ and suppose that the corrupt party has a budget $b \leq m(G)/2$. Then the central agency can identify a truthful node, regardless of the strategy of the corrupt party, and without knowledge of either $m(G)$ or $b$. Furthermore, the central agency's algorithm runs in linear time (in the number of edges in the graph*

[1] Unless explicitly specified, all graphs are undirected by default.

*G).*

Next, we consider the question from the viewpoint of the corrupt party: can the corrupt party efficiently compute the most economical way to allocate nodes to prevent the central agency from finding a truthful node? Concretely, we focus on a natural decision version of the question: given a graph $G$ and a upper bound on the number of possible corrupted nodes $k$, can the corrupt party prevent the central agency from finding a truthful node?

We actually focus on an easier question: can the corrupt party accurately compute $m(G)$, the minimum number of nodes that they need to control to prevent the central agency from finding a truthful node? Not only do we give evidence that computing $m(G)$ exactly is computationally hard, but we also provide evidence that $m(G)$ is hard to approximate. Specifically, we show that approximating $m(G)$ to any constant factor is NP-hard under the Small Set Expansion Hypothesis (SSEH); or in other words, that it is SSE-hard.

**Theorem 2.** *For every $\beta > 1$, there is a constant $\epsilon > 0$ such that the following is true. Given a graph $G = (V, E)$, it is SSE-hard to distinguish between the case where $m(G) \leq \epsilon \cdot |V|$ and $m(G) \geq \beta \cdot \epsilon \cdot |V|$. Or in other words, the problem of approximating the critical number of corrupt nodes for a graph to within any constant factor is SSE-hard.*

This Theorem immediately implies the following Corollary 1.

**Corollary 1.** *Assume the SSE Hypothesis and that $P \neq NP$. Fix any $\beta > 1$. There does not exist a polynomial-time algorithm that takes as input an arbitrary graph $G = (V, E)$ and outputs a set of nodes $S$ with size $|S| \leq O(\beta \cdot m(G))$, such that corrupting $S$ prevents the central agency from finding a truthful node.*

We note that in Corollary 1, the bad party's input is only the graph $G$: specifically, they do not have knowledge about the value of $m(G)$.

Our proof for Theorem 2 is similar to the proof of Austrin, Pitassi, and Wu [4] for the SSE-hardness of approximating treewidth. This is not a coincidence: in fact,

the "soundness" in their reduction involves proving that their graph does not have a good 1/2 vertex separator, where the notion of vertex separability (from [23]) is very related to the version we use to categorize the problem of hiding a truthful vertex. We give the proof of Theorem 2 in Section 2.3.2.

However, if one allows for an approximation factor of $O(\log|V|)$, then $m(G)$ can be approximated efficiently. Furthermore, this yields an approximation algorithm that the corrupt party can use to find a placement that hinders detection of a truthful node.

**Theorem 3.** *There is a polynomial-time algorithm that takes as input a graph $G = (V, E)$ and outputs a set of nodes $S$ with size $|S| \leq O(\log|V| \cdot m(G))$, such that corrupting $S$ prevents the central agency from finding a truthful node.*

The proof of Theorem 3, given in Section 2.3.2, uses a bi-criterion approximation algorithm for the $k$-vertex separator problem given by [63]. As alluded to in Section 2.1.1, Theorems 2 and 3 both rely on an approximate characterization of $m(G)$ in terms of a measure of vertex separability of the graph $G$, which we give in Section 2.3.

Additionally, we note that we can adapt Theorems 1 and 2 (as well as Corollary 1) to a more general setting, where the central agency wants to recover some arbitrary number of truthful nodes, where the number of nodes can be proportional to the size of the graph. We describe how to modify our proofs to match this more general setting in Section 2.5.

Together, Theorems 1 and 2 uncover a surprisingly positive result for corruption detection: it is computationally easy for the central agency to find a truthful node when the number of corrupted nodes is only somewhat smaller than the "critical" number for the underlying graph, but it is in general computationally hard for the corrupt party to hide the truthful nodes even when they have a budget that far exceeds the "critical" number for the graph.

**Results for Directed Graphs** As noted in [1], it is unlikely that real-world auditing networks are undirected. For example, it is likely that the FBI has the authority

to audit the Cambridge police department, but it is also likely that the reverse is untrue. Therefore, we would like the central agency to be able to find truthful nodes in directed graphs in addition to undirected graphs. We notice that the algorithm we give in Theorem 1 extends naturally to directed graphs.

**Theorem 4.** *Fix a directed graph $D$ and suppose that the corrupt party has a budget $b \leq m(D)/2$. Then the central agency can identify a truthful node, regardless of the strategy of the corrupt party, and without the knowledge of either $m(D)$ or $b$. Furthermore, the central agency's algorithm runs in linear time.*

The proof of Theorem 4 is similar to the proof of Theorem 1, and effectively relates the problem of finding a truthful node on directed graphs to a similar notion of vertex separability, suitably generalized to directed graphs.

**Results for Finding An Arbitrary Number of Good Nodes** In fact, the problem of finding one good node is just a special case of finding an arbitrary number of good nodes, $g$, on the graph $G$. We define $m(G, g)$ as the minimal number of bad nodes required to prevent the identification of $g$ good nodes on the graph $G$. We relate it to an analogous vertex separation notion, and prove the following two theorems, which are extensions of Theorems 1 and 2 to this setting.

**Theorem 5.** *Fix a graph $G$ and the number of good nodes to recover, $g$. Suppose that the corrupt party has a budget $b \leq m(G, g)/2$. If $g < |V| - 2b$, then the central agency can identify $g$ truthful nodes, regardless of the strategy of the corrupt party, and without knowledge either of $m(G, g)$ or $b$. Furthermore, the central agency's algorithm runs in linear time.*

**Theorem 6.** *For every $\beta > 1$ and every $0 < \delta < 1$ , there is a constant $\epsilon > 0$ such that the following is true. Given a graph $G = (V, E)$, it is SSE-hard to distinguish between the case where $m(G, \delta|V|) \leq \epsilon \cdot |V|$ and $m(G, \delta|V|) \geq \beta \cdot \epsilon \cdot |V|$. Or in other words, the problem of approximating the critical number of corrupt nodes such that it is impossible to find $\delta|V|$ good nodes within any constant factor is SSE-hard.*

The proof of Theorem 6 is similar to the proof of Theorem 1, and the hardness of approximation proof also relies on the same graph reduction and SSE conjecture. Proofs are presented in Section 2.5.

### 2.1.3 Related Work

The model of corruptions posed by [1] is identical to a model first suggested by Perparata, Metze, and Chien [81], who introduced the model in the context of detecting failed components in digital systems. This work (as well as many follow-ups, e.g. [55, 60]) looked at the problem of characterizing which networks can detect a certain number of corrupted nodes. Xu and Huang [96] give necessary and sufficient conditions for identifying a single corrupted node in a graph, although their characterization is not algorithmically efficient. There are many other works on variants of this problem (e.g. [92, 26]), including recovering node identities with one-sided or two-sided error probabilities in the local reports [66] and adaptively finding truthful nodes [45].

We note that our model of a computationally bounded corrupt party and our stipulation that the graph is fixed ahead of time rather than designed by the central agency, which are our main contributions to the model, seem more naturally motivated in the setting of corruptions than in the setting of designing digital systems. Even the question of identifying a single truthful node could be viewed as more naturally motivated in the setting of corruptions than in the setting of diagnosing systems. We believe there are likely more interesting theoretical questions to be discovered by approaching the PMC model through a corruptions lens.

The identifiability of a single node in the corruptions setting was studied in a recent paper of Mukwembi and Mukwembi [69]. They give a linear time greedy algorithm to recover the identify of a single node in many graphs, *provided that corrupt nodes always report other corrupt nodes as truthful.* Furthermore, this assumption allows them to reduce identifying all nodes to identifying a single node. They argue that such an assumption is natural in the context of corruptions, where corrupt nodes are selfishly incentivized not to out each other. However, in our setting, corrupt nodes

can not only betray each other, but are in fact incentivized to do so for the good of the overarching goal of the corrupt party (to prevent the central agency from identifying a truthful node). Given [69], it is not a surprise that the near-optimal strategies we describe for the corrupt party in this paper crucially rely on the fact that the nodes can report each other as corrupt.

Our problem of choosing the best subset of nodes to corrupt bears intriguing similarities to the problem of influence maximization studied by [56], where the goal is to find an optimal set of nodes to target in order to maximize the adoption of a certain technology or product. It is an interesting question to see if there are further similarities between these two areas. Additionally, social scientists have studied corruption extensively (e.g.[35], [71]), though to the best of our knowledge they have not studied it in the graph-theoretic way that we do in this paper.

### 2.1.4 Comparison to Corruption in Practice

Finally, we must address the elephant in the room. Despite our theoretical results, corruption *is* prevalent in many real-world networks, and yet in many scenarios it is not easy to pinpoint even a single truthful node. One reason for that is that some of assumptions do not seem to hold in some real world networks. For example, we assume that audits from the truthful nodes are not only non-malicious, but also perfectly reliable. In practice this assumption is unlikely to be true: many truthful nodes could be non-malicious but simply unable to audit their neighbors accurately. Further assumptions that may not hold in some scenarios include the notion of a central agency that is both uncorrupted and has access to reports from every agency, and possibly even the assumption that the number of corrupt nodes is less than $|V|/2$. In addition, networks $G$ may have very low critical numbers $m(G)$ in practice. For example, there could be a triangle (named, "President", "Congress" and "Houses") that is all corrupt and cannot be audited by any agent outside the triangle. It is thus plausible that a corrupt party could use the structure of realistic auditing networks for their corruption strategy to overcome our worst-case hardness result.

While this points to some shortcomings of our model, it also points out ways

to change policy that would potentially bring the real world closer to our idealistic scenario, where a corrupt party has a much more difficult computational task than the central agency. For example, we can speculate that perhaps information should be gathered by a transparent centralized agency, that significant resources should go into ensuring that the centralized agency is not corrupt, and that networks ought to have good auditing structure (without important agencies that can be audited by very few nodes).

## 2.2 Preliminaries

### 2.2.1 General Preliminaries

We denote undirected graphs by $G = (V, E)$, where $V$ is the vertex set of the graph and $E$ is the edge set. We denote directed graphs by $D = (V, E_D)$. When the underlying graph is clear, we may drop the subscripts. Given a vertex $u$ in an undirected graph $G$, we let $\mathcal{N}(u)$ denote the *neighborhood* (set of neighbors) of the vertex in $G$. Similarly, given a vertex $u$ in a directed graph $D$, let $\mathcal{N}(u)$ denote the set of *outgoing* neighbors of $u$: that is, vertices $v \in V$ such that $(u, v) \in E_D$.

#### Vertex Separator

**Definition 1.** ***(k-vertex separator)****([74],[14]) For any $k \geq 0$, we say a subset of vertices $U \subseteq V$ is k-vertex separator of a graph $G$, if after removing $U$ and incident edges, the remaining graph forms a union of connected components, each of size at most $k$.*

*Furthermore, let*

$$S_G(k) = \min\left(|U|\text{: } U \textit{ is a } k\textit{-vertex separator of } G\right)$$

*denote the size of the minimal $k$-vertex separator of graph $G$.*

**Small Set Expansion Hypothesis**

In this section we define the Small Set Expansion (SSE) Hypothesis introduced in [83]. Let $G = (V, E)$ be an undirected $d$-regular graph.

**Definition 2** (Normalized edge expansion)**.** *For a set $S \subseteq V$ of vertices, denote $\Phi_G(S)$ as the normalized edge expansion of $S$,*

$$\Phi_G(S) = \frac{|E(S, V\backslash S)|}{d|S|},$$

*where $|E(S, V\backslash S)|$ is the number of edges between $S$ and $V\backslash S$.*

*The Small Set Expansion Problem* with parameters $\eta$ and $\delta$, denoted $\mathrm{SSE}(\eta, \delta)$, asks whether $G$ has a small set $S$ which does not expand or all small sets of $G$ are highly expanding.

**Definition 3** ($(\mathrm{SSE}(\eta, \delta))$)**.** *Given a regular graph $G = (V, E)$, distinguish between the following two cases:*

- ***Yes*** *There is a set of vertices $S \subseteq V$ with $S = \delta|V|$ and $\Phi_G(S) \leq \eta$*
- ***No*** *For every set of vertices $S \subseteq V$ with $S = \delta|V|$ it holds that $\Phi_G(S) \geq 1 - \eta$*

The *Small Set Expansion Hypothesis* is the conjecture that deciding $\mathrm{SSE}(\eta, \delta)$ is NP-hard.

**Conjecture 1** (Small Set Expansion Hypothesis [83])**.** *For every $\eta > 0$, there is a $\delta > 0$ such that SSE($\eta, \delta$) is NP-hard.*

We say that a problem is *SSE-hard* if it is at least as hard to solve as the SSE problem. The form of conjecture most relevant to our proof is the following "stronger" form of the SSE Hypothesis. [84] showed that the SSE-problem can be reduced to a quantitatively stronger form of itself. In order to state this version, we first need to define the *Gaussian noise stability*.

**Definition 4.** *(Gaussian Noise Stability) Let $\rho \in [-1,1]$. Define $\Gamma_\rho : [0,1] \mapsto [0,1]$ by*

$$\Gamma_\rho(\mu) = Pr[X \leq \Phi^{-1}(\mu) \wedge Y \leq \Phi^{-1}(\mu)]$$

*where $X$ and $Y$ are jointly normal random variables with mean 0 and covariance matrix* $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

The only fact that we will use for stating the stronger form of SSEH is the asymptotic behavior of $\Gamma_\rho(\mu)$ when $\rho$ is close to 1 and $\mu$ bounded away from 0.

**Fact 1.** *There is a constant $c > 0$ such that for all sufficiently small $\epsilon$ and all $\mu \in [1/10, 1/2]$,*[2]

$$\Gamma_{1-\epsilon}(\mu) \leq \mu(1 - c\sqrt{\epsilon}).$$

**Conjecture 2** (SSE Hypothesis, Equivalent Formulation [84])**.** *For every integer $q > 0$ and $\epsilon, \gamma > 0$, it is NP-hard to distinguish between the following two cases for a given regular graph $G = (V, E)$:*

- ***Yes*** *There is a partition of $V$ into $q$ equi-sized sets $S_1, \cdots, S_q$ such that $\Phi_G(S_i) \leq 2\epsilon$ for every $1 \leq i \leq q$.*
- ***No*** *For every $S \subseteq V$, letting $\mu = |S|/|V|$, it holds that $\Phi_G(S) \geq 1 - (\Gamma_{1-\epsilon/2}(\mu) + \gamma)/\mu$,*

*where the $\Gamma_{1-\epsilon/2}(\mu)$ is the Gaussian noise stability.*

We present two remarks about the Conjecture 2 from [4], which are relevant to our proof of Theorem 2.

**Remark 1.** *[4] The **Yes** instance of Conjecture 2 implies that the number of edges leaving each $S_i$ is at most $4\epsilon|E|/q$, so the total number of edges not contained in one of the $S_i$ is at most $2\epsilon|E|$.*

[2]Note that the lower bound on $\mu$ can be taken arbitrarily close to 0. So the statement holds with $\mu \in [\epsilon', 1/2]$ for any constant $\epsilon' > 0$.

**Remark 2.** *[4] The* ***No*** *instance of Conjecture 2 implies that for $\epsilon$ sufficiently small, there exists some constant $c'$ such that $\Phi_G(S) \geq c'\sqrt{\epsilon}$, provided that $\mu \in [1/10, 1/2]$ and setting $\gamma \leq \sqrt{\epsilon}$. In particular, $|E(S, V\backslash S)| \geq \Omega(\sqrt{\epsilon}|E|)$, for any $|V|/10 \leq |S| \leq 9|V|/10$.* [3]

Remark 1 follows from the definition of normalized edge expansion and the fact that sum of degree is two times number of edges. Remark 2 follows from Fact 1. The strong form of SSE Hypothesis 2, Remark 1, and Remark 2 will be particularly helpful for proving our SSE-hardness of approximation result (Theorem 2).

### 2.2.2 Preliminaries for Corruption Detection on Networks

We model networks as directed or undirected graphs, where each vertex in the network can be one of two types: truthful or corrupted. At times, we will informally call truthful vertices "good" and corrupt vertices "bad." We say that the corrupt party has *budget* $b$ if it can afford to corrupt at most $b$ nodes of the graph. Given a vertex set $V$, and a budget $b$, the corrupt entity will choose to control a subset of nodes $B \subseteq V$ under the constraint $|B| \leq b$. The rest of the graph remains as truthful vertices, i.e., $T = V\backslash B \subseteq V$. We assume that there are more truthful than corrupt nodes ($b < |V|/2$). It is easy to see that in the case where $|B| \geq |T|$, the corrupt nodes can prevent the identification of even one truthful node, by simulating truthful nodes (see e.g. [1]).

Each node audits and reports its (outgoing) neighbors' identities. That is, each vertex $u \in V$ will report the type of each $v \in \mathcal{N}(u)$, which is a vector in $\{0,1\}^{|\mathcal{N}(u)|}$. Truthful nodes always report the truth, i.e., it reports its neighbor $v \in T$ if $v$ is truthful, $v \in B$ if $v$ is corrupt. The corrupt nodes report their neighbors' identities adversarially. In summary, a strategy of the bad agents is composed of a strategy to take over at most $b$ nodes on the graph, and reports on the nodes that neighbor them.

[3] Recall that Fact 1 is true for $\mu \in [\epsilon', 1/2]$ for any constant $\epsilon' > 0$. Therefore, Remark 2 can be strengthened and states, for any $\epsilon'|V| \leq |S| \leq (1-\epsilon')|V|$, $|E(S, V\backslash S)| \geq \Omega(\sqrt{\epsilon}|E|)$. This will be a useful fact for proving hardness of approximation of $m(G, g)$ for finding many truthful nodes in Section 2.5.

**Definition 5** (**Strategy for a corrupt party**). *A strategy for the corrupt party is a function that maps a graph $G$ and budget $b$ to a subset of nodes $B$ with size $|B| \leq b$, and a set of reports that each node $v \in B$ gives about its neighboring nodes, $\mathcal{N}(v)$.*

**Definition 6** (**Computationally bounded corrupt party**). *We say that the corrupt party is computationally bounded if its strategy can only be a polynomial-time computable function.*

The task for the central agency is to find a good node on this corrupted network, based on the reports. It is clear that the more budget the corrupt party has, the harder the task of finding one truthful node becomes. It was observed in [1] that, for any graph, it is not possible to find one good node if $b \geq |V|/2$. If $b = 0$, it is clear that the entire set $V$ is truthful. Therefore, given an arbitrary graph $G$, there exists a critical number $m(G)$, such that if the bad party has budget lower than $m(G)$, it is always possible to find a good node; if the bad party has budget greater than or equal to $m(G)$, it may not be possible to find a good node. In light of this, we define the critical number of bad nodes on a graph $G$. First, we formally define what we mean when we say it is impossible to find a truthful node on a graph $G$.

**Definition 7** (**Impossibility of finding one truthful node**). *Given a graph $G = (V, E)$, the bad party's budget $b$ and reports, we say that it is* impossible to identify one truthful node *if for every $v \in V$ there is a configuration of the identities of the nodes where $v$ is bad, and the configuration is consistent with the given reports, and consists of fewer than or equal to $b$ bad nodes.*

**Definition 8** (**Critical number of bad nodes on a graph $G$, $m(G)$**). *Given an arbitrary graph $G = (V, E)$, we define $m(G)$ as the minimum number $b$ such that there is a way to distribute $b$ corrupt nodes and set their corresponding reports such that it is impossible to find one truthful node on the graph $G$, given $G$, the reports and that the bad party's budget is at most $b$.*

For example, for a star graph $G$ with $|V| \geq 5$, the critical number of bad nodes is $m(G) = 2$. If there is at most 1 corrupt node on $G$, the central agency can always find

a good node, thus $m(G) \neq 1$. If there are at most 2 bad nodes on $G$, then the bad party can control the center node and one of the leaves. It is impossible for central agency to find one good node.

Given a graph $G$, by definition there exists some set of $m(G)$ nodes that can make it impossible to find a good node if they are corrupted. However, this does not mean that the corrupt party can necessarily find this set in polynomial time. Indeed, Theorem 2 establishes that they cannot always find this set in polynomial time if we assume the SSE Hypothesis (Conjecture 2) and that P $\neq$ NP.

## 2.3 Proofs of Theorems 1, 2, and 3

In the following section, we state our main results by first presenting the close relation of our problem to the $k$-vertex separator problem. Then we use this characterization to prove Theorem 1. This characterization will additionally be useful for the proofs of Theorems 2 and 3, which we will give in Section 2.3.2 and Section 2.3.3.

### 2.3.1 2-Approximation by Vertex Separation

**Lemma 1** (2-Approximation by Vertex Separation)**.** *The critical number of corrupt nodes for graph $G$, $m(G)$, can be bounded by the minimal sum of $k$-vertex separator and $k$, $\min_k(S_G(k)+k)$, up to a factor of 2. i.e.,*

$$\frac{1}{2}\min_k\left(S_G(k)+k\right) \leq m(G) \leq \min_k\left(S_G(k)+k\right)$$

*Proof of Lemma 1.* The direction $m(G) \leq \min_k S_G(k) + k$ follows simply. Let $k^* = \arg\min_k(S_G(k)+k)$. If the corrupt party is given $S_G(k^*)+k^*$ nodes to corrupt on the graph, it can first assign $S_G(k^*)$ nodes to the separator, thus the remaining nodes are partitioned into components of size at most $k^*$. Then it arbitrarily assigns one of the components to be all bad nodes. The bad nodes in the connected components report the nodes in the same component as good, and report any node in the separator as

bad. The nodes in the separator can effectively report however they want (e.g. report all neighboring nodes as bad). It is impossible to identify even one single good node, because all connected components of size $k$ can potentially be bad, and all vertices in the separator are bad.

The direction $1/2 \min_k(S_G(k) + k) \leq m(G)$ can be proved as follows. When there are $b = m(G)$ corrupt nodes distributed optimally in $G$, it is impossible to find a single good node by definition, and therefore, in particular, the following algorithm (Algorithm 1) cannot always find a good node:

**Algorithm 1** Finding one truthful vertex on undirected graph $G$

Input: Undirected graph $G$

- If the reports on edge $(u, v)$ does not equal to $(u \in T, v \in T)$, remove both $u, v$ and any incident edges. Remove a pair of nodes in each round, until there are no bad reports left.
- Call the remaining graph $H$. Declare the largest component of $H$ as good.

Run Algorithm 1 on $G$, and suppose the first step terminates in $i$ rounds, then:

- No remaining node reports neighbors as corrupt
- $|V|-2i$ nodes remain in graph
- $\leq b - i$ bad nodes remain in the graph, because each time we remove an edge with bad report, and one of the end points must be a corrupt vertex.

Note that if two nodes report each other as good, they must be the same type (either both truthful, or both corrupt.) Since graph $H$ only contains good reports, nodes within a connected component of $H$ have the same types. If there exists a component of size larger than $b - i$, it exceeds bad party's budget, and must be all good. Therefore, Algorithm 1 would successfully find a good node.

Since Algorithm 1 cannot find a good node, the bad party must have the budget to corrupt the largest component of $H$, which means it has size at most $b - i$. Hence, $S_G(b - i) \leq 2i$. Plugging in $b = m(G)$, we get that

$$m(G) = \frac{2i}{2} + b - i \geq \min_k (S_G(k)/2 + k) \geq \frac{1}{2} \min_k (S_G(k) + k),$$

where the first inequality comes from $2i \geq S_G(b - i)$. □

Furthermore, the upperbound in Lemma 1 additionally tells us that if corrupt party's budget $b \leq m(G)/2$, the set output by Algorithm 1 is guaranteed to be good.

**Theorem 1.** *Fix a graph $G$ and suppose that the corrupt party has a budget $b \leq m(G)/2$. Then the central agency can identify a truthful node, regardless of the strategy of the corrupt party, and without knowledge of either $m(G)$ or $b$. Furthermore, the central agency's algorithm runs in linear time (in the number of edges in the graph $G$).*

*Proof of Theorem 1.* Suppose the corrupt party has budget $b \leq m(G)/2$. Run Algorithm 1. We remove $2i$ nodes in the first step, and separate the remaining graph $H$ into connected components. Notice each time we remove an edge with bad report, at least one of the end point is a corrupt vertex. So we have removed at most $2b \leq m(G) \leq \lceil |V|/2 \rceil$ nodes. Therefore, the graph $H$ is nonempty, and the nodes in any connected component of $H$ have the same identity. Let $k^* \geq 1$ be the size of the maximum connected component of $H$. We can conclude that $S_G(k^*) \leq 2i$, since $2i$ is a possible size of $k^*$-vertex separator of $G$.

Notice there are at most $b - i \leq m(G)/2 - i$ bad nodes in $H$ by the same fact that at least one bad node is removed each round. By the upper bound in Lemma 1,

$$b - i \leq m(G)/2 - i \leq \min_k (S_G(k) + k)/2 - i \leq (2i + k^*)/2 - i \leq \frac{k^*}{2}.$$

Since $k^* \geq 1$, the connected component of size $k^*$ exceeds the bad party's remaining budget $k^*/2$, and must be all good.

Algorithm 1 is linear time because it loops over all edges and removes any "bad" edge that does not have reports $(T, T)$ (takes $\leq |E|$ time when we use a list with "bad" edges at the front), and counts the size of the remaining components ($\leq |V|$ time), and thus is linear in $|E|$. □

**Remark 3.** *Both bounds in Lemma 1 are tight. For the lower bound, consider a complete graph with an even number of nodes. For the upper bound, consider a complete bipartite graph with one side smaller than the other.*

To elaborate on Remark 3, for the lower bound, in a complete graph with $n$ nodes, the critical number of bad nodes is $n/2$, and $\min_k S_G(k) + k = n$.

For the upper bound, consider a complete bipartite graph $G = (V, E)$. The vertex set is partitioned into two sets $V = S_1 \cup S_2$ where the induced subgraphs on $S_1$ and $S_2$ consist of isolated vertices, and every vertex $u \in S_1$ is connected with every vertex $v \in S_2$. The smallest sum of $k$-vertex separator with $k$ is obtained with $k = 1$, i.e., $\min_k S_G(k) + k = \min\{|S_1|, |S_2|\} + 1$. We argue that this is also the minimal number of bad nodes needed to corrupt the graph. Without loss of generality , let $|S_1| < |S_2|$. If the bad party controls all of $S_1$ plus one node in $S_2$, it can prevent the identification of a good node. On the other hand, if the bad party controls $b < |S_1|+1$ nodes, then we can always identify a good node. Specifically, we are in one of the following cases:

1. The bad party does not control all of $S_1$. Then there will be a connected component of size $n - b > b$ that report each other as good, because the bad nodes cannot control all of $S_2$, and any induced subgraph of a complete bipartite graph with nodes on both sides is connected.

2. The bad party controls all of $S_1$. In this case, the largest connected component of nodes that all report each other as good is only 1. However, in this case, we conclude that the bad nodes must control all of $S_1$ and no other node (due to their budget). Hence, any node in $S_2$ is good.

We end by discussing that the efficient algorithm given in this section does not address the regime when the budget of the bad party, $b$, falls in $m(G)/2 < b \leq m(G)$. Though by definition of $m(G)$, the central agency can find at least one truthful node as long as $b \leq m(G)$, by, for example, enumerating all possible assignments of good/bad nodes consistent with the report, and check the intersection of the assignment of good nodes. However, it is not clear that the central agency has a polynomial time algorithm for doing this. Of course, one can always run Algorithm 1, check whether

the output set exceeds $b - i/2$, and concludes that the output set is truthful if that is the case. However, there is no guarantee that the output set will be larger than $b - i/2$ if $m(G)/2 < b \leq m(G)$. We propose the following conjecture:

**Conjecture 3.** *Fix a graph $G$ and suppose that the corrupt party has a budget $b$ such that $m(G)/2 < b \leq m(G)$. The problem of finding one truthful node given the graph $G$, bad party's budget $b$ and the reports is NP-hard.*

### 2.3.2 SSE-Hardness of Approximation for $m(G)$

In this section, we show the hardness of approximation result for $m(G)$ within any constant factor under the Small Set Expansion (SSE) Hypothesis [83]. Specifically, we prove Theorem 2.

**Theorem 2.** *For every $\beta > 1$, there is a constant $\epsilon > 0$ such that the following is true. Given a graph $G = (V, E)$, it is SSE-hard to distinguish between the case where $m(G) \leq \epsilon \cdot |V|$ and $m(G) \geq \beta \cdot \epsilon \cdot |V|$. Or in other words, the problem of approximating the critical number of corrupt nodes for a graph to within any constant factor is SSE-hard.*

In order to prove Theorem 2, we construct a reduction similar to [4], and show that the bad party can control auxiliary graph of the **Yes** case of SSE with $b = O(\epsilon|V'|)$ and cannot control the auxiliary graph of the **No** case of SSE with $b = \Omega(\epsilon^{0.51}|V'|)$.

Given an undirected $d$-regular graph $G = (V, E)$, construct an auxiliary undirected graph $G' = (V', E')$ in the following way [4]. Let $r = d/2$. For each vertex $v^i \in V$, make $r$ copies of $v^i$ and add to the vertex set of $G'$, denoted $v^i_1, \cdots, v^i_r$. Denote the resulting set of vertices as $\tilde{V} = V \times \{1, \cdots, r\}$. Each edge $e^k \in E$ of $G$ becomes a vertex in $G'$, denoted $e^k$. Denote this set of vertices as $\tilde{E}$. In other words, $V' = \tilde{V} \cup \tilde{E} = V \times \{1, \cdots, r\} \cup E$. There exists an edge between a vertex $v^i_j$ and a vertex $e^k$ of $G'$ if $v^i$ and $e^k$ were adjacent edge and vertex pair in $G$. Note that $G'$ is a bipartite $d$-regular graph with $d/2|V|+|E|= 2|E|$ vertices.

**Lemma 2.** *Suppose $q = 1/\epsilon$, and $G$ can be partitioned into $q$ equi-sized sets $S_1, \cdots, S_q$*

*such that* $\Phi_G(S_i) \leq 2\epsilon$ *for every* $1 \leq i \leq q$. *Then the bad party can control the auxiliary graph* $G'$ *with at most* $4\epsilon|E| = 2\epsilon|V'|$ *nodes.*

*Proof of Lemma 2.* Notice by Remark 1, the total number of edges in $G$ not contained in one of the $S_i$ is at most $2\epsilon|E|$.

This implies that a strategy for the bad party to control graph $G'$ is as follows. Control vertex $e^k \in \tilde{E}$ if $e^k \in E$ is not contained in any of the $S_i$s in $G$. Call the set of such vertices $E^* \subseteq \tilde{E}$. Let $S_i^* \subseteq V'$ be the set that contains all $r$ copies of nodes in $S_i \subseteq V$. Control one of the $S_i^* s$, say $S_1^*$. Control all the edge nodes in $\tilde{E}$ that are adjacent to $S_1^*$. Call this set $\mathcal{N}(S_1^*)$. The corrupt nodes in $S_1^* \cup \mathcal{N}(S_1^*)$ report their neighbors in $S_1^* \cup \mathcal{N}(S_1^*)$ as good, and report $E^*$ as bad. Nodes in $E^*$ can effectively report however they want; suppose they report every neighboring node as bad. Then, it is impossible to identify even one truthful node, since assigning any $S_i^*$ as corrupt is consistent with the report and within bad party's budget.

This strategy controls $|E^*|+|S_i^*|+|\mathcal{N}(S_1^*)\backslash E^*|$ nodes on $G'$. Note that $|\mathcal{N}(S_1^*)\backslash E^*|$ is equal to the number of edges that are totally contained in $S_1$ on $G$, which is bounded by $|S_1|\cdot d/2$ (that is if all edges adjacent to $S_1$ are totally contained in $S_1$). If $q = 1/\epsilon$, this strategy amounts to controlling $|E^*|+|S_i^*|+|\mathcal{N}(S_1^*) \setminus E^*| \leq 2\epsilon|E|+d/2 \cdot |V|/q + |V|/q \cdot d/2 = 4\epsilon|E| = 2\epsilon|V'|$ nodes on $G'$. Notice, this number is guaranteed to be smaller than $1/2|V'|$, as long as $q > 4$.

□

Note that, different from the argument in [4], we cannot take $r$ to be arbitrarily large (e.g. $> O(|V||E|)$). This is because when $r$ is large, $2\epsilon|E|+r \cdot |V|/q = O(\epsilon(|E|+|V'|)) = O(\epsilon|V'|)$, and will not be comparable with the $O(\sqrt{\epsilon}|E|)$ in Lemma 3.

**Lemma 3.** *Let* $G = (V, E)$ *be an undirected* $d$*-regular graph with the property that for every* $|V|/10 \leq |S| \leq 9|V|/10$ *we have* $|E(S, V \setminus S)| \geq \Omega(\sqrt{\epsilon}|E|)$. *If bad party controls* $O(\epsilon^{0.51}|E|) = O(\epsilon^{0.51}|V'|) < 1/2|V'|$ *nodes on the auxiliary graph* $G'$ *constructed from* $G$, *we can always find a truthful node on* $G'$.

*Proof of Lemma 3.* Assume towards contradiction that the bad party controls $O(\epsilon^{0.51}|E|)$ vertices of graph $G'$, and we can't identify a truthful node.

**Claim 1.** *If the bad party controls $O(\epsilon^{0.51}|E|)$ vertices of graph $G'$, and it is impossible to identify a truthful node, then there exists a set $C$ of size $O(\epsilon^{0.51}|E|)$ and separates $V'\backslash C$ into sets $\{T'_i\}_{i=1,\cdots,\ell}$, each of size $O(\epsilon^{0.51}|E|)$.*

*Proof of Claim 1.* Since the bad nodes can control $G'$ with $O(\epsilon^{0.51}|E|)$ vertices, $m(G') \leq O(\epsilon^{0.51}|E|)$. By the lower bound in Lemma 1, $\min_k(S_{G'}(k)+k) \leq 2m(G') \leq O(\epsilon^{0.51}|E|)$. Let $k^* = \arg\min_k(S_{G'}(k)+k)$. Then $k^* \leq O(\epsilon^{0.51}|E|)$, $S_{G'}(k^*) \leq O(\epsilon^{0.51}|E|)$. By definition of $S_{G'}(k^*)$, there exists a set of size $S_{G'}(k^*)$ whose removal separates the remainder of the graph $G'$ to connected components of size at most $k^*$. □

Let $C$ and $T'_i$ be the sets guaranteed by Claim 1. Note we have taken $r = d/2$, and thus $|\tilde{V}| = |\tilde{E}|$. In other words, half of the $V'$ are "vertex" vertices $\tilde{V}$, and half are "edge" vertices $\tilde{E}$. Therefore, with sufficiently small $\epsilon$, $|C \cap \tilde{V}| \leq |C| < 1/2|\tilde{V}|$, $|(\cup_{i=1}^{\ell} T'_i) \cap \tilde{V}| \geq 1/2|\tilde{V}|$, $|T'_i \cap \tilde{V}| \leq |T'_i| < 3/10|\tilde{V}|$ for every $i$. Therefore, we can merge the different $T'_i$s in Claim 1, and have two sets $T'_1$ and $T'_2$, such that $|T'_1 \cap \tilde{V}| \geq |\tilde{V}|/5$ and $|T'_2 \cap \tilde{V}| \geq |\tilde{V}|/5$. Furthermore, $T'_1$ and $T'_2$ are disjoint, and $T'_1, T'_2$, and $C$ cover $V'$.

Similar to the proof of Lemma 5.1 in [4], we let $T_1 \subseteq V$ (resp. $T_2 \subseteq V$) be the set of vertices $v \in V$ such that some copy of $v$ appears in $T'_1$ (resp. $T'_2$). Let $S \subseteq V$ be the set of vertices $v \in V$ such that *all* copies of $v$ appear in $C$. Since $|T'_1 \cap \tilde{V}|, |T'_2 \cap \tilde{V}| \geq |\tilde{V}|/5 = r|V|/5$, both $|T_1|, |T_2| \geq |V|/5$. Furthermore, we observe that $T_1 \cup T_2 \cup S = V$, which follows since $T'_1 \cup T'_2 \cup C = V'$. Now we can lower bound $|T_1 \cup T_2|$ as follows.

$$|T_1 \cup T_2| = |V\backslash S| \geq |V| - |C|/r \geq |V| - c\epsilon^{0.51}|E|/r = |V| - c\epsilon^{0.51}|V|,$$

where the first equality uses the fact that $T_1 \cup T_2 \cup S = V$ and that $T_1 \cup T_2$ is disjoint from $S$, and the following inequality uses the fact that $|S| \leq |C|/r$, which follows by definition.

Since $|T_1 \cup T_2|$ is sufficiently large, we can find a balanced partition of $T_1 \cup T_2$ into sets $S_1 \subseteq T_1$, $S_2 \subseteq T_2$, such that $S_1 \cap S_2 = \emptyset, S_1 \cup S_2 = T_1 \cup T_2$, and $|V|/10 \leq$

$|S_1|, |S_2| \leq 9|V|/10$. From the property of $G$ that $E(S, V \setminus S) \geq \Omega(\sqrt{\epsilon}|E|)$ in Lemma 3 and the fact that $G$ is $d$-regular, we know that

$$E(S_1, S_2) = E(S_1, V \setminus S_1) - E(S_1, S) \geq \alpha\sqrt{\epsilon}|E| - d(\epsilon^{0.51}|E|/r) = \alpha\sqrt{\epsilon}|E| - 2\epsilon^{0.51}|E| = \Omega(\sqrt{\epsilon}|E|),$$

for some constant $\alpha$. In the first equality we use the fact that $S_1, S_2, S$ form a partition of $V$. Thus $E(S_1, V \setminus S_1) = E(S_1, S_2 \cup S) = E(S_1, S_2) + E(S_1, S)$.

Note that since $S_1 \subseteq T_1$ and $S_2 \subseteq T_2$, and $T_1'$ and $T_2'$ do not have edge between them in $G'$, the edges $E(S_1, S_2)$ all have to land as "edge vertices" in $C$. In other words, for any $u \in S_1$, and $v \in S_2$, if $(u, v) \in E$, then the vertex $(u, v) \in V'$ has to be included in the set $C$, thus $|C| \geq \Omega(\sqrt{\epsilon}|E|)$.

This contradicts the fact that there are only $O(\epsilon^{0.51}|E|)$ vertices in $C$.

□

Combining Lemma 2 and Lemma 3, Theorem 2 follows in standard fashion. We give a proof here for completeness.

*Proof of Theorem 2.* Suppose for contradiction that there exists some constant $\beta > 0$ such that there is polynomial time algorithm $\mathcal{A}$ that does the following. For any $\epsilon' > 0$ and an arbitrary graph $G' = (V', E')$, it can distinguish between the case where $m(G') \leq \epsilon' \cdot |V'|$ and $m(G') \geq \beta \cdot \epsilon' \cdot |V'|$. Specifically, we will suppose this holds for $\epsilon' < \frac{1}{\beta^{2.05}}$. Then we can use this algorithm to decide the SSE problem as follows.

Fix $\epsilon < \frac{1}{1.5\beta^{2.05}}$, $q = 1/\epsilon$, $\gamma > 0$ sufficiently small ($\leq o(\sqrt{\epsilon})$ suffices). Let $G = (V, E)$ be an arbitrary input to the resulting instance of the SSE decision problem (from Conjecture 2). Construct the graph $G' = (V', E')$ from $G$ as done in the beginning of Section 2.3.2.

If $G$ was from the YES case of Conjecture 2, then $m(G') \leq 1.5\epsilon|V'|$ (Lemma 2). If $G$ was from the NO case of Conjecture 2, then $m(G') > \epsilon^{0.51}|V'|$ (Lemma 3). We can invoke our algorithm $\mathcal{A}$ to distinguish these two cases, by letting $\epsilon' = 1.5\epsilon$ and noting

that $\beta < (1/(\epsilon')^{0.49})$ by design, which would decide the problem in Conjecture 2 in polynomial time. □

Now, we can obtain the following Corollary 1 from Theorem 2.

**Corollary 1.** *Assume the SSE Hypothesis and that $P \neq NP$. Fix any $\beta > 1$. There does not exist a polynomial-time algorithm that takes as input an arbitrary graph $G = (V, E)$ and outputs a set of nodes $S$ with size $|S| \leq O(\beta \cdot m(G))$, such that corrupting $S$ prevents the central agency from finding a truthful node.*

In summary, the analysis in this section tells us that given an arbitrary graph, it is hard for bad party to corrupt the graph with minimal resources. On the other hand, if the budget of bad nodes is a factor of two less than $m(G)$, a good party can always be detected with an efficient algorithm, e.g. using Algorithm 1.

### 2.3.3 An $O(\log|V|)$ Approximation Algorithm for $m(G)$

In light of the SSE-hardness of approximation of $m(G)$ within any constant, and the close relation of $m(G)$ with $k$-vertex separator, we leverage the best known approximation result for $k$-vertex separator to propose an $O(\log n)$ approximation algorithm for $m(G)$. It is useful as a test for central authorities for measuring how corruptible a graph is. Notably, it is also a potential algorithm for (computationally restricted) bad party to use to decide which nodes to corrupt.

The paper [63] presents an bicritera approximation algorithm for $k$-vertex separator, with the guarantee that for each $k$, the algorithm finds a subset $B_k \subseteq V$ such that $|B_k| \leq O(\frac{\log k}{\epsilon}) \cdot S_G(k)$, and the induced subgraph $G_{V \setminus B_k}$ is divided into connected components each of size at most $k/(1-2\epsilon)$ vertices.

**Proposition 1** (Theorem 1.1, [63])**.** *For any $\epsilon \in (0, 1/2)$, there is a polynomial-time $(\frac{1}{1-2\epsilon}, O(\frac{\log k}{\epsilon}))$- bicriteria approximation algorithm for $k$-vertex separator.*

Interested readers can refer to [63] Section 3 for the description of the algorithm. Leveraging this algorithm for $k$-vertex separator, we can obtain a polynomial-time

algorithm for seeding corrupt nodes and preventing the identification of a truthful node.

**Theorem 3** ($O(\log|V|)$ Approximation Algorithm)**.** *There is a polynomial-time algorithm that takes as input a graph $G = (V, E)$ and outputs a set of nodes $S$ with size $|S| \leq O(\log|V| \cdot m(G))$, such that corrupting $S$ prevents the central agency from finding a truthful node.*

*Proof.* The algorithm is as follows. Call the bicriteria algorithm for approximating $k$-vertex separator in [63] $n$ times, once for each $k$ in $k = 1, \cdots, n$, where $n = |V|$. Each time the algorithm outputs a set of vertices $B_k$ that divides the remaining graph into connected components with maximum size $g(k)$. Choose the $k^*$ for which the algorithm outputs the smallest value of $\min_k |B_k| + g(k)$. The bad party can control $B_{k^*}$ and one of the remaining connected components (the size of which is at most $g(k^*)$), and be sure to prevent the identification of one good node, by the same argument that lead to the upper bound in Lemma 1.

We now prove that $|B_{k^*}| + g(k^*)$ is an $O(\log|V|)$ approximation for the quantity of consideration $\min_k S_G(k) + k$. For each $k$, we denote our approximation for $S_G(k) + k$ as $f(k) := |B_k| + g(k)$. Then by the guarantee given in Proposition 1, we know

$$f(k) = |B_k| + g(k) \leq O\left(\frac{\log k}{\epsilon}\right) \cdot S_G(k) + \frac{1}{1 - 2\epsilon} k \leq O\left(\frac{\log k}{\epsilon}\right) \cdot (S_G(k) + k).$$

Thus

$$\min_k f(k) \leq \min_k O\left(\frac{\log k}{\epsilon}\right) \cdot (S_G(k) + k) \leq O\left(\frac{\log n}{\epsilon} \min_k (S_G(k) + k)\right) \leq O\left(\log n \cdot m(G)\right).$$

The last inequality follows from the fact that that $\min_k (S_G(k) + k)/2 \leq m(G) \leq \min_k (S_G(k) + k)$ in Lemma 1, and by taking $\epsilon$ to be a fixed constant, e.g. $\epsilon = 1/3$. So $\min_k f(k)$ provides an $O(\log n)$ approximation of $m(G)$. The algorithm consists of $n$ calls of the polynomial-time algorithm in Proposition 1, so is also polynomial-time. □

## 2.4 Directed Graphs

Here we present the variant of our problem on directed graphs. As discussed in [1], this is motivated by the fact that in various auditing situations, it may not be natural that any $u$ will be able to inspect $v$ whenever $v$ inspects $u$.

Given a directed graph $D = (V, E_D)$, we are asked to to find $m(D)$, the minimal number of corrupted agents needed to prevent the identification of a single truthful agent. Firstly, since undirected graphs are special cases of directed graphs, it is clear that the worst case hardness of approximation results still hold. In this section, we will define a analogous notion of vertex separator relevant to corruption detection for directed graphs, and state the version of Theorem 1 for directed graphs.

**Definition 9** (**Reachability Index**). *On a directed graph $D = (V, E_D)$, say a vertex $s$ can reach a vertex $t$ if there exists a sequence of adjacent vertices (i.e. a path) which starts with $s$ and ends with $t$. Let $R_D(v)$ be the set of vertices that can reach a vertex $v$. Define the* ***reachability index*** *of $v$ as $|R_D(v)|$, or in other words, as the total number of nodes that can reach $v$.*

Based on the notion of reachability index, we design the following algorithm, Algorithm 2, for detecting one good node on directed graphs:

**Algorithm 2** Finding one truthful vertex on directed graph $D$

Input: Directed graph $D$

- If node $u$ reports node $v$ as corrupt, remove both $u, v$ and any incident edges (incoming and outgoing). Remove a pair of nodes in each round. Continue until there are no bad reports left.
- Call the remaining graph $H = (V_H, E_H)$. Declare a vertex in $H$ with maximum reachability index as good.

Run Algorithm 2 on directed graph $D$, and suppose the first step terminates in $i$ rounds. Then:

- No remaining node reports out-neighbors as corrupt
- $|V|-2i$ nodes remain in graph
- $\leq b-i$ bad nodes remain in the graph, because each round in step 1 removes at least one bad node.

The main idea is that, if there exists a node $v$ with reachability index larger than $b-i$, at least $b-i$ nodes claim (possibly indirectly) that $v$ is good, which means at least one good node also reports $v$ as good, and thus $v$ must be good. In the rest of the section, we use this observation to generalize Theorem 1.

We define a notion similar to $k$-vertex separator on directed graphs, show that our notion provides a 2-approximation for $m(D)$ when $D$ is a directed graph, and that the equivalent of Theorem 1 also holds in the directed case.

**Definition 10** (**$k$-reachability separator**)**.** *We say a set of vertices $S \subseteq V$ is a $k$-reachability separator of a directed graph $D = (V, E_D)$ if after the removal of $S$ and any adjacent edges, all vertices in the remaining graph are of reachability at most $k$.*

Since in an undirected graph, any pair of vertices can reach each other if and only if they belong to the same connected component, one can check that $k$-reachability separator on an undirected graph is exactly equivalent to a $k$-vertex separator. Thus we use a similar notation, $S_D(k)$, to denote the size of the minimal $k$-reachability separator on $D$.

**Lemma 4** (2-Approximation Lemma on Directed Graphs)**.**

$$\frac{1}{2}\min_k(S_D(k)+k) \leq m(D) \leq \min_k(S_D(k)+k)$$

*Proof.* The direction $m(D) \leq \min_k S_D(k)+k$ is proved as follows. Let $k^* = \arg\min_k(S_D(k)+k)$. If the corrupt party is given $\min_k(S_D(k)+k)$ nodes to allocate on $D$, it can first assign $S_D(k^*)$ nodes to a $k^*$-reachability separator $C$, such that the remaining nodes have reachability index at most $k^*$. Then it arbitrarily assigns one of the vertices $v^*$ with maximum reachability index plus its $R_H(v^*)$ as bad. The bad nodes in $R_H(v^*)$

report any neighbor in the separator $C$ as bad and any other neighbor as good. The nodes in the separator can effectively report however they want (e.g. report all neighboring nodes as bad).

It is impossible to detect a single good node, because every node $v$ can only be reached by $R_H(v)$ and $C$. For every $v \in H$, it being assigned as corrupt or good is consistent with the reports. If $v$ is corrupt, $R_H(v)$ is also assigned as corrupt, thus all nodes in $H$ receive good reports from $R_H(v)$, bad reports from $C$ and give bad reports to $C$. If $v$ is truthful, all nodes still receive and give the same reports. So for every $v \in V_H$, assigning $R_H(v)$ as bad, and $V_H \setminus R_H(v)$ as good is consistent with the observed reports. It is impossible to find a good node in $H$ by definition.

The proof for $1/2 \min_k(S_D(k) + k) \leq m(D)$ is given by Algorithm 2. Let there be $m(D)$ bad nodes distributed optimally on the graph. By definition, these nodes prevent the identification of a good node. Run Algorithm 2, and suppose the first step terminates in $i$ rounds. This means we have removed at least $i$ bad nodes, and there are at most $m(D) - i$ bad nodes left on $H$. If there exists a node on $H$ with reachability $m(D) - i$, then this node must be truthful, since there are not enough bad nodes left to corrupt all the nodes that can reach it, and all the reports in the remaining graph are good. Thus $|R(v)| < m(D) - i$ for any $v$. Therefore, the set of $2i$ removed nodes must be an $m(D) - i$ reachability separator. Hence, we can bound $m(D)$ as follows.

$$m(D) = (m(D) - i) + 2i/2 \geq \min_k(k + S_D(k)/2) \geq \frac{1}{2} \min_k(S_D(k) + k)$$

where the first inequality follows from the fact that $2i \geq S_D(m(D) - i)$. □

**Theorem 4.** *Fix a directed graph $D$ and suppose that the corrupt party has a budget $b \leq m(D)/2$. Then the central agency can identify a truthful node, regardless of the strategy of the corrupt party, and without the knowledge of either $m(D)$ or $b$. Furthermore, the central agency's algorithm runs in linear time.*

*Proof of Theorem 4.* Suppose the corrupt party has budget $b \leq m(D)/2$. Run Algo-

rithm 2. Notice each time we remove an edge with bad report, at least one of the end point is a corrupt vertex. So we have removed at most $2b \leq m(D) \leq \lceil |V|/2 \rceil$ nodes. Therefore, the graph $H$ is nonempty. Let $k^* \geq 1$ be the maximum reachability index in $H$. Since $b \leq m(D)/2$, and there are no bad reports in $H$, the reachability index of a bad node in graph $H$ is at most $m(D)/2 - i \leq \min_k(S_D(k) + k)/2 - i \leq (2i + k^*)/2 - i = k^*/2 < k^*$.

Then a vertex with reachability index $k^*$ must be found by Algorithm 2, and must be a truthful node. The linear runtime $O(|E_D|)$ follows from the same analysis as in the proof of Theorem 1. □

## 2.5 Finding an Arbitrary Fraction of Good Nodes on a Graph

Being able to detect one good node may seem limited, but in fact, the same arguments and construction can be adapted to show that approximating the critical number of bad nodes to prevent detection of any arbitrary $\delta$ fraction of good nodes is SSE-hard. In this section, we propose the definition of $g$-remainder $k$-vertex separator, a vertex separator notion related to identifying arbitrary number of good nodes, present a 2-approximation result, and prove hardness of approximation with arguments similar to proof of Theorem 2 in Section 2.3.2.

We abuse notation and define $m(G, g)$ to be the minimal number of bad nodes needed to prevent the identification of $g$ nodes.

**Definition 11** ($m(G, g)$)**.** *We define $m(G, g)$ as the minimal number of bad nodes such that it is impossible to find $g$ good nodes in $G$. In particular, $m(G) = m(G, 1)$.*

**Definition 12** ($g$**-remainder** $k$**-vertex Separator**)**.** *Consider the following separation property: after the removal of a vertex set $S$, the remaining graph $G_{V \setminus S}$ is a union of connected components, where connected components of size larger than $k$ sum up to size less than $g$. We call such a set $S$ a $g$-remainder $k$-vertex separator of $G$.*

*For any integer $0 < k, g < |V|$, we denote the minimal size of such a set as $S_G(k, g)$. In particular, a minimal $k$-vertex separator is a 1-remainder $k$-vertex separator, i.e., $S_G(k) = S_G(k, 1)$.*

**Theorem 5.** *Fix a graph $G$ and the number of good nodes to recover, $g$. Suppose that the corrupt party has a budget $b \leq m(G, g)/2$. If $g < |V| - 2b$, then the central agency can identify $g$ truthful nodes, regardless of the strategy of the corrupt party, and without knowledge either of $m(G, g)$ or $b$. Furthermore, the central agency's algorithm runs in linear time.*

---

**Algorithm 3** Finding $g$ truthful vertices on an undirected graph $G$

---

Input: Undirected graph $G$

- If the reports on edge $(u, v)$ does not equal to $(u \in T, v \in T)$, remove both $u, v$ and any incident edges. Remove a pair of nodes in each round, until there are no bad reports left.
- Suppose the previous step terminates in $i$ rounds. In the remaining graph $H$, rank the connected component from large to small by size. Declare the largest component as good and remove the declared component until we have declared $g$ nodes as good.

---

*Proof of Theorem 5.* We claim that central agency can use Algorithm 3, and output at least $g$ good nodes if $b \leq m(G, g)/2$. Step 1 of Algorithm 1 must terminate after removing fewer than $m(G, g)$ nodes, because each round has to remove at least one bad node, and there are only $m(G, g)/2$ bad nodes in total. Let the number of nodes removed be $m(G, g) - \delta$, so at least $m(G, g)/2 - \delta/2 \geq b - \delta/2$ are corrupt. Thus at most $\delta/2$ bad nodes remain in the graph $H$.

Assume towards contradiction that only $y < g$ nodes output by Algorithm 1 are good. This means that the $m(G, g) - \delta$ removed nodes separate the graph $G$ into connected components where all components with size larger than $\delta/2$ sum to fewer than $g$. Then $m(G, g) - \delta = m(G, y)$ for $y < g$, contradicting the fact that $m(G, g)$ is the minimum budget needed to prevent identification of $g$ nodes.

□

In fact, just like in Section 2.3, Algorithm 3 additionally gives us a characterization of $m(G, g)$ in terms of the size of the smallest $g$-remainder $k$-vertex separator of a graph, for an appropriately chosen value of $k$.

**Lemma 5** (2-Approximation by Vertex Separation)**.** *The minimal sum of $g$-remainder $k$-vertex separator and $k$, $\min_k (S_G(k, g) + k)$, bounds the critical number of bad nodes $m(G, g)$ up to a factor of 2. i.e.,*

$$\frac{1}{2} \min_k S_G(k, g) + k \leq m(G, g) \leq \min_k S_G(k, g) + k.$$

*Proof of Lemma 5.* The upper bound follows simply. Let $k^* = \arg\min_k S_G(k, g) + k$. Given a budget $b = \min_k S_G(k, g) + k$, the bad party can remove a set of size $S_G(k^*, g)$ and separate the graph into connected components of size at most $k^*$, except for fewer than $g$ nodes. Control one of the connected components of size at most $k^*$, and construct the reports similarly as in Lemma 1. Then the central agency can only identify fewer than $g$ good nodes.

For the lower bound, suppose there are $b = m(G, g)$ bad nodes distributed optimally on $G$ and thus it's impossible to find $g$ good nodes by definition. Run Algorithm 3. Suppose the first step terminates in $i$ rounds. After the removal of $2i$ nodes, the graph must be separated into connected components smaller than $b - i$, except for fewer than $g$ nodes. Then $2i \geq S_G(b - i, g)$. Therefore,

$$\frac{1}{2} \min_k (S_G(k, g) + k) \leq \min_k \left( \frac{S_G(k, g)}{2} + k \right) \leq \frac{1}{2} S_G(b-i, g) + (b-i) \leq \frac{2i}{2} + b - i = m(G, g)$$

□

Now using the characterization given by $g$-remainder $k$-vertex separator, we are ready to prove that it is SSE-hard to approximate the budget needed to prevent any arbitrary number of good nodes, i.e., $m(G, g)$ for any $g < |V|/3$.

**Theorem 6.** *For every $\beta > 1$ and every $0 < \delta < 1$, there is a constant $\epsilon > 0$ such that the following is true. Given a graph $G = (V, E)$, it is SSE-hard to distinguish between the case where $m(G, \delta|V|) \leq \epsilon \cdot |V|$ and $m(G, \delta|V|) \geq \beta \cdot \epsilon \cdot |V|$. Or in other words, the problem of approximating the critical number of corrupt nodes such that it is impossible to find $\delta|V|$ good nodes within any constant factor is SSE-hard.*

We first prove Theorem 6 for $0 < \delta < 1/3$. The proof in this regime follows similar constructions and arguments as in the proof of Theorem 2. Note that the proof extends naturally for any $0 < \delta < 1/2$. This is effectively because the range for $\mu$ in Remark 2 can be made to $[\epsilon', 1/2]$, for any constant $\epsilon' > 0$. Further explanation is provided in proof for Lemma 7.

Firstly, we construct $G'$ based on $G$ as in Section 2.3.2. Lemma 2 immediately implies that:

**Lemma 6.** *Suppose $q = 1/\epsilon$, and $G$ can be partitioned into $q$ equi-sized sets $S_1, \cdots, S_q$ such that $\Phi_G(S_i) \leq 2\epsilon$ for every $1 \leq i \leq q$. The bad party can prevent the identification of one good node, and thus $\delta|V'|$ good nodes, on the auxiliary graph $G'$ with $O(\epsilon|E|) = O(\epsilon|V'|)$ nodes.*

We reprove the analogous lemma to Lemma 3.

**Lemma 7.** *Let $G = (V, E)$ be an undirected $d$-regular graph with the property that for every $|V|/10 \leq |S| \leq 9|V|/10$ we have $|E(S, V \setminus S)| \geq \Omega(\sqrt{\epsilon}|E|)$. If bad party controls $O(\epsilon^{0.51}|E|) = O(\epsilon^{0.51}|V'|) < 1/2|V'|$ nodes on the auxiliary graph $G'$ constructed from $G$, we can always find $\delta|V'|$ truthful nodes on $G'$, for any $\delta < 1/3$.*

*Proof of Lemma 7.* Let $g = \delta|V'|$. Assume towards contradiction that the bad party controls $O(\epsilon^{0.51}|E|)$ vertices in $G'$, and we cannot identify $g$ truthful nodes.

**Claim 2.** *If the bad party controls $O(\epsilon^{0.51}|E|)$ vertices of graph $G'$, and we can't identify $g$ truthful node, then there exists a set $C$ of size $O(\epsilon^{0.51}|E|)$ and separates $V' \backslash C$ into sets $\{T'_i\}_{i=1,\cdots,\ell}$, each of size $|T'_i| \leq O(\epsilon^{0.51}|E|)$, and sets $\{A'_j\}_{j=1,\cdots,K}$, each of size $|A'_j| > \Omega(\epsilon^{0.51}|E|)$, and $|\cup_j^K A'_j| < g$.*

*Proof of Claim 2.* Since the corrupt party can control $G'$ with $O(\epsilon^{0.51}|E|)$ vertices, $m(G', g) \leq O(\epsilon^{0.51}|E|)$. By Lemma 5 $\min_k S_{G'}(k, g) + k \leq 2m(G', g) \leq O(\epsilon^{0.51}|E|)$. Let $k^* = \arg\min_k S_{G'}(k, g) + k$. Then $k^* \leq O(\epsilon^{0.51}|E|)$, $S_{G'}(k^*, g) \leq O(\epsilon^{0.51}|E|)$. By definition of $S_{G'}(k^*, g)$, there exists a set of size $S_G(k^*)$ after whose removal separates the remainder of the graph $G$ to connected components of size at most $k^*$ except for fewer than $g$ nodes. Thus components of size larger than $\Omega(\epsilon^{0.51}|E|)$ contain fewer than $g$ nodes. □

Let $T' = \cup_{i=1}^{\ell} T'_i$, $A' = \cup_{j=1}^{K} A'_j$. Since $|C| = O(\epsilon^{0.51}|E|) = O(\epsilon^{0.51}|\tilde{V}|)$, and $C \cup T' \cup A' = V'$, for small enough $\epsilon$, $|(T' \cup A') \cap \tilde{V}| \geq 9|\tilde{V}|/10$. From the assumption that we can't identify $g$ truthful nodes, $|A'| < g \leq |V'|/3$. Otherwise, we can claim the entire $A'$ as good and identify $g$ truthful nodes. Thus $|A' \cap \tilde{V}| \leq |V'|/3 \leq 2/3|\tilde{V}|$.[4]

Additionally, use the fact that $|T'_i \cap \tilde{V}| < |\tilde{V}|/10$ for every $i$, with sufficiently small $\epsilon$, we can merge various sets in $\{\{A_j\}_{j=1,\cdots,K}, \{T_i\}_{i=1,\cdots,\ell}\}$ and get two sets $V'_1$ and $V'_2$, such that $|V'_1 \cap \tilde{V}|, |V'_2 \cap \tilde{V}| \geq |\tilde{V}|/10$, and $V'_1$ and $V'_2$ are separated by $C$.

Now, let $V_1 \subseteq V$ (resp. $V_2 \subseteq V$) be the set of vertices $v \in V$ such that some copy of $v$ appears in $V'_1$ (resp. $V'_2$). Let $S \subseteq V$ be the set of vertices $v \in V$ such that all $r$ copies of $v$ appears in $C$. Since $|V'_1 \cap \tilde{V}|, |V'_2 \cup \tilde{V}| \geq |\tilde{V}|/10 = r|V|/10$, both $|V_1|, |V_2| \geq |V|/10$. Furthermore, we observe that $V_1 \cup V_2 \cup S = V$, which follows from $V'_1 \cup V'_2 \cup C = V'$. Now we can lower bound $|V_1 \cup V_2|$ as follows.

$$|V_1 \cup V_2| = |V \backslash S| \geq |V| - |C|/r \geq |V| - c\epsilon^{0.51}|E|/r = |V| - c\epsilon^{0.51}|V|$$

The first equality again follows from the fact that $V_1 \cup V_2 \cup S = V$, and that $V_1 \cup V_2$ is disjoint from $S$, and the second inequality follows by definition of $S$.

Since $V_1 \cup V_2$ is sufficiently large, we can find a balanced partition of $V_1 \cup V_2$ into sets $S_1 \subseteq V_1$, $S_2 \subseteq V_2$, $S_1 \cap S_2 = \emptyset, S_1 \cup S_2 = V_1 \cup V_2$, $|V|/10 \leq |S_1|, |S_2| \leq 9|V|/10$. From the property of $G$ that $E(S, V \setminus S) \geq \Omega(\sqrt{\epsilon}|E|)$ in Lemma 3 and the fact that $G$ is $d$-regular, we know that

[4] If we use the fact that $|A'| < g \leq (|V'| - \epsilon'|V|)/2$, for some constant $\epsilon'$, then $|A' \cap \tilde{V}| \leq (|V'| - \epsilon'|V'|)/2 \leq (1 - \epsilon')|\tilde{V}|$. We can merge $\{\{A_j\}, \{T_i\}\}$ to two sets $V'_1$, $V'_2$ such that $|V'_1 \cap \tilde{V}|, |V'_2 \cap \tilde{V}| \geq \epsilon'|\tilde{V}|$. The rest of the proof still goes through.

$$E(S_1, S_2) = E(S_1, V \backslash S_1) - E(S_1, S) \geq \alpha\sqrt{\epsilon}|E| - d(\epsilon^{0.51}|E|/r) = \alpha\sqrt{\epsilon}|E| - 2\epsilon^{0.51}|E| = \Omega(\sqrt{\epsilon}|E|),$$

for some constant $\alpha$. In the first equality, we use the fact that $S_1 \cup S_2 \cup S = V$, and $S_1, S_2, S$ are disjoint. Thus $E(S_1, V \backslash S_1) = E(S_1, S_2 \cup S) = E(S_1, S_2) + E(S_1, S)$.

Note that since $S_1 \subseteq V_1$ and $S_2 \subseteq V_2$, and $T_1'$ and $T_2'$ do not have edges between them in $G'$, the edges $E(S_1, S_2)$ all have to land as "edge vertices" in $C$. Formally, $E(S_1, S_2) \subseteq \tilde{E} \cap C$. In other words, for any $u \in S_1$, and $v \in S_2$, if $(u, v) \in E$, then the vertex $(u, v) \in V'$ has to be included in the set $C$, thus $|C| \geq \Omega(\sqrt{\epsilon}|E|)$.

This contradicts the fact that there are only $O(\epsilon^{0.51}|E|)$ vertices in $C$. □

Using Lemma 6 and Lemma 7, we can again obtain Theorem 6 for $0 < \delta < 1/2$, with the same argument for the proof of Theorem 2 in Section 2.3.2.

When $1/2 \leq \delta < 1$, we construct an auxiliary graph in the following way. Take as input any graph $G = (V, E)$. Let $h = \delta/(1-\delta)|V|$, construct $G' = G \cup h$-clique. Note $h = \delta|V'|$. Then, we claim that the critical number of bad nodes such that it is impossible to detect $\delta|V'|+1$ good nodes on $G'$ is the same as the critical number of bad nodes such that it is impossible to find one good node on $G$.

**Claim 3.** *Given any graph $G$, $1/2 \leq \delta < 1$ and $G'$ as constructed,*

$$m(G', \delta|V'|+1) = m(G).$$

*Proof.* Firstly, observe that

$$\delta|V'| = \delta(|V|+h) = \delta(|V|+\delta/(1-\delta)|V|) = h.$$

Therefore, one way to prevent identification of $\delta|V'|+1$ good nodes on $G'$ is to prevent identification of one good node on $G$. Since the $h$-clique is of size at least $|V'|/2$, and

report each other as good, they will be detected as good nodes. This strategy requires bad party to have budget $b = m(G)$. Thus $m(G', \delta|V'|+1) \leq m(G)$.

The direction $m(G', \delta|V'|+1) \geq m(G)$ follows by the fact that the strategy above is optimal. In order to prove this, we make the following observation:

**Claim 4.** *Given any graph $G$ and $g \leq |V|$,*

$$m(G) \leq m(G, g) + g - 1$$

*Proof of Claim 4.* One way to prevent identification of one good node is to corrupt $m(G, g)$ nodes plus the (at most) $g - 1$ detected good nodes. Call the set of the $g - 1$ or fewer detected nodes $S$. Notice that any node in $G \backslash S$ that is adjacent to $S$ are reported as bad by $S$. If not, this node has the same identity with $S$, and should be detected as good as well. Therefore, the bad party is able to corrupt the set $S$ without incurring any change in the reports, since all edges incident to $S$ now have both endpoints corrupt and so the reports are arbitrary. Previously, the set $S$ were good in any configuration of identities consistent with the reports and the budget. But now, the bad party's budget increases by at least $g - 1 \geq |S|$, and any configuration with the set $S$'s identity changed to all bad is also consistent with the reports.

Therefore, no node is good in all configurations, and so no node can be detected as good. This strategy requires $m(G, g) + g - 1$ nodes and prevents identification of one good node. Since $m(G)$ is the minimal number of bad nodes so that it is impossible to detect one good node, $m(G) \leq m(G, g) + g - 1$. □

Now we continue to prove the $m(G', \delta|V'|+1) \geq m(G)$ direction of Claim 3. Assume towards contradiction that there exists a strategy that controls at least one node in the $h$-clique, prevents identification of $h + 1$ good nodes, and requires fewer than $m(G)$ bad nodes in total. Suppose this strategy assigns $a$ nodes in the $h$-clique as bad, where $1 < a < m(G) \leq |V|/2 \leq h/2$. Then $h - a > h/2 > m(G) > b$. There-

fore, the rest of the $h$-clique forms a connected component with only good reports, and is of size $h - a$, which is larger than the bad party's budget $b < m(G)$, thus are declared as good. As a result, the bad party must prevent identification of $a+1$ good nodes in $G$ with budget strictly less than $m(G) - a$. This contradicts the fact that $m(G) - a \leq m(G, a) - 1 < m(G, a + 1)$ by Claim 4.

Therefore, the strategy of controlling $m(G)$ nodes on $G$ and let the $h$-clique be detected as good is an optimal strategy, $m(G', \delta|V'|+1) = m(G)$. □

Now, with Claim 3, we conclude that for any $1/2 \leq \delta < 1$, approximating $m(G, \delta|V|)$ within any constant must be SSE-hard. If not, we will obtain an efficient algorithm for approximating $m(G)$ by constructing a graph $G'$ by adding a $\frac{\delta}{1-\delta}|V|$ clique to any graph $G$, for some $\delta$, and approximate $m(G)$ by approximating $m(G', \delta|V'|+1)$, which is just $m(G', \delta'|V'|)$ for some other $0 < \delta' < 1$.

Theorem 6 implies a similar corollary about the SSE-hardness of seeding the nodes on a graph $G$ given any constant multiple of the critical number $m(G, \delta|V|)$ to prevent detection of any arbitrary fraction of good nodes.

**Corollary 2.** *Assume the SSE Hypothesis and $P \neq NP$. Fix any $\beta > 1$, and $0 < \delta < 1$. There does not exist a polynomial-time algorithm that takes as input an arbitrary graph $G = (V, E)$ and outputs a set of nodes $S$ with size $|S| \leq O(\beta \cdot m(G, \delta|V|))$, such that corrupting $S$ prevents the central agency from finding $\delta|V|$ truthful nodes.*

# Chapter 3

# A Geometric Model for Opinion Polarization

## 3.1 Introduction

Opinion polarization is a widely acknowledged social phenomenon, especially in the context of political opinions [33, 90, 51], leading to recent concerns over "echo chambers" created by mass media [82] and social networks [25, 75, 7, 6, 39]. The objective of this paper is to propose a simple, multi-dimensional geometric model of the dynamics of polarization where the evolution of correlations between opinions on different topics plays a key role.

Many models have been proposed to explain how polarization arises, and this remains an active area of research [73, 5, 72, 49, 65, 10, 27, 28, 59, 78, 86]. Our attempt aims at simplicity over complexity. As opposed to a large majority of previous works, our model does not require social network-based mechanism. Instead, we focus on influences of advertising or political *campaigns* that reach a wide segment of the population.

We develop a high-dimensional variant of *biased assimilation* [64] and use it as our main behavioral assumption. The bias assimilation for one topic states that people tend to be receptive to opinions they agree with, and antagonistic to opinions they disagree with.

The *multi-dimensional* setting reflects the fact that campaigns often touch on many topics. For example, in the context of American politics, one might wonder why there exists a significant correlation between opinions of individuals on, say, abortion access, gun rights and urgency of climate change [80]. Our model attempts to illustrate how such correlations between opinions can arise as a (possibly unintended) effect of advertising exploiting different topics and social values.

In mathematical terms, we consider a population of agents with preexisting opinions represented by vectors in $\mathbb{R}^d$. Each coordinate represents a distinct topic, and the value of the coordinate reflects the agent's opinion on the topic, which can be positive or negative. As discussed more fully in Section 3.1.4, we assume that all opinions lie on the Euclidean unit sphere. This reflects an assumption that each agent has the same "budget of importance" of different topics. We then consider a sequence of *interventions* affecting the opinions. An intervention is also a unit vector in $\mathbb{R}^d$, representing the set of opinions expressed in, e.g., an advertising campaign or "news cycle".

We model the effect of intervention $v$ on an agent's opinion $u$ in the following way. Supposing an agent starts with opinion $u \in \mathbb{R}^d$, after receiving an intervention $v$ it will update the opinion to the unit vector proportional to

$$w = u + \eta \cdot \langle u, v \rangle \cdot v \ , \tag{3.1}$$

where $\eta > 0$ is a global parameter that controls the influence of an intervention. Most of our results do not depend on a choice of $\eta$ and in our examples we often take $\eta = 1$ for the sake of simplicity. Smaller values of $\eta$ could model campaigns with limited persuasive power. This and other design choices are discussed more extensively in Section 3.1.4.

Intuitively, the agent evaluates the received message in context of its existing opinion, and assimilates this message weighted by its "agreement" with it. Our model exhibits biased assimilation in that if the intervening opinion $v$ is positively correlated with an agent's opinion $u$, then after the update the agent opinion moves towards $v$,

and conversely, if $v$ is negatively correlated with $u$, then the update moves $u$ away from $v$ and towards the opposite opinion $-v$.

One way to think of the intervention is as an exposure to persuasion by a political actor, like a political campaign message. A different way, in the context of marketing, is a product advertisement that exploits values besides the quality of the product. In that context, we can think of one of the $d$ coordinates of the opinion vector as representing opinion on a product being introduced into the market and the remaining coordinates as representing preexisting opinions on other (e.g., social or political) issues. Then, an intervention would be an advertising effort to connect the product with a certain set of opinions or values [93]. Some examples are corporate advertising campaigns supporting LGBT rights [91] or gun manufacturers associating their products with patriotism and conservative values [87]. Another scenario of an intervention is a company (e.g., a bank or an airline [37]) announcing its refusal to do business with the gun advocacy group NRA. Such advertising strategies can have a double effect of convincing potential customers who share relevant values and antagonizing those who do not.

Our main results show that such interventions, even if intending mainly to increase sales and without direct intention to polarize, can have a side effect of increasing the extent of polarization in the society. For example, it might be that, in a population with initial opinions distributed uniformly, a number of interventions introduces some weak correlations. In our model, these correlations can be profitably exploited by advertisers in subsequent interventions. As a side effect, the interventions strengthen the correlations and increase polarization.

For example, suppose that after various advertising campaigns, we observe that people who tend to like item A (say, electric cars) tend to be liberal, and people who like a seemingly unrelated item B (say, firearms) tend to be conservative. This may result from the advertisers exploiting some obvious connections, e.g., between electric cars and responding to climate change, and between firearms and respect for the military. Subsequently, future advertising efforts for electric cars may feature other values associated with liberals in America to appeal to potential consumers: an

advertisement might show a gay couple driving to their wedding in an electric car. Similarly, future advertisements for firearms may appeal to conservative values for similar reasons. The end result can be that the whole society becomes more polarized by the incorporation of political topics into advertisements.

Throughout the paper, we analyze properties of our model in a couple of scenarios. With respect to the interventions, we consider two scenarios: either there is one entity (an *influencer*) trying to persuade agents to adopt their opinion or there are two competing influencers pushing different agendas. With respect to the time scale of intervations, we also consider two cases: the influencer(s) can apply arbitrarily many interventions, i.e., the *asymptotic* setting, or they need to maximize influence with a limited number of interventions, i.e., the *short-term* setting. The questions asked are: (i) What sequence of interventions should be applied to achieve the influencer's objective? (ii) What are the computational resources needed to compute this optimal sequence? (iii) What are the effects of applying the interventions on the population's opinion structure? We give partial answers to those questions. The gist of them is that in most cases, applying desired interventions increases the polarization of agents.

### 3.1.1 Model definition

The formal definition of our model is simple. We consider a group of $n$ *agents*, whose opinions are represented by $d$-dimensional unit vectors, where each coordinate corresponds to a topic. We will look into how those opinions change after receiving a sequence of *interventions*. Each intervention is also a unit vector in $\mathbb{R}^d$, representing the opinion contained in a message that the influencer (e.g., an advertiser) broadcast to the agents. Our model features one parameter: $\eta > 0$, signifying how strongly an intervention influences the opinions.

The interventions $v^{(1)}, \ldots, v^{(t)}, \ldots$ divide the process into discrete time steps. Initially, the agents start with opinions $u_1^{(1)}, \ldots, u_n^{(1)}$. Subsequently, applying intervention $v^{(t)}$ updates the opinion of agent $i$ from $u_i^{(t)}$ to $u_i^{(t+1)}$.

After each intervention, the agents update their opinions by moving towards or away from the intervention vector, depending on whether or not they agree with it

(which is determined by the inner product between the intervention vector $v^{(t)}$ and the opinion vector), and normalizing suitably. The update rule is given by

$$u_i^{(t+1)} = \frac{w_i^{(t+1)}}{\left\| w_i^{(t+1)} \right\|} , \qquad \text{where} \qquad w_i^{(t+1)} = u_i^{(t)} + \eta \langle u_i^{(t)}, v^{(t)} \rangle \cdot v^{(t)} . \quad (3.2)$$

We note that, by expanding out the definition of $w_i^{(t+1)}$,

$$\|w_i^{(t+1)}\|^2 = \langle w_i^{(t+1)}, w_i^{(t+1)} \rangle = 1 + (2\eta + \eta^2) \langle u_i^{(t)}, v^{(t)} \rangle^2 \quad (3.3)$$

In particular, this implies that $\|w_i^{(t+1)}\| \geq 1$, and consequently that $u_i^{(t+1)}$ is well-defined. The norm in (3.2) and everywhere else throughout is the standard Euclidean norm. Note that applying $v^{(t)}$ or $-v^{(t)}$ to an opinion $u_i^{(t)}$ results in the same updated opinion $u_i^{(t+1)}$.

### 3.1.2 Example

To illustrate our model, let us consider an empirical example with $\eta = 1$. Suppose an advertiser is marketing a new product. The opinion of the population has four dimensions. The population consists of 500 agents, each with initial opinions $u_i^{(1)} = (u_{i,1}, u_{i,2}, u_{i,3}, 0) \in \mathbb{R}^4$ subject to $u_{i,1}^2 + u_{i,2}^2 + u_{i,3}^2 = 1$. The opinion on the new product is represented by the fourth coordinate, which is initially set to zero for all agents. These starting opinions are sampled independently at random from the uniform distribution on the sphere. A typical arrangement of initial opinions is shown under $t = 1$ in Figure 3-1.

Suppose the advertiser chooses to repeatedly apply an intervention that couples the product with the preexisting opinion on the first coordinate. More concretely, let the intervention vector be

$$v = (\beta, 0, 0, \alpha) , \qquad \text{where} \qquad \alpha = \frac{3}{4} , \beta = \sqrt{1 - \alpha^2} .$$

In that case, an application of the intervention $v$ to an opinion $u_i^{(1)} = (u_{i,1}, u_{i,2}, u_{i,3}, 0)$

results in $\langle u_i^{(1)}, v\rangle = \beta u_{i,1}$ and

$$u_i^{(2)} = \frac{w_i^{(2)}}{\|w_i^{(2)}\|}\ , \qquad w_i^{(2)} = ((1+\beta^2)u_{i,1}, u_{i,2}, u_{i,3}, \beta\alpha u_{i,1})\ , \qquad \|w_i^{(2)}\|^2 = 1 + 3\beta^2 u_{i,1}^2\ .$$

Note that after applying the intervention the first and last coordinates have the same sign. In subsequent time step, the intervention $v$ is applied again to the updated opinions $u_i^{(2)}$ and so on.

The evolution of opinions over five consecutive applications of $v$ in this process is illustrated in Figure 3-1. The interventions increase the affinity for the product for some agents while antagonizing others. Furthermore, they have a side effect of polarizing the agents' opinions also on the first three coordinates. A similar example is included in Appendix B.2.

### 3.1.3 Outline of our results

We analyze the strategy of influencers in several settings.

In an **"asymptotic scenario"**, the influencer wants to apply an infinite sequence of interventions $v^{(1)}, v^{(2)}, \ldots$, that maximizes how many out of the $n$ agent opinions converge to the target vector $v$. As is standard, we say that a sequence of vectors $u^{(1)}, \cdots, u^{(t)}, \ldots$ converges to a vector $v$ if $\lim_{t\to\infty}||u^{(t)} - v|| = 0$. One way to interpret this scenario is that a campaigner wants to establish a solid base of support for their party platform.

In a **"multiple-influencer scenario"**, two influencers (such as two companies or two parties) who have different objectives apply their two respective interventions on the population in a certain order. We ask how the opinions change under such competing influences. This scenario can be interpreted as two parties campaigning their agendas to the population.

In a **"short-term scenario"**, the influencer is advancing a product/subject which is expressed in the last coordinate of opinion vectors $u_{i,d}$. The influencer assumes some fixed threshold $0 < T < 1$ and an upper bound $K$ on the number of interventions, and asks, given $n$ opinions $u_1, \ldots, u_n$, how to choose $v^{(1)}, \cdots, v^{(K)}$ in order to maximize

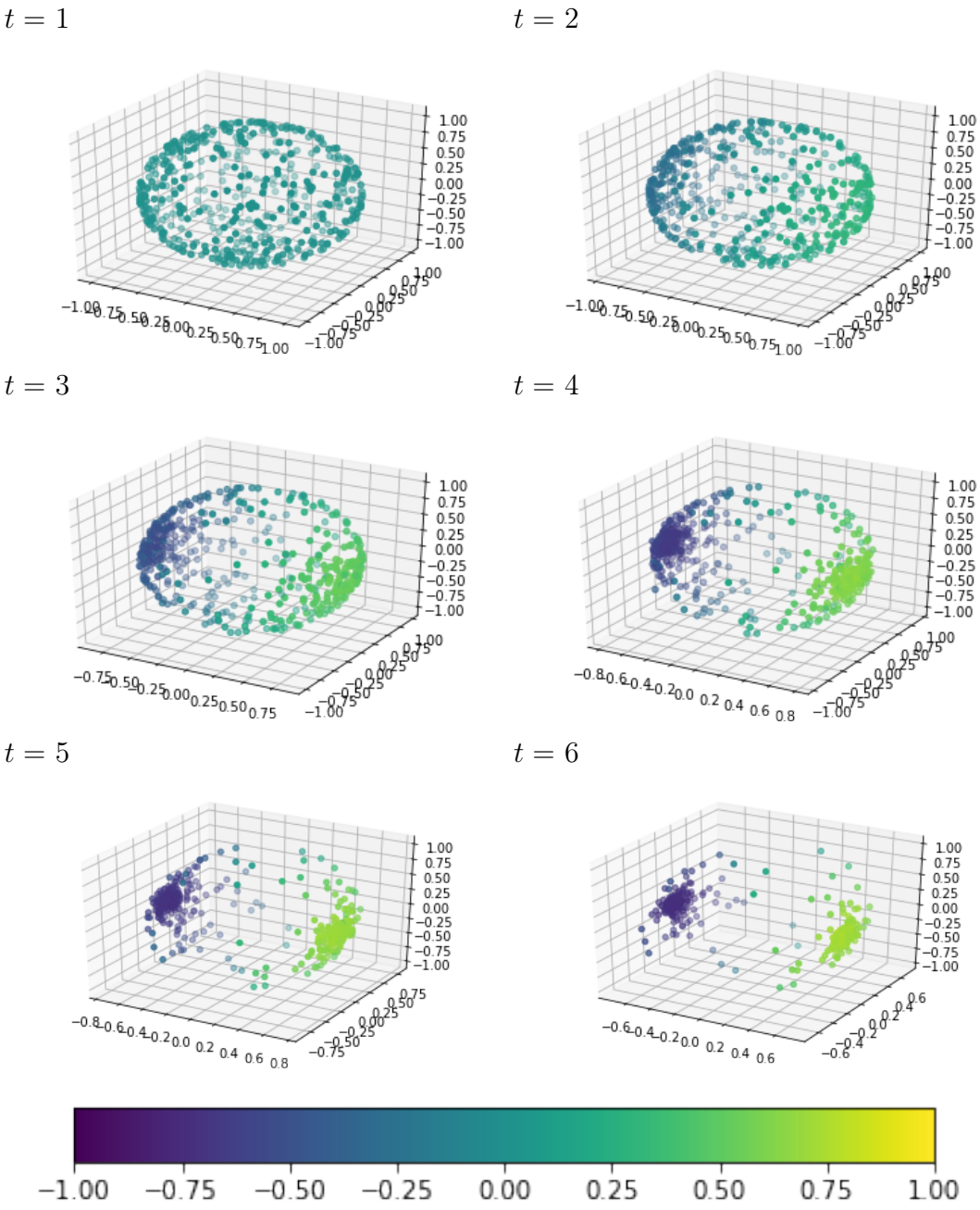


Figure 3-1: Graphical illustration of the example discussed in Section 3.1.2. Since we are working in $d = 4$, we illustrate the first three dimensions as spatial positions and the fourth dimension with a color scale. Initially the opinions are uniformly distributed on the sphere, with the fourth dimension equal to 0 (no opinion) everywhere. Consecutive applications of the intervention $v = (\sqrt{7}/4, 0, 0, 3/4)$ in $\mathbb{R}^4$ result in polarization both in spatial dimensions and in the color scale.

the number of time-$(K+1)$ opinions $u_1^{(K+1)}, \ldots, u_n^{(K+1)}$ with $u_{i,d}^{(K+1)} > T$. One interpretation is that advertisers only have a limited number of opportunities to publicize their products to consumers, and consumers with $u_i^{(K+1)} > T$ will decide to buy the product after the interventions $v^{(1)}, \cdots, v^{(K)}$ are applied.

We briefly summarize our results for these scenarios. In Section 3.3 we start by showing that random interventions lead to a strong form of polarization. More precisely, assuming uniformly distributed initial opinions, we prove that applying an independent uniformly random intervention at each time step leads the opinions to form two equally-sized clusters converging to a pair of (moving) antipodal points.

In Section 3.4 we consider the asymptotic scenario, where there is one influencer with a desired campaign agenda $v$ and unlimited numbers of interventions at its disposal. We ask which sequence of interventions maximizes the number of opinions that converge to the agenda $v$. Somewhat surprisingly, we show that such optimal strategy does not necessarily promote the campaign agenda directly at every step. Instead, it finds a hemisphere containing the largest number of initial opinions, concentrates the opinions in this hemisphere around an arbitrary point, and only in the last stage nudges them gradually towards the target agenda. We then show that it is computationally hard to approximate this densest hemisphere (and therefore the optimal strategy) to any constant factor. Again, strong polarization emerges from our dynamic: there exists a pair of antipodal points such that all opinions converge to one of them.

In Section 3.5 we study the short-term scenario where one influencer is allowed only one intervention. In Section 3.5.1, we describe a case study with one influencer and two agents in the population. We assume that the influencer wants to increase the correlations of agent opinions with the target opinion $v$ above a given threshold $T > 0$. We show consequences of optimal interventions depending on if the influencer can achieve this objective for one or both agents. In Section 3.5.2, we consider a similar scenario, but with a large number of agents. In that case, it surprisingly turns out that the problem of finding optimal intervention in this short-term setting is related to the problem analyzed in the asymptotic setting. Finding the optimal

intervention is equivalent to finding a spherical cap containing the largest number of initial opinions.

In Section 3.6, we study two competing influencers. At each time step, one of the influencers is selected at random to apply its intervention. One might hope that having multiple advertisers can make the resulting opinions more spread-out, but we prove that this not the case. We show that, as time goes to infinity, all opinions converge to the convex cone between the two intervention vectors. Furthermore, we show that the if the correlation between the interventions is high enough, the strong form of polarization emerges: the opinions of the population concentrate around two antipodes moving around in the convex cones of the two interventions.

### 3.1.4 Design choices

Our goal in this work is to provide a simple, elegant and analyzable model demonstrating how correlations between different topics and natural interventions lead to polarization. That being the case, there are many societal mechanisms related to polarization that we do not discuss here.

First, in contrast to majority of existing literature, we present a mechanism independent from opinion changes induced by interactions between individuals. Second, we do not address aspects such as replacement of the population or unequal exposure and effects of the interventions. We do not consider any external influences on the population in addition to the interventions. Our model does not align with (limited) theoretical and empirical research suggesting that in certain settings exposure to conflicting views can decrease polarization [79, 68, 41, 40] or works that question the overall extent of polarization in the society [34, 11].

In general we assume that the influencers have full knowledge of the agent opinions. This is not a realistic assumption and in fact our results in Section 3.4 show that in some settings the optimal influencer strategy is infeasible to compute even with the full knowledge of opinions. On the other hand, we observe polarization also in settings where the influencers apply interventions that are agnostic to the opinions, for example with purely random interventions in Section 3.3 or competing influencers

in Section 3.6.

We sometimes discuss the uniform distribution of initial opinions on $\mathbb{R}^d$. We do this as the uniform distribution may be viewed as the most diverse and establishing polarization starting from the uniform distribution hints that we are modelic a generic phenomenon. Most of our results do not make assumptions about the initial distribution.

We assume that any group of topics can be combined into an intervention with the effect given by (3.1). A more plausible model might feature some "internal" (content) correlations between topics in addition to "external" (social) correlations arising out of the agents' opinion structure. For example, topics may have innate connections, causing inherent correlations between corresponding opinions (e.g., being positive on renewable energy and recycling). Furthermore, there are certain topics (e.g., undesirability of murder) on which nearly all members of the population share the same inclination. As a matter of fact, it is common for marketing strategies to exploit unobjectionable social values (see, e.g., [93]). However, we presume that under suitable circumstances (e.g., due to inherent correlations we just mentioned) the "polarizing" topics might present a more appealing alternative for a campaign. Our model concerns such a case, where the "unifying" topics might be excluded from the analysis. We note that other works have also suggested that focusing on polarizing topics may be appealing for campaigns [78].

Below we discuss a couple of specific design choices in more detail:

**Euclidean unit ball** We make an assumption that all opinions and interventions lie on the Euclidean unit ball. Note that the interpretation of this representation is somewhat ambiguous. The magnitude of an opinion on a given subject $u_{i,k}$ might signify the strength of the opinion, the confidence of the agent or relative importance of the subject to the agent. While these are different measures, there are psychological reasons to expect that, e.g., "issue interest" and "extremity of opinion" are correlated [62, 9, 10]. Especially taking the magnitudes as signifying the relative importance, we believe that the assumption that this "budget of importance" for any given agent

is fixed is quite natural. That being said, we are also motivated by simplicity and tractability.

Multiple ways of relaxing or modifying this assumption are possible. While we do not study these variants in this paper, we now discuss them very briefly. At least empirically, our basic findings about ubiquity of polarization seem to remain valid for those modified models.

Perhaps the simplest modification is to use the same update rule as in (3.2) with a different norm (eg., $\ell_1$ or $\ell_\infty$ norm). Such variant would also assume that opinions and interventions lie on the unit sphere of the respective norm. Our experiments suggest that, qualitatively, both $\ell_1$ and $\ell_\infty$ variants behave similarly to the Euclidean norm.

In another direction, rather than having all opinions on the unit sphere, fixed, but different norms $z_i$ can be specified for different agents. Then, the update rule (3.2) could be modified as

$$w_i^{(t+1)} = u_i^{(t)} + \eta \cdot \left\langle \frac{u_i^{(t)}}{z_i}, v^{(t)} \right\rangle \cdot v^{(t)} \ ,$$

with normalization preserving $\|u_i^{(t+1)}\| = z_i$. As long as the values of $z_i$ are bounded from below and above, the resulting dynamic is essentially identical and our results carry over to this more general setup.

Yet another possibility is to consider opinion unit vectors $u \in \mathbb{R}^{d+1}$ with $u_{d+1} \geq 0$ and interpret the first $d$ coordinates as opinions and the last coordinate as "unused budget". Therefore, large values of $u_{d+1}$ signify generally uncertain opinions and small values of $u_{d+1}$ correspond to strong opinions. There are multiple possible rules for interventions, where an intervention can have $d$ or $d+1$ coordinates, and with different treatments of the last coordinate. We leave the details for another time.

**Effects of applying $v$ and $-v$** In our model, an effect of an intervention $v$ is exactly the same as for the opposite intervention $-v$. This might look like a cynical assumption about human nature, but arguably it is not entirely inaccurate. For

example, experiments on social media show that not only exposure to similar ideas (the "echo chamber" effect), but also exposure to opposing opinions causes beliefs to become more polarized [6]. This is even more apparent if a broader notion of an intervention is considered. Using a recent example, social media platforms banning or disassociating from certain statements can have a polarizing effect [12]. Furthermore, in our model this effect occurs only if all the components of an opinion are negated.

A related, more general objection is that direct persuasion is not possible in our model. If an agent has an opinion $u$ with $\langle u, v\rangle < 0$, directly applying $v$ only makes the situation worse. Instead, an effective influencer needs to apply interventions utilizing different subjects to gradually move $u$ through a sequence of intermediate positions towards $v$. Our answer is that we posit that a lot of, if not all, persuasion actually works that way: to convince that "$x$ is good", one argues that "$x$ is good, since it is quite like $y$, which we both already agree is good".

**Notions of polarization** While the notion of polarization is clear when discussing one topic, it is not straightforward to interpret in higher dimensions. Let $S \subseteq \mathbb{R}^d$ be a set of $n$ opinions. Writing $u = (u_1, \ldots, u_d)$ for $u \in S$, a natural measure of polarization of $S$ on a single topic $i$ is

$$\rho_i(S) = \frac{1}{|S|^2} \max_{T \subset S} \sum_{u \in T, u' \in S \setminus T} (u_i - u'_i)^2,$$

and we may generalize it to higher dimensions by measuring the polarization as:

$$\rho(S) = \frac{1}{|S|^2} \max_{T \subset S} \sum_{u \in T, u' \in S \setminus T} \|u - u'\|^2.$$

It is clear from the definition that

$$\max_i \rho_i(S) \leq \rho(S) \leq \sum_i \rho_i(S).$$

If we consider an example set $S_1$ with $n/2$ opinions at $u$ and $n/2$ opinions at $-u$, then clearly $\rho(S_1) = \sum_i \rho_i(S_1)$, but in any other example, the upper bound will

not be tight. For example, if $S_2$ is the set of the $2^d$ vertices of a hypercube, i.e., $S_2 = 1/\sqrt{d} \cdot \{-1, 1\}^d$, then $\rho_i(S_2) = 1/d$ for all $i$, but $\rho(S_2)$ converges to $1/2$ as $n \to \infty$. This corresponds to the fact that while the society is completely polarized on each topic, two random individuals will agree on about half of the topics. In Section 3.2 we refer to such a situation as exhibiting *issue radicalization*, but no *issue alignment*.

Ultimately, in many of our results we do not worry about these issues, since we observe a strong form of polarization, where all opinions converge to two antipodal points.

### 3.1.5 Other variants

Other than discussed above, there are many possible variants that can lead to interesting future work. These include:

- "Targeting", where the influencer can select subgroups of the population and apply interventions groupwise.
- Models where the strength of an intervention $\eta$ varies across agents and/or time steps.
- Perturbing preferences with noise after each step.
- Replacement of the population, e.g., introducing new agents with "fresh" opinions or removing agents that stayed in the population for a long time or who already "bought" the product, i.e., exceeded the threshold $u_{i,d} > T$. For example, this could correspond to "one-time" purchase product like a house or a fridge, or situations where the customer's opinion is more difficult to change as time passes.
- Models where the initial opinions are not observable or partially observable.
- Expanding the model by adding peer effects and social network structure and exploring the resulting dynamics of polarization and opinion formation. This

can be done in different ways and we expect that polarization will feature in many of them. For example, [38] show polarization for random interventions in what they term the "party model".

- Strategic competing influencers: in the studied scenarios with competing influencers, we assume that they apply fixed interventions. One can ask: suppose the influencers have their own target opinions, what is each campaigner's optimal sequence of messages in face of the other campaigner? Then, resulting equilibrium of opinion formation could be analyzed. This can be modeled as a dynamic game where the game state is the opinion configuration and optimal strategies may be derived using sequential planning and control.

## 3.2 Related works

As mentioned, there is a multitude of modeling and empirical works studying opinion polarization in different contexts [73, 5, 8, 72, 49, 65, 70, 10, 27, 28, 59, 86, 78, 6]. Broadly speaking, previous works have proposed various possible sources for polarization, including peer interactions, bias in individuals' perceptions, and global information outlets.

There is an extensive line of models of opinion exchange on networks with peer interactions, where individuals encounter neighboring individuals' opinions and update their own opinions based on, e.g., pre-defined friend/hostile relations [89], or the similarity and relative strength of opinions [68], etc. This branch of work often attributes polarization to homophily of one's social network [27] that is induced by the self-selective nature of social relations and segregation of like-minded people [95] and exacerbated by the echo chamber effect of social media [75].

A parallel proposed mechanism points to psychological biases in individuals' opinion formation processes. One example is biased assimilation [64, 27, 10, 6]: the tendency to reinforce one's original opinions regardless if other encountered opinions align with them or not. For example, [6] observed that even when social media users are assigned to follow accounts that share opposing opinions, they still tend to hold

their old political opinions and often to a more extreme degree. On the modeling side, [27] showed that DeGroot opinion dynamics with the biased assimilation property on a homophilous network may lead to polarization.

Existing works have also proposed models where polarization occurs even when information is shared globally [97, 70]. For example, [70] propose a model where competition for readership between global information outlets causes news to become polarized in a single-dimensional setting. Another example is [97], a classical work on the formation of mass opinion. It theorizes that each individual has political dispositions formed in their own life experience, education and previous encounters that intermediate between the message they encounter and the political statement they make. Therefore, hearing the same political message can cause different thinking processes and changes in political preferences in different individuals.

It is noteworthy that the majority of previous work focuses on polarization on a single topic dimension. Two exceptions are [10], which studies biased assimilation with opinions on multiple topics and [11] that observed non-trivial correlations between people's attitudes on different issues. We note that [10] uses a different updating rule to observe dynamics that differ from our work: in their simulations, polarization on one issue typically does not result in polarization on others. There is also a class of models [5, 72, 65] that concern multi-dimensional opinions where an opinion on a given topic takes one of finitely many values (e.g., $+$ or $-$). These models do not seem to have a geometric structure of opinion space similar to ours and usually focus on formation of discrete groups in the society rather than total polarization. Another model [76] uses a geometric (affine) rule of updating multi-dimensional opinions. Unlike us, they seem to be modeling pre-existing, "intrinsic" correlations between topics rather than the emergence of new ones and they are concerned mostly with convergence and stability of their dynamics.

A related paper [78] contains a geometric model of opinion (preference) structures. Both this and our model propose mechanisms through which information outlets acting for their own benefit can lead to increased disagreement in the society. The key difference to our model is that their population's preferences are static and do

not update, but the outlets are free to choose what information to offer to their customers. By contrast, in our model, the influencers have pre-determined ideologies and compete to align agents' opinions with their own. In other words, [78] focuses on modeling of competitive information acquisition, and our paper on modeling the influence of marketing on the public opinion.

Our model suggests that under the conditions of biased assimilation, opinion manipulation by one or several global information outlets can unintentionally lead to a strong form of polarization in multi-dimensional opinion space. Not only do people polarize on individual issues, but also their opinions on previously unrelated issues become correlated. This form of polarization is known as *issue alignment* [11] in political science and sociology literature. Issue alignment refers to an opinion structure where the population's opinions on multiple (relatively independent) issues correlate. It is related to *issue radicalization*, where the opinions polarize for each issue separately. Compared to issue radicalization, issue alignment is theorized to pose more constraints on the opinions an individual can take, resulting in polarized and clustered mass opinions even when the public opinions are not extreme in any single topic, and presenting more obstacles for social integration and political stability [11]. In light of this, one way to view our model is as a mathematical mechanism by which this strong form of polarization can arise and worsen due to companies', politicians', and the media's natural attempts to gain support from the public.

On the more technical side, we note that our update equation bears similarity to Kuramoto model [53] for synchronization of oscillators on a network in the control literature. In this model, each oscillator $i$ is associated with the point $\theta_i$ on the two-dimensional sphere, and $i$ updates its point continuously as a function of its neighbors' points $\theta_j$:

$$\dot{\theta}_i = \omega_i + \frac{K}{N} \sin(\theta_j - \theta_i),$$

where $K$ is the *coupling strength* and $N$ is the number of nodes in the network. In two dimensions, our model can be compared to Kuramoto model with $\omega_i = 0$ on

a star graph, with the influencers at the center of the star connected to the entire population, where the influencers' opinions do not change and the update strength is qualitatively similar to $\sin((\theta_v - \theta_u)/2)$ (see (3.15)). However, we note a crucial difference: in the Kuramoto dynamic, $\theta_i$ always moves towards $\theta_j$, i.e. nodes always move towards synchronization, but in our dynamic, opinions $\theta_i$ are allowed to move further away from $\theta_j$ when the angle between their opinions are obtuse. In addition, the central node in our model can be strategic in choosing its positions, while the central node in Kuramoto model follows the synchronization dynamics of the system. We think this property provides a better model for opinion interactions.

**Subsequent work** A work by Gaitonde, Kleinberg and Tardos [38], announced after we posted the preprint of this paper, proposes a framework that generalizes our random interventions scenario from Section 3.3. They prove several interesting results, including a strong form of polarization under random interventions in some related models. They also shed more light on the scenario of dueling influencers from Section 3.6, showing that in case the dueling interventions are orthogonal, the resulting dynamics exhibits a weaker kind of polarization.

## 3.3 Asymptotic scenario: random interventions polarize opinions

In this section, we analyze the long-term behavior of our model in a simple random setting. We assume that, for given dimension $d$ and parameter $\eta$, at the initial time $t = 1$ we are given $n$ opinion vectors $u_1^{(1)}, \ldots, u_n^{(1)}$. Subsequently, we sample a sequence of interventions $v^{(1)}, v^{(2)}, \ldots$, each $v^{(t)}$ iid from the uniform distribution on the unit sphere $S^{d-1}$. At time $t$ we apply the random intervention $v^{(t)}$ to every opinion vector $u_i^{(t)}$, obtaining a new opinion $u_i^{(t+1)}$.

We want to show that the opinions $\{u_i^{(t)}\}$ almost surely polarize as time $t$ goes to infinity. We need to be careful about defining the notion of polarization: since the interventions change at every time step, the opinions cannot converge to a fixed

vector. Instead, we show that for every pair of opinions the angle between them converges either to 0 or to $\pi$. More formally:

**Theorem 7.** *Consider the model of iid interventions described above for some $d \geq 2$, $\eta > 0$ and initial opinions $u_1^{(1)}, \ldots, u_n^{(1)}$. For any $1 \leq i < j \leq n$ and $t \to \infty$,*

$$\Pr\left[\|u_i^{(t)} - u_j^{(t)}\| \to 0 \vee \|u_i^{(t)} + u_j^{(t)}\| \to 0\right] = 1 \,.$$

This leads to a corollary which follows by applying the union bound (with probability 0 in each term) for each pair of opinions $u_i^{(t)}, u_j^{(t)}$:

**Corollary 3.** *For any $d \geq 2$, $\eta > 0$, initial opinions $u_1^{(1)}, \ldots, u_n^{(1)}$ and a sequence of uniform iid interventions, almost surely, there exists $J \subseteq \{1, \ldots, n\}$ such that the diameter of the set*

$$\left\{(-1)^{\mathbb{1}[i \in J]} \cdot u_i^{(t)} : i \in \{1, \ldots, n\}\right\}$$

*converges to zero.*

**Remark 4.** *Consider initial opinions of $n$ agents that are independently sampled from a distribution $\Gamma$ that is symmetric around the origin, in the sense that $\Gamma(-A) = \Gamma(A)$ for every set $A \subseteq S^{d-1}$. Then, with high probability, the opinions converge to two polarized clusters of size roughly $n/2$. Indeed, consider sampling $n$ independent vectors $u_1, \ldots, u_n$ from $\Gamma$ and $n$ independent signs $\sigma_1, \ldots, \sigma_n \in \{\pm 1\}$. Then $\sigma_1 u_1, \ldots, \sigma_n u_n$ are independent samples from $\Gamma$. Moreover, if the sizes of the two clusters for $u_1, \ldots, u_n$ are $r$ and $n - r$ then the size of each cluster for $\sigma_1 u_1, \ldots, \sigma_n u_n$ is distributed according to $\mathrm{Bin}(r, 1/2) + \mathrm{Bin}(n - r, 1/2) = \mathrm{Bin}(n, 1/2)$ (this is due to the observation that $u_i$ and $-u_i$ converge to the opposite clusters).*

**Remark 5.** *For simplicity we do not elaborate on this later, but we note that, both empirically and theoretically, the convergence in our results is quite fast. This concerns Theorem 7, as well as the results presented in the subsequent sections.*

We now proceed to the proof of Theorem 7:

### 3.3.1 Notation and main ingredients

Before we proceed with explaining the proof, let us make a general observation that we will use frequently. Let $d \geq 2$ and $\eta > 0$ and let $f : S^{d-1} \times S^{d-1} \to S^{d-1}$ be the function mapping an opinion $u$ and an intervention $v$ to an updated opinion $f(u, v)$, according to (3.2) and (3.3). It should be clear that this function is invariant under isometries: namely, for any real unitary transformation $A : S^{d-1} \to S^{d-1}$ we have

$$f(Au, Av) = Af(u, v) \; . \tag{3.4}$$

In our proofs we will be often using (3.4) to choose a convenient coordinate system.

Let us turn to Theorem 7. Again, let $d \geq 2$ and $\eta > 0$. Without loss of generality we will consider only two starting opinions called $u_1^{(1)}$ and $u_2^{(1)}$. To prove Theorem 7, we need to show that almost surely one of the vectors $u_1^{(t)} - u_2^{(t)}$ and $u_1^{(t)} + u_2^{(t)}$ vanishes.

We proceed by using martingale convergence. Specifically, let

$$\alpha_t := \arccos\langle u_1^{(t)}, u_2^{(t)} \rangle \; .$$

That is, $0 \leq \alpha_t \leq \pi$ is the primary angle between $u_1^{(t)}$ and $u_2^{(t)}$.

The proof rests on two claims. First, $\alpha_t$ is a martingale:

**Claim 5.** $\mathrm{E}[\alpha_{t+1} \mid \alpha_t] = \alpha_t$.

Second, we show a property which has been called "variance in the middle" [20]:

**Claim 6.** *For every $\varepsilon > 0$, there exists $\delta > 0$ such that,*

$$\varepsilon \leq \alpha_t \leq \pi/2 \implies \Pr\left[\alpha_{t+1} < \alpha_t - \delta \mid \alpha_t\right] > \delta \; , \tag{3.5}$$

*and, symmetrically,*

$$\pi/2 \leq \alpha_t \leq \pi - \varepsilon \implies \Pr\left[\alpha_{t+1} > \alpha_t + \delta \mid \alpha_t\right] > \delta \; . \tag{3.6}$$

These two claims imply Theorem 7 by standard tools from the theory of martin-

gales (eg., [94]):

*Claims 5 and 6 imply Theorem 7.* As a consequence of applying Claim 6 $\lceil \pi/\delta \rceil$ times, we obtain that for every $\varepsilon > 0$ there exist $k_0 \in \mathbb{N}$ and $\eta < 1$ such that

$$\varepsilon \leq \alpha_t \leq \pi - \varepsilon \implies \Pr\left[\forall 1 \leq k \leq k_0 : \varepsilon \leq \alpha_{t+k} \leq \pi - \varepsilon \mid \alpha_t\right] \leq \eta \,.$$

Subsequently, it follows that for any fixed $\varepsilon > 0$ and $T \in \mathbb{N}$,

$$\Pr\left[\forall t \geq T : \varepsilon \leq \alpha_t \leq \pi - \varepsilon\right] = 0 \,. \tag{3.7}$$

By Claim 5, the sequence of random variables $\alpha_t$ is a bounded martingale and therefore almost surely converges. Accordingly, let $\alpha^* := \lim_{t\to\infty} \alpha_t$. We now argue that $\Pr[0 < \alpha^* < \pi] = 0$. To that end,

$$\begin{aligned}
\Pr[0 < \alpha^* < \pi] &\leq \sum_{s=1}^{\infty} \Pr\left[\frac{1}{s} < \alpha^* < \pi - \frac{1}{s}\right] \leq \sum_{s=1}^{\infty} \Pr\left[\exists T : \forall t \geq T : \frac{1}{2s} < \alpha_t < \pi - \frac{1}{2s}\right] \\
&\leq \sum_{s=1}^{\infty} \sum_{T=1}^{\infty} \Pr\left[\forall t \geq T : \frac{1}{2s} < \alpha_t < \pi - \frac{1}{2s}\right] = 0 \,,
\end{aligned}$$

where we applied (3.7) in the last line. Hence, almost surely, either $\alpha^* = 0$, which is equivalent to $\|u_1^{(t)} - u_2^{(t)}\| \to 0$ or $\alpha^* = \pi$, equivalent to $\|u_1^{(t)} + u_2^{(t)}\| \to 0$. □

In the subsequent sections we proceed with proving Claims 5 and 6. In Section 3.3.2 we prove Claim 5 for $d = 2$. In Section 3.3.3 we show the same claim for $d \geq 3$ by a reduction to the case $d = 2$. Finally, in Section 3.3.4 we use a continuity argument to prove Claim 6.

In the following proofs, we fix $d$, $\eta$, time $t$ and the opinions of two agents at that time. For simplicity, we will denote the relevant vectors as $u := u_1^{(t)}$, $u' := u_2^{(t)}$ and $v := v^{(t)}$.

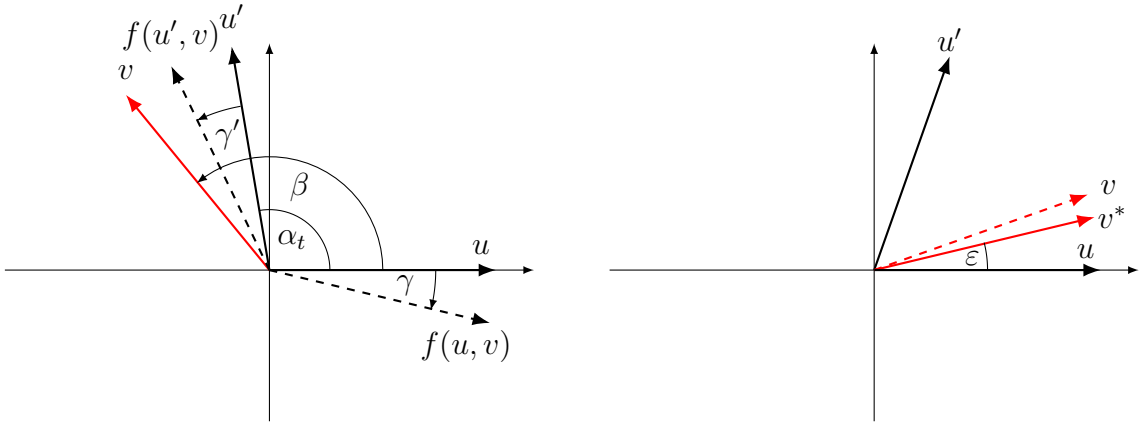


Figure 3-2: On the left an illustration of the vectors and angles in the proof of Claim 5. On the right an illustration for the proof of Claim 6.

### 3.3.2 Proof of Claim 5 for $d = 2$

It follows from (3.4) that we can assume wlog that $u = (1, 0)$ and $u' = (\cos \alpha_t, \sin \alpha_t)$ (recall that by definition $0 \leq \alpha_t \leq \pi$ holds). Let us write the random intervention vector as $v = (\cos \beta, \sin \beta)$, where the distribution of $\beta$ is uniform in $[0, 2\pi)$. We will also write (cf. Figure 3-2 for an overview)

$$f(u, v) = (\cos \gamma, \sin \gamma) \,, \qquad f(u', v) = (\cos(\alpha_t + \gamma'), \sin(\alpha_t + \gamma')) \,, \qquad \gamma, \gamma' \in [-\pi, \pi) \,.$$

Note that $\gamma$ is a function of the intervention angle $\beta$, and it should be clear that $\gamma(\beta) = -\gamma(-\beta)$. Accordingly, the distribution of $\gamma$ is symmetric around zero and in particular $\mathrm{E}\,\gamma = 0$ (where the expectation is over $\beta$). Applying (3.4), it also follows $\mathrm{E}\,\gamma' = 0$.

Let $\widehat{\alpha} := \alpha_t + \gamma' - \gamma$. Since $\widehat{\alpha}$ is equal to the directed angle from $f(u, v)$ to $f(u', v)$, one might think that we have just established $\mathrm{E}[\alpha_{t+1} \mid \alpha_t] = \alpha_t$. However, recall that we defined $\alpha_{t+1}$ to be the value of the (primary) *undirected* angle between $f(u, v)$ and $f(u', v)$. In particular, it holds that $0 \leq \alpha_{t+1} \leq \pi$, but we cannot assume that about $\widehat{\alpha}$. On the other hand, it is clear that if $0 \leq \widehat{\alpha} \leq \pi$, then indeed $\alpha_{t+1} = \widehat{\alpha}$. Therefore, in the following we will show that $0 \leq \widehat{\alpha} \leq \pi$ always holds, which implies $\mathrm{E}[\alpha_{t+1} \mid \alpha_t] = \mathrm{E}[\widehat{\alpha} \mid \alpha_t] = \alpha_t + \mathrm{E}\,\gamma' - \mathrm{E}\,\gamma = \alpha_t$.

To that end, we start with showing a weaker bound $-\pi < \widehat{\alpha} < 2\pi$. To see

this, we first establish that $-\pi/2 < \gamma, \gamma' < \pi/2$. The argument for $\gamma$ is as follows: if $\langle u, v\rangle \geq 0$, then $f(u, v)$ is a convex combination of $u$ and $v$. Therefore, an intervention cannot move $f(u, v)$ away from $u$ by an angle of more than $\pi/2$. If $v \neq u$, then also $f(u, v) \neq v$, so in fact the angle must be strictly less, that is $-\pi/2 < \gamma < \pi/2$. If $\langle u, v\rangle < 0$, then $-\pi/2 < \gamma < \pi/2$ follows from the same argument applied to $-v$ (since the effect of both interventions is the same). Finally, $-\pi/2 < \gamma' < \pi/2$ holds by (3.4) and the same proof. Since we know $0 \leq \alpha_t \leq \pi$, we obtain $-\pi < \widehat{\alpha} < 2\pi$.

Since we know $-\pi < \widehat{\alpha} < 2\pi$, the inequality $0 \leq \widehat{\alpha} \leq \pi$ is equivalent to $\sin \widehat{\alpha} \geq 0$. Geometrically, this property means that the ordered pair of vectors $(u, v)$ has the same orientation as the pair $(f(u, v), f(u', v))$. To avoid case analysis, we prove this claim by a calculation:

**Claim 7.** $\sin \widehat{\alpha} \geq 0$.

*Proof.* We defer the proof to Appendix B.1. □

### 3.3.3 Proof of Claim 5 for $d \geq 3$

In this case we will write the random intervention vector as $v = v^{\|} + v^{\perp}$ where $v^{\|}$ is projection of $v$ onto the span of $u$ and $u'$. In particular, $v^{\|}$ and $v^{\perp}$ are orthogonal. We will now prove a stronger claim $\mathrm{E}\left[\alpha_{t+1} \mid \alpha_t, \|v^{\|}\|\right] = \alpha_t$.

Accordingly, condition on the value $\|v^{\|}\| = R$. Observe that, by symmetry, vector $v^{\|}$ is distributed uniformly in the two-dimensional space $\mathrm{span}\{u, u'\}$ among vectors of norm $R$. In other words, we can write $v^{\|} = RV$, where $V$ is a uniform two-dimensional unit length vector.

Denote the non-normalized vectors after intervention as

$$\widehat{u} := u + \eta\langle u, v^{\|}\rangle(v^{\|} + v^{\perp}) \,, \qquad \widehat{u}' := u' + \eta\langle u', v^{\|}\rangle(v^{\|} + v^{\perp}) \,.$$

Let $c := 2\eta + \eta^2$. We proceed with calculations:

$$\begin{aligned}
\langle \widehat{u}, \widehat{u}' \rangle &= \langle u, u' \rangle + c\langle u, v^{\parallel} \rangle \langle u', v^{\parallel} \rangle = \langle u, u' \rangle + cR^2 \langle u, V \rangle \langle u', V \rangle \ , \\
\|\widehat{u}\|^2 &= 1 + c\langle u, v^{\parallel} \rangle^2 = 1 + cR^2 \langle u, V \rangle \ , \\
\|\widehat{u}'\|^2 &= 1 + c\langle u', v^{\parallel} \rangle^2 = 1 + cR^2 \langle u', V \rangle \ .
\end{aligned}$$

Note that all these formulas are valid also for $d = 2$, with the only difference that $R = 1$ holds deterministically in that case.

Since $c(\eta) = 2\eta + \eta^2$ is a bijection on $\mathbb{R}_{>0}$, there exists $\widehat{\eta} > 0$ such that $cR^2 = 2\widehat{\eta} + \widehat{\eta}^2$. Accordingly, for any $d \geq 3$ and $\eta > 0$, the joint distribution of $\langle \widehat{u}, \widehat{u}' \rangle$, $\|\widehat{u}\|$ and $\|\widehat{u}'\|$ conditioned on $\alpha_t = \arccos(\langle u, u' \rangle)$ and $\|v^{\parallel}\| = R$ is the same as their joint distribution for $d = 2$ and $\widehat{\eta}$, conditioned on the same value of $\alpha_t$.

Since $\alpha_{t+1} = \arccos\left(\frac{\langle \widehat{u}, \widehat{u}' \rangle}{\|\widehat{u}\| \|\widehat{u}'\|}\right)$, the same correspondence holds for the distribution of $\alpha_{t+1}$ conditioned on $\alpha_t$ and $\|v^{\parallel}\| = R$. Therefore, $\mathrm{E}\left[\alpha_{t+1} \mid \alpha_t, \|v^{\parallel}\| = R\right] = \alpha_t$ follows by Claim 5 for $d = 2$, which we already proved. □

### 3.3.4 Proof of Claim 6

Again we use (3.4) to choose a coordinate system and assume wlog that $u = (1, 0, \ldots, 0)$ and $u' = (\cos \alpha_t, \sin \alpha_t, 0, \ldots, 0)$. Our objective is to show that, with probability at least $\delta$, we will have $\alpha_{t+1} - \alpha_t > \delta$ (in case $\alpha_t \leq \pi/2$) or $\alpha_{t+1} - \alpha_t < -\delta$ (in case $\alpha_t \geq \pi/2$). To start with, we will show that by symmetry we need to consider only the first case $\alpha_t \leq \pi/2$.

Note that the intervention function $f$ exhibits a symmetry $f(-u, v) = -f(u, v)$. Furthermore, we also have $\arccos\langle u, u' \rangle = \pi - \arccos\langle u, -u' \rangle$. Consequently,

$$\begin{aligned}
\alpha_{t+1} - \alpha_t &= \arccos\langle f(u, v), f(u', v) \rangle - \arccos\langle u, u' \rangle \\
&= \pi - \arccos\langle f(u, v), f(-u', v) \rangle - (\pi - \arccos\langle u, -u' \rangle) \\
&= -(\arccos\langle f(u, v), f(-u', v) \rangle - \arccos\langle u, -u' \rangle) \ .
\end{aligned}$$

As a result, indeed it is enough that we prove (3.5) and then (3.6) follows by replacing

$u'$ with $-u'$.

Consider vector $v^* := (\cos\varepsilon, \sin\varepsilon, 0, \ldots, 0)$ (see Figure 3-2). We will now show that if $\varepsilon \le \alpha_t \le \pi/2$ and the intervention $v$ is sufficiently close to $v^*$, then $v$ decreases the angle between $u$ and $u'$. To that end, let us use a metric on $S^{d-1}$ given by

$$D(u, v) := \arccos\langle u, v\rangle \ .$$

Note that this metric is strongly equivalent to the standard Euclidean metric restricted to $S^{d-1}$. We can now use the triangle inequality to write

$$\begin{aligned}\alpha_{t+1} &= D(f(u, v), f(u', v)) \\ &\le D(f(u, v), f(u, v^*)) + D(f(u, v^*), v^*) + D(v^*, f(u', v^*)) + D(f(u', v^*), f(u', v)) \ .\end{aligned} \tag{3.8}$$

Let us bound the terms in (3.8) one by one.

First, since, by (3.2), $f(u, v^*)$ is a strict convex combination of $u$ and $v^*$ (note that in our coordinate system neither $u$ nor $v^*$ depends on $\alpha_t$), we have

$$D(f(u, v^*), v^*) = d(\varepsilon) < D(u, v^*) = \varepsilon \ .$$

Similarly,

$$D(v^*, f(u', v^*)) \le D(v^*, u') = \alpha_t - \varepsilon \ .$$

Second, since $f$ is continuous, $D(v, v^*) < \delta'$ for small enough $\delta' > 0$ implies that both $D(f(u, v), f(u, v^*))$ and $D(f(u', v^*), f(u', v))$ are as small as needed (for example, less than $(\varepsilon - d(\varepsilon))/4$).

All in all, we have that for some $\delta' = \delta'(\varepsilon) > 0$,

$$\begin{aligned}D(v, v^*) < \delta' \implies \alpha_{t+1} &< \frac{\varepsilon - d(\varepsilon)}{4} + d(\varepsilon) + (\alpha_t - \varepsilon) + \frac{\varepsilon - d(\varepsilon)}{4} \\ &= \alpha_t - \frac{\varepsilon - d(\varepsilon)}{2} \ .\end{aligned}$$

However, clearly, the event $D(v, v^*) < \delta'$ has some positive probability $\delta''$. Therefore, taking $\delta := \min(\delta''/2, (\varepsilon - d(\varepsilon))/2)$, we have

$$\Pr\left[\alpha_{t+1} < \alpha_t - \delta \mid \alpha_t\right] > \delta \,,$$

as claimed in (3.5). □

## 3.4 Asymptotic scenario: finding densest hemisphere

In this section we study the asymptotic scenario with one influencer who wishes to propagate a campaign agenda $v^* \in \mathbb{R}^d$. We assume that the influencer can use an unlimited number of interventions and its objective is to make the opinions of as many agents as possible to converge to $v^*$. More specifically, in this section we denote the initial opinions of agents at time $t = 1$ by $u_1, \ldots, u_n$. Given these preexisting opinions of $n$ agents, we want to find a sequence of interventions, $v^{(1)}, v^{(2)}, v^{(3)} \ldots$ that maximizes the number of agents whose opinions converge to $v^*$.

The thrust of our results is that finding a good strategy for the influencer is computationally hard. However, both the optimal strategy and some natural heuristics result in the polarization of agents.

### 3.4.1 Equivalence of optimal strategy to finding densest hemisphere

We first argue that the problem of finding an optimal strategy is equivalent to identifying an open hemisphere that contains the maximum number of agents. An *(open) hemisphere* is an intersection of the unit sphere with a *homogeneous open halfspace* of the form $\left\{x \in \mathbb{R}^d : \langle x, a\rangle > 0\right\}$ for some $a \in \mathbb{R}^d$.

**Theorem 8.** *For any $v^*$, there exists a strategy to make at least $k$ agents converge to $v^*$ if and only if there exists an open hemisphere containing at least $k$ of the opinions $u_1, \ldots, u_n$.*

A surprising aspect of Theorem 8 is that the maximum number of agents that can be persuaded does not depend on the target vector $v^*$. As we argue in Remark 6, this is somewhat plausible in the long-term setting with unlimited number of interventions. We also note that the number of interventions required to bring the opinions up to a given level of closeness to $v^*$ *does* depend on $v^*$.

*Proof of Theorem 8.* First, we prove that the hemisphere condition is sufficient for the existence of a strategy to make the agents' opinions converge (Claim 8). Then we prove the trickier direction: that the hemisphere condition is also *necessary* for the existence of such a strategy (Claim 12).

**Claim 8.** *If opinions $u_1, \ldots, u_k$ are contained in an open hemisphere, then there is a sequence of interventions making all of $u_1, \ldots, u_k$ converge to $v^*$.*

*Proof.* By definition of open hemisphere, there is a vector $a \in \mathbb{R}^d$ such that $\langle u_i, a\rangle > 0$ for every agent $i = 1, \ldots, k$. By (3.2), it is clear that repeated application of $a$ makes all the points converge to $a$ as time $t \to \infty$ .

After all the points are clustered close enough to $a$, by a similar argument they can be "moved around" together towards another arbitrary point $v^*$. For example, if $\langle v^*, a\rangle > 0$, the intervention $v^*$ can be applied repeatedly. If $\langle v^*, a\rangle \leq 0$, one can proceed in two stages, first applying an intervention proportional to $(v^* + a)/2$, and then applying $v^*$. □

**Remark 6.** *As a possible interpretation of the mechanism in Claim 8, it is not unheard of in campaigns on political issues to use an analogous strategy. First, build a consensus around a (presumably compromise) opinion. Then, "nudge" it little by little towards another direction.*

*In an extreme case one can imagine this mechanism even flipping the opinions of two polarized clusters. One example of this could be the reversal of the opinions on certain issues of 20th century Republican and Democratic parties in the US (this particular phenomenon can be found in many texts, e.g. [61]).*

To prove the other direction of Theorem 8, we will rely on the notions of conical combination and convex cone. A *conical combination* of points $u_1, \ldots, u_n \in \mathbb{R}^d$ is any point of the form $\sum_{i=1}^n \alpha_i u_i$ where $\alpha_i \geq 0$ for every $i$. A *convex cone* is a subset of $\mathbb{R}^d$ that is closed under finite conical combinations of its elements. Given a finite set of points $S \subseteq \mathbb{R}^d$, the convex cone *generated* by $S$ is the smallest convex cone that contains $S$.

**Claim 9.** *Suppose that for a given sequence of interventions, the opinions $u_1, \ldots, u_n$ converge to the same point $v^*$. Then, for any unit vector $u_{n+1}$ that lies in the convex cone of $u_1, \ldots, u_n$, we have that $u_{n+1}$ also converges to $v^*$.*

*Proof.* It suffices to prove that if at time $t$ an opinion $u_{n+1}^{(t)}$ lies in the convex cone of other opinions $u_1^{(t)}, \ldots, u_n^{(t)}$, then after applying one intervention $v^{(t)}$ the new opinion $u_{n+1}^{(t+1)}$ lies in the convex cone of $u_1^{(t+1)}, \ldots, u_n^{(t+1)}$. Then the claim follows by induction.

To prove this, we can simply write out $u_{n+1}^{(t+1)}$, using the relation $u_{n+1}^{(t)} = \sum_{i=1}^n \lambda_i u_i^{(t)}$ (where we use the notation $u \propto v$ to mean that $u = c \cdot v$ for some constant $c > 0$):

$$\begin{aligned} u_{n+1}^{(t+1)} &\propto u_{n+1}^{(t)} + \eta \left\langle u_{n+1}^{(t)}, v^{(t)} \right\rangle \cdot v^{(t)} \\ &= \sum_{i=1}^n \lambda_i u_i^{(t)} + \eta \cdot \sum_{i=1}^n \lambda_i \left\langle u_i^{(t)}, v^{(t)} \right\rangle \cdot v^{(t)} \\ &= \sum_{i=1}^n \lambda_i \left( u_i^{(t)} + \eta \cdot \left\langle u_i^{(t)}, v^{(t)} \right\rangle \cdot v^{(t)} \right) \\ &= \sum_{i=1}^n \lambda_i \cdot c_i u_i^{(t+1)} \end{aligned} \tag{3.9}$$

where the constants in (3.9) are $c_i := \left\| u_i^{(t)} + \eta \cdot \langle u_i^{(t)}, v^{(t)} \rangle \cdot v^{(t)} \right\|$. Specifically, they are all nonnegative. □

**Claim 10.** *Suppose there are two opinions $u_1, u_2$ that are antipodal, i.e., $u_1 = -u_2$. Then these two opinions will remain antipodal in future time steps. In particular, they will never converge to a single point.*

*Proof.* This follows directly from (3.2), noting that, for any intervention $v$, we have $u_1 + \eta \cdot \langle u_1, v \rangle \cdot v = -\left(u_2 + \eta \cdot \langle u_2, v \rangle \cdot v\right)$. □ □

We will also use the following consequence of the separating hyperplane theorem:

**Claim 11.** *A collection of unit vectors $a_1, \ldots, a_n$ cannot be placed in an open hemisphere if and only if the zero vector lies in the convex hull of $a_1, \ldots, a_n$.*

Now we are ready to establish the reverse implication in Theorem 8.

**Claim 12.** *Suppose that we start with agent opinions $u_1, \ldots, u_n$ and that there is no hemisphere that contains $M$ of those opinions. Then, there is no strategy that makes $M$ of the opinions converge to the same point.*

*Proof.* Assume towards contradiction that there exists a strategy that makes $M$ opinions converge to the same point, and assume wlog that they are $u_1, \ldots, u_M$. By assumption, we know that there is no hemisphere that contains all of $u_1, \ldots, u_M$, hence, by Claim 11, there is a convex combination of $u_1, \ldots, u_M$ that equals 0. Therefore, there is also a conical combination of $u_1, \ldots, u_{M-1}$ that equals $-u_M$, where wlog we assume that the coefficient on $u_M$ is initially nonzero. By Claim 9, we conclude that if $u_1, \ldots, u_{M-1}$ converge to the same point, then so does $-u_M$. But that means that $-u_M$ and $u_M$ converge to the same point, which is a contradiction by Claim 10. □

That concludes the proof of Theorem 8. □

**Remark 7.** *One consequence of Theorem 8 is that if the agent opinions are initially distributed uniformly on the unit sphere, and if the number of agents $n$ is large compared to the dimension $d$, an optimal strategy converging as many opinions as possible to $v^*$ results, with high probability, in dividing the population into two groups of roughly equal size, where the opinions inside each group converge to one of two antipodal limit opinions (i.e., $v^*$ and $-v^*$). Furthermore, this optimal strategy, which, as discussed below, might not be easy to implement, will not perform significantly better than a very simple strategy of fixing a random intervention and applying it repeatedly. Of course the simple strategy will also polarize the agents into two approximately equally large groups.*

### 3.4.2 Computational equivalence to learning halfspaces

Theorem 8 implies that an optimal strategy for the influencer is to compute the open hemisphere that is the densest, i.e., it contains the most opinions, and then apply the procedure from Claim 8 to converge the opinions from this hemisphere to $v^*$. In this section we study the computational complexity of this problem. While different approaches are possible, we focus on hardness of approximation and worst-case complexity.

**Definition 13** (Densest hemisphere)**.** *The input to the* densest hemisphere *problem consists of parameters $n$ and $d$ and a set of $n$ unit vectors $D = \{u_1, \ldots, u_n\}$ with $u_i \in S^{d-1}$. The objective is to find vector $a \in S^{d-1}$ maximizing the number of points from $D$ that belong to the open halfspace $\left\{x \in \mathbb{R}^d : \langle x, a\rangle > 0\right\}$.*

We analyze the computational complexity of the densest hemisphere problem in terms of the number of vectors $n$, regardless of dimension $d$. In particular, the computationally hard instances that exist as we will show in Theorem 9 have high dimension, without any guarantees beyond $d \leq n$ (which can always be assumed wlog). On the other hand, the algorithms from Theorem 10 run in time polynomial in $n$ uniformly for all values $d \leq n$.

In contrast, the case of finding densest hemisphere in fixed dimension $d$ can be solved efficiently. For example, an optimal solution can be found by considering $O(n^d)$ halfspaces defined by $d$-tuples of input vectors. We omit further details.

Our main result in this section relies on equivalence of the densest hemisphere problem and the problem of *learning noisy halfspaces*. Applying a work by Guruswami and Raghavendra [43] we will show that it is computationally difficult to even approximate the densest hemisphere up to any non-trivial constant factor:

**Theorem 9.** *Unless P=NP, for any $\varepsilon > 0$, there is no polynomial time algorithm $A_\varepsilon$ that distinguishes between instances of densest hemisphere problem such that, letting $D := \{u_1, \ldots, u_n\}$:*

- *Either there exists a hemisphere $H$ such that $|D \cap H|/n > 1 - \varepsilon$.*

- *Or for every hemisphere $H$ we have $|D \cap H|/n < 1/2 + \varepsilon$.*

*Consequently, unless P=NP, for any $\varepsilon > 0$ there is no polynomial time algorithm $A_\varepsilon$ that, given an instance $D$ that has a hemisphere with density more than $1 - \varepsilon$, always outputs a hemisphere with density more than $1/2 + \varepsilon$.*

In other words, even if guaranteed the existence of an extremely dense hemisphere, no polynomial time algorithm can do significantly better than choosing an arbitrary hyperplane and outputting the one of its two hemispheres that contains the larger number of points. At the same time, [17] (relying on earlier work [16]) shows that there exists an algorithm that finds a dense hemisphere provided that this hemisphere is stable in the sense that it remains dense even after a small perturbation of its separating hyperplane:

**Theorem 10** ([17])**.** *For every $\mu > 0$, there exists a polynomial time algorithm $A_\mu$, that, given an instance $D = \{u_1, \ldots, u_n\}$ of the densest hemisphere problem, provides the following guarantee:*

*Let $a \in S^{d-1}$ be the vector that maximizes the size of intersection $|D \cap H_{a,\mu}|$ for halfspace $H_{a,\mu} = \{x : \langle x, a\rangle > \mu\}$. Then, the algorithm $A_\mu$ outputs a hemisphere corresponding to a homogeneous halfspace $H_{a'} = \{x : \langle x, a'\rangle > 0\}$ such that $|D \cap H_{a'}| \geq |D \cap H_{a,\mu}|$.*

We emphasize that the only inputs to the algorithms are $n$, $d$ and the set of vectors $D$, and that the complexity is measured as a function of $n$. For example, the algorithm $A_\mu$ runs in polynomial time for every $\mu > 0$, but the running time is not uniformly polynomial in $1/\mu$.

In the remainder of this section we elaborate on how to obtain Theorem 9 from known results. To that end, we start with defining the related problem of finding maximum agreement halfspace.

**Definition 14** (Maximum Agreement Halfspace)**.** *In the problem of* maximum agreement halfspace*, the inputs are parameters $n$ and $d$, and a labeled set of points $D =$*

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathbb{R}^d \times \{\pm 1\}$. *The objective is to find a halfspace* $H = \{x : \langle x, a\rangle > c\}$ *for some* $a \in \mathbb{R}^d$ *and* $c \in \mathbb{R}$ *which maximizes the* agreement

$$A(D, H) = \frac{\sum_{i=1}^{n} \mathbb{1}\left[y_i \cdot x_i \in H\right]}{n} .$$

There is a strong hardness of approximation result for maximum halfspace agreement [43] (see also [32, 24, 15, 2] for related work):

**Theorem 11** ([43])**.** *Unless P=NP, for any* $\varepsilon > 0$*, there is no polynomial time algorithm* $A_\varepsilon$ *that distinguishes the following cases of instances of maximum agreement halfspace problem:*

- *There exists a halfspace* $H$ *such that* $A(D, H) > 1 - \varepsilon$*.*
- *For every halfspace* $H$ *we have* $A(D, H) < 1/2 + \varepsilon$*.*

As in Theorem 9, the hard instances are not guaranteed to have any dimension bounds beyond trivial $d \leq n$.

As pointed out in [17], there exists a reduction from the maximum agreement halfspace problem to the densest hemisphere problem that preserves the quality of solutions. Since this reduction is only briefly sketched in [17], we describe it below.

The reduction proceeds as follows: Given a labeled set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathbb{R}^d \times \{\pm 1\}$, we map it to $D' = \{x'_1, \ldots, x'_n\} \in \mathbb{R}^{d+1}$ using the formula

$$x'_i = \frac{1}{\sqrt{1 + \|x_i\|^2}} \cdot (y_i x_i, 1) .$$

In other words, we proceed in three steps: first, we negate each point that came with negative label $y_i = -1$. Then, we add a new coordinate and set its value to 1 for every point $x_i$. Finally, we normalize each resulting point so that it lies on the unit sphere in $S^d$.

This is a so-called "strict reduction", which is expressed in the following claim:

**Claim 13.** *The solutions (halfspaces) for an instance* $D$ *of Maximum Agreement Halfspace are in one-to-one correspondence with solutions (hemispheres) for the re-*

*duced instance $D'$ of Densest Hemisphere. Furthermore, for a corresponding pair of solutions $(H, H')$ the agreement $A(D, H)$ is equal to the density $|D' \cap H'|/n$.*

*Proof.* It is more convenient to think of solutions for $D'$ as homogeneous, open halfspaces $H' = \{x \in \mathbb{R}^{d+1} : \langle x, a\rangle > 0\}$.

With that in mind, we map a solution to the maximum agreement halfspace problem $H = \{x \in \mathbb{R}^d : \langle x, a\rangle > c\}$ to a solution to the densest hemisphere problem $H' = \{(x, x_{d+1}) \in \mathbb{R}^{d+1} : \langle (x, x_{d+1}), (a, -c))\rangle > 0\}$. Clearly, this is a one-to-one mapping between open halfspaces in $\mathbb{R}^d$ and homogeneous open halfspaces in $\mathbb{R}^{d+1}$.

Furthermore, it is easy to verify that $y_i \cdot x_i \in H$ if and only if $x'_i \in H'$ and therefore $A(D, H) = |D' \cap H'|/n$. □

Theorem 9 follows from Theorem 11 and Claim 13 by standard (and straightforward) arguments from complexity theory.

## 3.5 Short-term scenario: polarization as externality

The analysis of the asymptotic setting with unlimited interventions tells us what is feasible and what is not. A fundamentally different question is how to persuade as many as possible with a limited number of interventions. This is motivated by bounded resources or time that usually allow only limited placements of campaigns and advertisements. Furthermore, arguably only the initial interventions can be considered effective: in the long run the opinions might shift due to external factors and become more unpredictable and harder to control. Therefore, in this section we discuss strategies where the influencer has only one intervention at its disposal, and its goal is to get as many agents as possible to exceed certain "threshold of agreement" with its preferred opinion. Throughout this section, we fix $\eta = 1$ in Equation 3.1, so an opinion $u$ is updated to be proportional to $w = u + \langle u, v\rangle \cdot v$.

Both scenarios we discuss in this section describe a situation where a "new" product or idea is introduced. Therefore, we assume that the agents have some preexisting opinions in $\mathbb{R}^{d-1}$ and that they are neutral as to the new idea, with the $d$-th coordinate

set to zero for every agent. Our results indicate significant potential for polarization in such a situation. This is in spite of the fact that the influencer might only care about persuading a number of agents towards the new subject, without intention to polarize.

Since we are dealing with scenarios with only one intervention, we use the following notational convention: an initial opinion of agent $i$ is denoted $u_i$ and the opinion after intervention is denoted $\widetilde{u}_i$.

### 3.5.1 One intervention, two agents: polarization costs

We consider a simple example that features only two agents and one influencer who is allowed one intervention. We imagine a new product, such that the agents are initially agnostic about it, i.e., $u_{i,d} = 0$ for $i = 1, 2$. Given an intervention $v$, we are interested in two issues: First, what will be new opinions of agents about the product $\widetilde{u}_{i,d}$? Second, assuming that the initial correlation between opinions is $c = \langle u_1, u_2 \rangle$, what will be the new correlation $\widetilde{c} = \langle \widetilde{u}_1, \widetilde{u}_2 \rangle$? We think of the correlation as a measure of agreement between the agents and therefore interpret differences in correlation as changes in the extent of polarization.

In order to answer these questions, we introduce notions of two- and one-agent interventions corresponding to two natural strategies:

**Definition 15.** *The* two-agent intervention *is an intervention that maximizes* $\min(\widetilde{u}_{1,d}, \widetilde{u}_{2,d})$. *The* one-agent intervention *maximizes* $\max(\widetilde{u}_{1,d}, \widetilde{u}_{2,d})$.

The motivation for this definition is as follows. Assume that there exists a threshold $T > 0$ such that agent $i$ is going to make a positive decision (e.g., buy the product or vote a certain way) if its coordinate $\widetilde{u}_{i,d}$ exceeds $T$. Then, if the influencer cares only about inducing agents to make the decision, it has two natural choices for the intervention. One option is the case where it is possible to induce two decisions, i.e., achieve $\widetilde{u}_{1,d}, \widetilde{u}_{2,d} > T$. By continuity considerations, it is not difficult to see that an intervention that achieves this can be assumed to maximize $\min(\widetilde{u}_{1,d}, \widetilde{u}_{2,d})$ with $\widetilde{u}_{1,d} = \widetilde{u}_{2,d}$ (such intervention is also optimal if the influencer bets on convincing both

agents without knowing $T$). The other case is to appeal only to one of the agents, disregarding the second agent and concentrating only on achieving, say, $\widetilde{u}_{1,d} > T$.

Let $c = \langle u_1, u_2 \rangle$ be the initial correlation between opinions and let $c_{\text{two}}$ and $c_{\text{one}}$ be the correlations after applying, respectively, the two- and one-agent interventions. Our main result in this section is:

**Proposition 2.** *Let* $\rho := c_{\text{two}} - c_{\text{one}}$ *be a value that we call the* polarization cost. *Then, we always have* $\rho \geq 0$ *with exact values given as*

$$c_{\text{two}} = 1 - \frac{\sqrt{2}(1-c)}{\sqrt{3c+5}}\,, \qquad\qquad c_{\text{one}} = \frac{c\sqrt{2}}{\sqrt{c^2+1}}\,. \tag{3.10}$$

The values of $\rho, c_{\text{two}}$ and $c_{\text{one}}$ as functions of $c$ are illustrated in Figure 3-3. Proposition 2 states that the one-agent intervention always results in smaller correlation than the two-agent intervention. Note that we made a modeling assumption that the influencer will always choose an intervention as opposed to doing nothing. This is consistent with a scenario where the influencer's objective is to increase the opinions above the threshold $T$. In that case doing nothing is certain to give no gain to the influencer.

The main conclusion of this theorem is consistent with our other results. In the setting we consider, in the absence of any external mitigation, the self-interested influencer without direct intention to polarize might be incentivized to choose the intervention that increases polarization. If polarization is regarded as undesirable, the polarization cost can be thought of as the externality imposed on the society.

Looking at Figures 3-3 and 3-4, this effect seems most pronounced for initial correlation around $c \approx -0.5$, where the one-agent intervention increases polarization, the polarization cost is large and the range of thresholds $T$ for which the influencer profits from the one-agent strategy is relatively large. This suggests that a situation where the society is already somewhat polarized is particularly vulnerable to spiraling out of control. It also suggests that situations where the level of commitment required for the decision (i.e., the threshold $T$) is large increase the risk of polarization.

We also note that this overall picture is complicated by the case of positive initial

correlation $c > 0$. In that case both two- and one-agent interventions actually increase the correlation between the agents, even though the two-agent intervention does so to a larger extent. The analysis leading to the proof of Proposition 2 is contained in Appendix B.3.

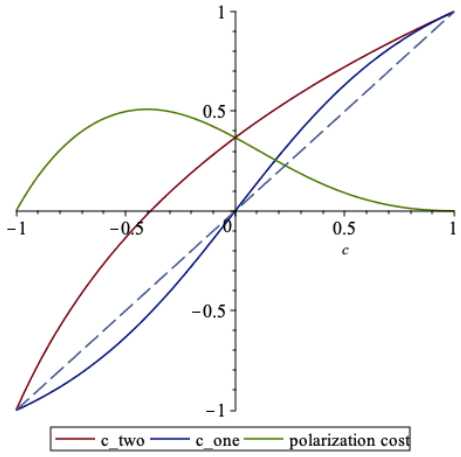


Figure 3-3: Illustration of the polarization cost as a function of the initial correlation $c$. The dashed line is the initial correlation included as a reference point. The red and blue lines are correlations after applying two- and one-agent interventions respectively. The green line shows the polarization cost $c_{\text{two}} - c_{\text{one}}$.

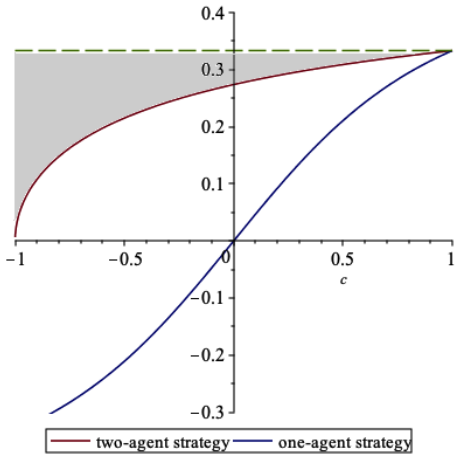


Figure 3-4: The after-intervention opinions of both agents $\widetilde{u}_{i,d}$ as functions of initial correlation $c$. The red line represents the opinion of either agent after applying the two-agent intervention. The blue line is the opinion of the second agent after the one-agent intervention. For reference, the dashed line (1/3) shows the opinion of the first agent in the one-agent intervention (which does not depend on $c$). The grey area represents the range of thresholds $T$ where it is preferable for the influencer to apply the one-agent intervention.

### 3.5.2 One intervention, many agents: finding the densest spherical cap

A more general version of the problem of persuading with limited number of interventions features $n$ agents with opinions $u_1, \ldots, u_n \in \mathbb{R}^d$. The influencer is given a threshold $0 \leq T < 1$ and can apply one intervention $v$ with the objective of maxi-

mizing the number of agents such that $\widetilde{u}_{i,d} > T$. As before, we assume that intially $u_{i,d} = 0$ and that $T$ can be interpreted as a threshold above which a consumer decides to buy the newly advertised product, or more generally take a desired action, such as voting, donating, etc.

Interestingly, we show that this problem is equivalent to a generalization of the densest hemisphere problem from the long-term scenario discussed in Section 3.4. More precisely, it is equivalent to finding a densest *spherical cap* of a given radius (that depends on the threshold $T$) in $d-1$ dimensions.

We give the technical statement in the proposition below. We make an assumption $0 \leq T < 1/3$, since $1/3$ is the maximum value that can be achieved in the $d$-th coordinate by a single intervention, cf. Figure 3-4. In order to state Proposition 3, we slightly abuse notation and write vectors $u \in \mathbb{R}^d$ as $u = (u^*, u_d)$ for $u^* \in \mathbb{R}^{d-1}$, $u_d \in \mathbb{R}$.

**Proposition 3.** *In the setting above, let*

$$c := \frac{2T}{1-3T^2}\,, \qquad z := \frac{\sqrt{\sqrt{1+3c^2}-1}}{\sqrt{3}c}\,, \qquad \beta := \arccos(z)\,.$$

*Then, the number of agents with $\widetilde{u}_{i,d} > T$ is maximized by applying an intervention*

$$v := (\cos\beta \cdot v^*, \sin\beta) \tag{3.11}$$

*for a unit vector $v^* \in \mathbb{R}^{d-1}$ that maximizes the number of agents satisfying*

$$\langle u_i^*, v^* \rangle > c\,.$$

The proof of Proposition 3 is contained in Appendix B.4. Note that the solution to this short-term problem for $T$ going to zero approaches the densest hemisphere solution to the long-term problem discussed in Section 3.4.

## 3.6 Asymptotic effects of two dueling influencers: two randomized interventions polarize

Finally, we analyze a scenario where there are two influencers with differing agendas, represented by different[1] intervention vectors $v$ and $v'$. We consider the *randomized* setup, where at each time step, one of the influencers is randomly chosen to apply their intervention. We demonstrate that this setting also results, in most cases and in a certain sense, in the polarization of agents.

Recall that a convex cone of two vectors $v$ and $v'$ is the set $\{\alpha v + \beta v' : \alpha, \beta \geq 0\}$. A precise statement that we prove is:

**Theorem 12.** *Let $\langle v, v' \rangle > 0$ and let a starting opinion $u^{(1)}$ be such that $\langle u^{(1)}, v \rangle \neq 0$ or $\langle u^{(1)}, v' \rangle \neq 0$. Then, as $t$ goes to infinity and almost surely, either the Euclidean distance between $u^{(t)}$ and the convex cone generated by $v$ and $v'$ or between $u^{(t)}$ and the convex cone generated by $-v$ and $-v'$ goes to 0.*

In order to justify the assumptions of Theorem 12, note that if an agent starts with an opinion $u^{(1)}$ such that

$$\langle u^{(1)}, v \rangle = \langle u^{(1)}, v' \rangle = 0 \,, \tag{3.12}$$

applying $v$ or $v'$ never changes their opinion. In Theorem 12 we show that if (3.12) does not hold and, additionally, $\langle v, v' \rangle > 0$, (if $\langle v, v' \rangle < 0$ we can exchange $v'$ with $-v'$ without changing the effects of any interventions), the opinion vector with probability 1 ends up either converging to the convex cone generated by $v$ and $v'$ or the convex cone generated by $-v$ and $-v'$. In particular, since vectors $u$ for which (3.12) holds form a set of measure 0, if $n$ initial opinions are sampled iid from an absolutely continuous distribution, almost surely all opinions converge to the convex cones (which are themselves sets of measure 0 for $d > 2$).

Furthermore, this notion of polarization is strengthened if the correlation between the two interventions is large. As in Theorem 7, the best we can hope for is that

[1] We also assume that $v \neq -v'$, as otherwise the intervention effects are the same in our model.

for each pair of opinions either the distance between $u_1^{(t)}$ and $u_2^{(t)}$ or between $u_1^{(t)}$ and $-u_2^{(t)}$ converges to 0. Letting $V := \text{span}\{v, v'\}$ and $W := V^\perp$ and writing any vector $u$ as a sum of its respective projections $u = u_V + u_W$, we show:

**Theorem 13.** *Suppose that $\langle v, v' \rangle > 1/\sqrt{2+\eta}$ and let $u_1^{(1)}, u_2^{(1)}$ be such that $(u_1^{(1)})_V \neq 0$, $(u_2^{(1)})_V \neq 0$. Then, almost surely, either $\|u_1^{(t)} - u_2^{(t)}\|$ converges to 0, or $\|u_1^{(t)} + u_2^{(t)}\|$ converges to 0.*

In other words, the stronger notion of convergence, same as in Section 3.3 with uniformly drawn random interventions, reappears in case the correlation between two interventions $v$ and $v'$ is larger than $1/\sqrt{2+\eta}$. In particular, we have strong convergence for any $\eta > 0$ and $\langle v, v' \rangle \geq \sqrt{2}/2 \approx 0.71$, and for $\eta = 1$ for $\langle v, v' \rangle > \sqrt{3}/3 \approx 0.58$. Our experiments suggest that this convergence occurs also for other non-zero values of the correlation $\langle v, v' \rangle$, but we do not prove it here.

Also note that same in spirit as Remark 3.1, the usual argument from symmetry shows that if the initial opinions are independent samples from a symmetric distribution, then with high probability the opinions divide into two clusters of roughly equal size.

The case when $v$ and $v'$ are orthogonal is different. As we mentioned, if $\langle v, v' \rangle > 0$, i.e., the angle between $v$ and $v'$ is less than $\pi/2$, then all opinions converge to the two "narrow" convex cones, respectively between $v$ and $v'$ and between $-v$ and $-v'$ — namely, the pairs of vectors among $v, v', -v$, and $-v'$ between which there are acute angles. Similarly, if $\langle v, v' \rangle < 0$, then the opinions converge to two cones between $v$ and $-v'$ and between $-v$ and $v'$. In case $\langle v, v' \rangle = 0$ the four convex cones form right angles, so such a result is not possible.

However, we can still show that an initial opinion $u^{(1)}$ converges to the same quadrant in which it starts with respect to $v$ and $v'$. Namely, for all $t$, we have that $\text{sgn}\left(\left\langle u^{(t)}, v \right\rangle\right) = \text{sgn}\left(\left\langle u^{(1)}, v \right\rangle\right)$ and $\text{sgn}\left(\left\langle u^{(t)}, v' \right\rangle\right) = \text{sgn}\left(\left\langle u^{(1)}, v' \right\rangle\right)$, and furthermore the distance between $u^{(t)}$ and the subspace $V$ goes to 0 with $t$:

**Proposition 4.** *Suppose that $\langle v, v' \rangle = 0$ and let an initial opinion $u^{(1)}$ be such that $\langle u^{(1)}, v \rangle \neq 0$ and $\langle u^{(1)}, v' \rangle \neq 0$. Then, almost surely, the following facts hold:*

1. $\|u_W^{(t)}\| \to 0$ *as* $t \to \infty$.

2. *For all* $t$, $\operatorname{sgn}\left(\left\langle u^{(t)}, v\right\rangle\right) = \operatorname{sgn}\left(\left\langle u^{(1)}, v\right\rangle\right)$ *and* $\operatorname{sgn}\left(\left\langle u^{(t)}, v'\right\rangle\right) = \operatorname{sgn}\left(\left\langle u^{(1)}, v'\right\rangle\right)$.

Fascinatingly, Gaitonde, Kleinberg and Tardos [38] showed subsequently to our initial preprint that strong polarization does not occur for orthogonal interventions. Specifically, they proved that two opinions in $S^{d-1}$ with random interventions chosen iid from the standard basis $\{e_1, \ldots, e_d\}$ do not polarize in the sense of $u_1^{(t)} - u_2^{(t)}$ or $u_1^{(t)} - u_2^{(t)}$ vanishing, but they do exhibit a weaker form of polarization. We refer to their paper for more details.

In order to prove Theorem 12, we first show that the distance between $u^{(t)}$ and $V$ almost surely goes to 0 as $t \to \infty$, by showing that the norm of the projection of $u^{(t)}$ onto $W$ converges to 0. Then, we demonstrate that the convex cone spanned by $v$ and $v'$ is absorbing: when the projection of $u^{(T)}$ onto $V$ falls in the cone, then the projections of $u^{(t)}$ for $t \geq T$ always stay in the cone as well.

Finally, we show that almost surely the projection of $u^{(t)}$ onto $V$ eventually enters either the cone spanned by $v$ and $v'$, or the cone spanned by $-v$ and $-v'$. More concretely, we show that at any time $t$, there is a sequence of $T$ interventions that lands the projection of $u^{(t+T)}$ in one of the cones, for some $T$ that is independent of $t$. Since this sequence occurs with probability $2^{-T}$, which is independent of $t$, the opinion almost surely eventually enters one of the cones.

### 3.6.1 Proofs of Theorem 12 and Proposition 4

We start with the fact the opinions converge to the subspace $V$ spanned by the two intervention vectors. Recall that $V = \operatorname{span}\{v, v'\}$ and that $W = V^{\perp}$. In the following we will write $\langle v, v' \rangle = \cos\theta$ for $0 < \theta \leq \pi/2$.

**Proposition 5.** *Let* $\langle v, v' \rangle \geq 0$ *and take an opinion vector* $u$ *such that* $\|u_V\| = c \geq 0$. *Furthermore, let* $\widetilde{u}$ *be the vector resulting from randomly intervening on* $u$ *with either* $v$ *or* $v'$. *Then:*

1. $\|\widetilde{u}_W\|^2 \leq \|u_W\|^2$.

*2. With probability at least 1/2,* $\|\widetilde{u}_W\|^2 \leq \|u_W\|^2 \cdot (1-\xi)$*, where*

$$\xi = \min\left(\frac{1}{2}, (\eta + \eta^2/2) \cdot \frac{c^2\theta^2}{16}\right) .$$

*Proof.* Recall from (3.2)–(3.3) that if $\overline{v} \in \{v, v'\}$ is the intervention vector, then

$$\widetilde{u} = k(u + \eta \langle u, \overline{v}\rangle \cdot \overline{v})$$

where $k = \sqrt{\frac{1}{1+(2\eta+\eta^2)\cdot\langle u,\overline{v}\rangle^2}}$ is the normalizing constant. Observe that when we project onto $W$, the component in the direction of $\overline{v}$ vanishes, so we have that

$$\widetilde{u}_W = k \cdot u_W ,$$

and the first claim easily follows since $k \leq 1$.

To establish the second point, we need to show that with probability $1/2$ we have $k^2 < 1$ or, equivalently, $\langle u, \overline{v}\rangle^2 = \langle u_V, \overline{v}\rangle^2 > 0$. Since $\theta \neq 0$, the projected vector $u_V$ cannot be orthogonal both to $v$ and $v'$ (cf. Figure 3-5). More precisely, for at least one of $\overline{v} \in \{v, v'\}$ the primary angle between $u_V$ and $\overline{v}$ (or $-\overline{v}$) must be at most $\pi/2 - \theta/2$ and consequently

$$|\langle u_V, \overline{v}\rangle| \geq \|u_V\| \cdot |\cos(\pi/2 - \theta/2)| \geq c \cdot \theta/4 ,$$

resulting in

$$k^2 = \frac{1}{1 + (2\eta + \eta^2) \cdot \langle u_V, \overline{v}\rangle^2} \leq \max\left(\frac{1}{2}, 1 - (\eta + \eta^2/2) \cdot \frac{c^2\theta^2}{16}\right) . \qquad \square$$

Next, we show that the convex cone of vectors $v$ and $v'$ is absorbing:

**Proposition 6.** *Let* $\langle v, v'\rangle \geq 0$ *and take* $u$ *to be an opinion vector and* $\widetilde{u}$ *to be a vector resulting from intervening on* $u$ *with either* $v$ *or* $v'$*. If* $u_V$ *is a conical combination of* $v$ *and* $v'$*, then also* $\widetilde{u}_V$ *is such a conical combination.*

*Proof.* Assume wlog that the vector applied is $v$ and let $k$ be the same constant as in

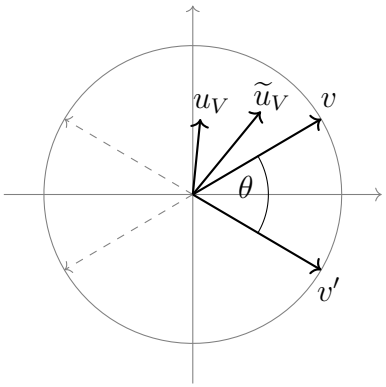


Figure 3-5: Projection onto the subspace $V = \text{span}\{v, v'\}$.

the proof of Proposition 5. Then,

$$\widetilde{u}/k = u + \eta \cdot \langle u, v \rangle \cdot v = u_V + \eta \cdot \langle u_V, v \rangle \cdot v + u_W \ .$$

Therefore, $\widetilde{u}_V$ can be written as a nonnegative linear combination of $u_V$ and $v$, where we use the fact that $\langle u_V, v \rangle$ is nonnegative, which follows since $u_V$ is a conical combination of $v$ and $v'$, and $\langle v, v' \rangle \geq 0$. □

Next, we prove that when $\langle v, v' \rangle > 0$, the opinion $u^{(t)}$ not only approaches subspace $V$, but also a specific area of $V$, namely, either $\text{cone}(v, v')$ or $\text{cone}(-v, -v')$.

**Proposition 7.** *Let $\langle v, v' \rangle > 0$ and consider a vector $u^{(t)}$ such that $\|u_V^{(t)}\| \geq c > 0$. Then, there exists $T := T(c, \theta, \eta)$ such that with probability at least $2^{-T}$, vector $u_V^{(t+T)}$ will either be a conical combination of $v$ and $v'$* or *a conical combination of $-v$ and $-v'$.*

*Proof.* First, for any vector $u^{(t)}$ such that $\|u_V^{(t)}\| \geq c > 0$, at least one of $v, v', -v, -v'$ has positive inner product with $u^{(t)}$ (and $u_V^{(t)}$) which can be lower bounded by a function of $c$ and $\theta$ (see Figure 3-5). Take such a vector and call it $\overline{v}$. By the argument from Proposition 5, applying $\overline{v}$ repeatedly will bring $u^{(t+T)}$ arbitrarily close to it. More precisely, for every $\varepsilon > 0$, there exists $T_1 = T_1(c, \theta, \eta, \varepsilon)$ such that $\|u_V^{(t+T_1)} - \overline{v}\| < \varepsilon$ and $\|u^{(t+T_1)} - \overline{v}\| < \varepsilon$ both hold.

Furthermore, since $\langle v, v' \rangle > 0$, there exists $\varepsilon > 0$ such that if $\|u^{(t)} - \overline{v}\| < \varepsilon$, then applying the other intervention vector ($v$ or $v'$) once guarantees that $u_V^{(t+1)}$ enters the

convex cone between $v$ and $v'$ or, respectively, between $-v$ and $-v'$. In particular, if $u_V^{(t)}$ already is in the convex cone, then applying either intervention will keep it inside by Proposition 6. On the other hand, if $u_V^{(t)}$ is not yet in the cone, but at the distance at most $\varepsilon$ to $\overline{v}$, then applying the other intervention will bring it inside the cone (see Figure 3-5).

Therefore, there exists a sequence of $T(c, \theta, \eta) = T_1 + 1$ interventions that make $u_V^{(t+T)}$ enter cone$(v, v')$ or cone$(-v, -v')$. Clearly, this sequence occurs with probability $2^{-T}$. □

We are now ready to prove Theorem 12.

*Proof of Theorem 12.* Let $\|u_V^{(1)}\| = c > 0$. Proposition 5 tells us that the squared norm of the projection $u_W^{(t)}$ onto subspace $W = V^\perp$ never increases, and with probability $1/2$ decreases by the multiplicative factor $1 - \xi(c, \eta, \theta) < 1$. By induction (note that $\xi$ increases with $c$), $u_W^{(t)}$ converges to 0, and consequently $\|u^{(t)} - u_V^{(t)}\|$ converges to 0, almost surely.

In order to show that convergence to one of the two convex cones occurs, we apply Proposition 7. Since at *any time step* $t$, there exists a sequence of $T$ choices that puts $u_V^{(t+T)}$ in one of the convex cones, and since $T$ depends only on the starting parameters $c$, $\theta$, and $\eta$, we get that $u_V^{(t)}$ almost surely eventually enters one of the cones. By Proposition 6 and induction, once $u_V^{(t)}$ enters a convex cone, it never leaves. □

Proposition 4 follows as a corollary of Propositions 5 and 6:

*Proof of Proposition 4.* The first statement is an inductive application of Proposition 5, exactly the same as in the proof of Theorem 12.

The second statement follows from noting that out of four orthogonal pairs of vectors $\{v, v'\}$, $\{v, -v'\}$, $\{-v, v'\}$, or $\{-v, -v'\}$, there is exactly one such that $u_V^{(1)}$ is a (strict) conical combination of this pair (by assuming $\langle u^{(1)}, v\rangle \neq 0$ and $\langle u^{(1)}, v'\rangle \neq 0$ we avoid ambiguity in case $u_V^{(1)}$ is parallel to $v$ or $v'$). By the same argument as in Proposition 6 and by induction, if the initial projection $u_V^{(1)}$ is strictly inside one of the convex cones, the projection $u_V^{(t)}$ remains strictly inside forever. □

### 3.6.2 Proof of Theorem 13

Consider the subspace $V = \text{span}\{v, v'\}$ with some coordinate system (cf. Figure 3-5) imposed on it. As is standard, a unit vector $u \in V$ can be represented in this system by its angle $\alpha(u) \in [0, 2\pi)$ as measured counterclockwise from the positive $x$-axis.

Given a unit vector $\overline{v} \in V$, let $f_{\overline{v}} : [0, 2\pi) \to [0, 2\pi)$ be the function with the following meaning: given a unit vector $u \in V$ with angle $\alpha = \alpha(u)$, the value $f_{\overline{v}}(\alpha) = \alpha(\widetilde{u})$ represents the angle of vector $\widetilde{u}$ resulting from applying intervention $\overline{v}$ to vector $u$. Note that $\alpha(\overline{v})$ is a fixed point of $f_{\overline{v}}$. Also, by Proposition 6, both functions $f_v$ and $f_{v'}$ map the interval corresponding to $\text{cone}(v, v')$ to itself.

The main part of our argument is the following lemma, which we prove last:

**Lemma 8.** *If $\langle v, v' \rangle = \cos\theta > 1/\sqrt{2+\eta}$, then functions $f_v$ and $f_{v'}$ restricted to the convex cone of $v$ and $v'$ are contractions, i.e., there exists $k = k(\theta, \eta) < 1$ such that for all vectors $u, u' \in \text{cone}(v, v')$, letting $\alpha := \alpha(u), \beta := \alpha(u'), \overline{v} \in \{v, v'\}$, we have*

$$|f_{\overline{v}}(\beta) - f_{\overline{v}}(\alpha)| \leq k \cdot |\beta - \alpha| \ , \tag{3.13}$$

*where the distances $|f_{\overline{v}}(\beta) - f_{v^*}(\alpha)|$ and $|\beta - \alpha|$ are in the metric induced by $S^1$, i.e., "modulo $2\pi$".*

*Proof of Theorem 13.* Lemma 8 implies that the angle distance between two opinions $u_1^{(t)}, u_2^{(t)} \in V$ starting in the convex cone deterministically converges to 0 as $t$ goes to infinity. Of course, this is equivalent to their Euclidean distance $\|u_1^{(t)} - u_2^{(t)}\|$ converging to 0. We now make a continuity argument to show that such convergence almost surely occurs also for general $u_1^{(t)}, u_2^{(t)} \in S^{d-1}$. To this end, we let $g_v, g_{v'} : S^{d-1} \to [0, 2\pi)$ as natural extensions of $f_v, f_{v'}$: the value $g_{\overline{v}}(u)$ denotes the angle of the projection $\widetilde{u}_V$ of the new opinion onto $V$, after applying $\overline{v}$ on opinion $u$ (cf. Figure 3-5). Note that the value $g_{\overline{v}}(u)$ depends only on the angle $\alpha(u_V)$ and the orthogonal projection length $\|u_W\|$:

$$g_{\overline{v}}(u) = g_{\overline{v}}(\alpha(u_V), \|u_W\|) \ .$$

In this parametrization, for $u \in V$ we have $f_{\overline{v}}(\alpha(u)) = g_{\overline{v}}(u) = g_{\overline{v}}(\alpha(u), 0)$.

By Theorem 12, for any starting opinions $u_1^{(1)}$ and $u_2^{(1)}$ having non-zero projections onto $V$, almost surely there exists a $t$ such that $(u_1^{(t)})_V$ and $(u_2^{(t)})_V$ end up inside (possibly different) convex cones. We consider the case of $u_1^{(t)}$ and $u_2^{(t)}$ both in cone$(v, v')$, other three cases being analogous. Furthermore, almost surely, $\|(u_1^{(t)})_W\|$ and $\|(u_2^{(t)})_W\|$ converge to 0. Hence, it is enough that we show that almost surely $|\alpha((u_1^{(t)})_V) - \alpha((u_2^{(t)})_V)|$ (in $S^1$ distance) converges to zero.

To this end, let $\delta > 0$. By uniform continuity of $g_v$, we know that for small enough value of $r$, we have

$$|g_v(\alpha, r) - g_v(\alpha, 0)| < \frac{1-k}{4} \cdot \delta$$

for every $\alpha \in [0, 2\pi)$, where $k$ is the Lipschitz constant from (3.13). Therefore, almost surely, for $t$ large enough, for $u_1^{(t)}$ and $u_2^{(t)}$ parameterized as $u_1^{(t)} = (\alpha_1, r_1)$ and $u_2^{(t)} = (\alpha_2, r_2)$ we have

$$\begin{aligned}|g_v(\alpha_1, r_1) - g_v(\alpha_2, r_2)| &\leq |g_v(\alpha_1, r_1) - g_v(\alpha_1, 0)| + |g_v(\alpha_1, 0) - g_v(\alpha_2, 0)| + |g_v(\alpha_2, 0) - g_v(\alpha_2, r_2)| \\ &\leq \frac{1-k}{4} \cdot \delta + k \cdot |\alpha_1 - \alpha_2| + \frac{1-k}{4} \cdot \delta \leq \left(k + \frac{1-k}{2}\right) \cdot \max(|\alpha_1 - \alpha_2|, \delta) \ .\end{aligned}$$

Since $k + (1-k)/2 < 1$, and applying the same argument to $f_{v'}$, we conclude by induction that the distance $|\alpha_1(t) - \alpha_2(t)|$ must decrease and stay below $\delta$ in a finite number of steps. Since $\delta > 0$ was arbitrary, it must be that $|\alpha_1(t) - \alpha_2(t)|$ converges to 0, concluding the proof of Theorem 13. □

It remains to prove Lemma 8:

*Proof.* Proof of Lemma 8. Recall that we assumed a two-dimensional coordinate system on $V$. Let $f := f_{(1,0)}$, i.e., $f$ corresponds to the intervention along the $x$-axis in this coordinate system. Clearly, functions $f_v$ and $f_{v'}$ are cyclic shifts of $f$ modulo $2\pi$. More precisely, we have

$$f_{\overline{v}}(\alpha) = \alpha(\overline{v}) + f(\alpha - \alpha(\overline{v})) \ , \tag{3.14}$$

where arithmetic in (3.14) is modulo $2\pi$. Furthermore, $f$ is symmetric around the

intervention vector, i.e., $f(\alpha) = 2\pi - f(2\pi - \alpha)$ for $0 < \alpha \leq \pi$. Hence, to prove that $f_v$ and $f_{v'}$ restricted to $\text{cone}(v, v')$ are contractions, it is enough that we show that $f$ restricted to the interval $[0, \theta]$ is a contraction (recall that we assumed $\cos^2(\theta) > 1/(2+\eta)$).

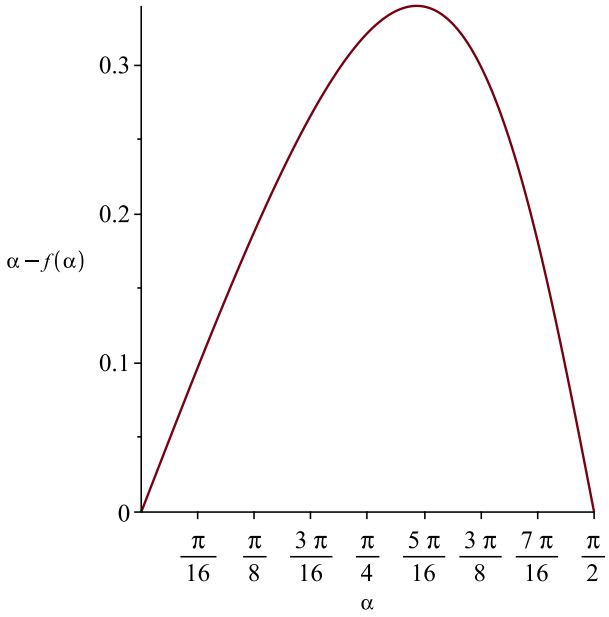


Figure 3-6: The graph of the "pull function" $\alpha - f(\alpha)$ in case $\eta = 1$.

To that end, we use (3.2) to calculate the formula for $f$ for $0 \leq \alpha \leq \pi/2$ as

$$f(\alpha) = \arccos\left(\frac{(1+\eta)\cos\alpha}{\sqrt{1+(2\eta+\eta^2)\cos^2\alpha}}\right) . \tag{3.15}$$

More computation using elementary calculus (we omit the details) establishes that, additionally, for every $0 \leq \alpha < \beta \leq \pi/2$:

1. $f(\alpha) \leq \alpha$. In other words, applying the intervention brings vector $u$ closer to the intervention vector.

2. $f(\alpha) < f(\beta)$, i.e., applying the intervention does not change relative ordering of vectors wrt the intervention vector.

3. If $\beta \leq \theta^* := \arccos\left(\sqrt{\frac{1}{2+\eta}}\right)$, then $0 \leq \alpha - f(\alpha) < \beta - f(\beta)$, i.e., in absolute

terms, the "pull" on a vector is stronger the further away it is from the intervention vector (until the correlation reaches the threshold $1/\sqrt{2+\eta}$, cf. Figure 3-6).

The preceding items taken together imply that for every $0 \leq \alpha < \beta \leq \theta^*$ we have $0 < f(\beta) - f(\alpha) < \beta - \alpha$. To conclude that $f$ is a contraction, we observe that $f$ and its derivative $f'$ are continuous on the interval $[0, \theta^*]$. If there exist sequences $(\alpha_k)$ and $(\beta_k)$ in $[0, \theta]$ for $\theta < \theta^*$ such that $|f(\alpha_k) - f(\beta_k)|/|\beta_k - \alpha_k|$ converges to 1, then, by compactness, there exist convergent sequences $\alpha_k \to \alpha^*$ and $\beta_k \to \beta^*$ such that $|f(\alpha_k) - f(\beta_k)|/|\beta_k - \alpha_k| \to 1$. Then,

1. Either $\alpha^* \neq \beta^*$ and by continuity we get $f(\beta^*) - f(\alpha^*) = \beta^* - \alpha^*$, contradicting the third property above.

2. Or $\alpha^* = \beta^*$, which by continuity of $f'$ implies $f'(\alpha^*) = 1$ for some $0 \leq \alpha^* < \theta^*$. But that would imply that the derivative of $\alpha - f(\alpha)$, i.e., $1 - f'(\alpha)$, vanishes at $\alpha^* < \theta^*$, again contradicting the third property above (see also Figure 3-6). $\square$

# Chapter 4

# A Median Voter Theory for Liquid Democracy

## 4.1 Introduction

### 4.1.1 Liquid Democracy and Problem Set-up

Liquid democracy is an election paradigm first proposed in [67] that attracts increasing attention and experimentation from Pirate Party's implementation to recent years [13, 22, 36, 58, 44, 54, 77, 21, 98, 46]. It is a delegable proxy voting scheme where each voter can choose to vote directly or delegate to another voter on a given issue. The delegation is transitive.

Previous works have examined liquid democracy both theoretically and empirically. Most existing works analyze liquid democracy for decision problems where a "correct" outcome exists, and analyze its voting quality compared to traditional direct democracy under this *epistemic* assumption [54, 46, 44]. The intuitive advantage of liquid democracy in this setting is clear: delegating to others with more domain knowledge can lead to better collective decision-making with lower aggregate effort and time invested in learning about the issue at hand for each individual. Under such assumptions, several papers have studied delegation models where voters delegate to others with higher voting accuracy, randomly, or with more sophisticated mechanisms

on underlying social networks [54, 46, 44, 21].

However, there is an orthogonal dimension of interest for democracy that is equally important and in fact prevalent in practice, where no "correct" outcome exists, and voters possess heterogeneous preferences according to which they evaluate policy candidates. This setting reflects many realistic elections including general presidential elections and many public policy regulation decisions involving stakeholders with contrasting interests. Heterogeneous preferences is also a common assumption for models of voting in game theory and political economy literature [31, 42, 88, 30]. Whether and how liquid democracy is advantageous in this ideological setting is an interesting open direction. Philosophically, liquid democracy's effectiveness in this domain seems well-grounded: voters not only want to delegate to voters with better knowledge, but also those with similar preferences.

Relatively few existing works study liquid democracy under heterogeneous preferences. One such work is [21]. It models delegation and voting in liquid democracy as a game where voters' preferences are modeled as discrete-valued types $\{0, 1\}$, i.e., each voter prefers either the outcome 0 or 1. In addition, voters also vary in their ability to cast votes that correctly match their preferences, defined as accuracy $q_i$, the likelihood of voter $i$ casting a vote matching her type, after paying a learning cost $e_i$. The goal for each voter is to maximize the probability that she casts a vote manifesting her true type (i.e., *correctly expressing*). The authors derive pure-strategy Nash equilibrium where members of a connected components delegate to voters of the same type with the highest voting accuracy. Therefore, domain knowledge is still the fundamental force that motivates the delegation.

We take a different angle in modeling, aiming to emphasize how a voter in a continuous preference spectrum delegate when learning about policies are costly (for example, would a voter delegate to the less extreme voter, or a more extreme voter?). By doing so, we capture a key motivation behind the proposal of liquid democracy: voters may not have enough information about the policy at hand to make informed choices, but may have information about other voters they are familiar with. One point of departure from the previous game-theoretic model is the continuous-valued

voter preference in our model. In our model, voters have heterogeneous and continuous ideological preferences, reflecting reality more closely, and also similar to the single-peaked preferences prevalent in game theory.

Different from the previous rational model of liquid democracy [21, 98], we adopt an outcome-oriented aspect in the modeling of utility. In [21], each voter maximizes his probability of casting a vote according to his type, instead of maximizing the probability that his type wins. Maximizing the probability of correct expression is not always equivalent to maximizing probability of the best outcome. In the current model, in contrary, the utility is derived from the result of the election directly.

Another point of departure from previous model is the continuous policy prior. We derive equilibrium results with policy candidates drawn from a probability distribution instead of being fixed $\{0, 1\}$. The modeling choice is motivated by the question of whether a formed delegation network can be reused for various decision making; so even facing multiple rounds of voting, as long as voters still expect policies to be drawn from a similar pool, the delegation they form can be extended for multiple rounds of voting.

Transitivity is a key feature and assumption in liquid democracy. "Trust" is assumed to be transitive, as for liquid democracy to work properly in an ideal case, voters delegating to neighbors they trust ideally leads to a voter they trust at the end of the chain. This assumption is not problematic in models with correct outcomes and voters delegating to better informed individuals. However, it will be interesting to see whether this transitivity works out when voters have heterogeneous preferences, because myopic and locally sound delegation could potentially lead to someone with opposite or much more extreme preferences. In this paper, we will show that this transitivity holds directionally. In particular, whether to delegate to less extreme or more extreme voters depend on the learning cost and one's stance in the political spectrum, and chains break naturally once reaching someone with deciding voting power. Or more precisely, the coalition-proof Nash equilibriums derived in this paper can be constructed by local delegations.

To this end, the current paper aims to bring the element of preference and ideology

into the picture and model how rational delegation plays out among a population of voters with their political stance and ideologies varying continuously. We will first recapture some familiar incentive structure for voting in populations with continuous preferences, and show how this incentive structure motivates the formation of different delegation networks in equilibrium.

### 4.1.2 Main Results

We study pure-strategy coalition-proof Nash equilibrium in the a liquid democracy voting game where voters have heterogenous and continuous preferences. We are able to recapture some of the classical results on voting derived in theory and observed in practice such as median voter theorem and the relationship between turnout and voters' political stance in a left-right spectrum. Political economy literature has designated extensive debate on both why voters turn out and who have higher probability of turning out. For example, there are arguments and supports both for moderate voters to have low turnout incentives as in "swing-voters curse" and extreme voters to have lower voting incentives as no policies represent them well [42]. Our model shows that the form of the policy prior is one possible mechanism for the various incentive structure to form. For example, low incentive of learning for moderate voters could occur naturally when a group of voters with single-peaked preferences facing unknown policies drawn from a single probability distributions (Proposition 8).

Echoing classical results that policy favored by the median voter wins in elections where voters have single-peaked preferences [50, 19], when allowed delegation, delegating to median forms coalition-proof Nash equilibriums (when median voters have enough incentive to turn out), as shown in Theorem 14 and 15. However, as learning cost increases, a region of disincentivised voters forms in the middle of the political spectrum. In particular, when learning cost is moderate, Theorem 16 shows that new structure of coalition-proof NE emerges where extreme voters delegate inward and moderate voters delegate outward to the most moderate voter who is still incentivised to learn. Non-trivial delegation to opposite political spectrum occurs in order to rule out more unfavored coalitions.

The analysis indicates that liquid democracy provides some remedy for moderate voters' expression of preference through delegation when learning cost is moderate, but may cause all motivated voters to vote randomly when cost of learning exceeds a threshold.

### 4.1.3 Related Work

#### Theoretical and empirical studies on liquid democracy

Several models have been proposed to analyze liquid democracy previously, with focuses on settings where a correct outcome exists [54, 46, 44]. Both supportive and unfavoring conclusions are drawn. Procaccia etal. [54, 44] present condition for the mechanism to lead to super voters which will distort the collective social choice accuracy. Halpern etal. [46] classified various delegating mechanisms and provides condition for the final voting outcome to surpass direct majority vote. Bloembergen etal. [21], a previous rational model for delegation characterizes graph conditions for the underlying social network so that voting equilibrium exists and characterize the equilibrium delegation path on the graphs.

Liquid democracy has also been implemented and experimented in real-world decision makings, including Pirate party in Germany and Google [58, 47, 85, 13]. Previous empirical studies have examined liquid democracy's performance in practice in corporate, non-corporate organizations and in political parties[58, 47, 85]. It has been observed that a common drawback of liquid democracy is the emergence of super voter that occupies disproportional decision weights.

#### Median voter theorem and political economy

Median voter model first appeared in [50] as an informal assertion and observation that political parties gravitate towards the position occupied by median voters. On the other hand, from a social choice theory perspective (thus viewing the policies as given instead of strategically generated), [19] first proved Median voter theorem, showing that in ranked preference elections, any voting rule that satisifies Condorcet

criterion elects candidate closest to the the median's preference.

At the core of median voter theories in is a connection between some characteristics of the voting population and the policy outcome. It abstracts away other features of the political process and provide testable implications from some characteristics of voting population to the policy outcome [29]. Thus it will be interesting to see whether and how delegative voting like liquid democracy retains this bridge.

**Game-theoretic modeling of voting: incentives, turnout and outcome**

The political economy literature has designated extensive attention to voting, providing plural of models that intend at explaining voting behavior, including voter turnouts, incentives, and winning policies. For example, it has been observed both theoretically and empirically that more moderate voters may have lower turn-out rate [42, 88, 31]. This resonates with the incentive structure that the current model has.

## 4.2 Model

Consider a population of $n$ agents $\mathcal{N} = \{1, 2, \cdots, n\}$, where $n$ is an odd number.[1] The agents are to vote and collectively elect a policy from two candidates $p_0, p_1 \in \mathbb{R}$. We use $v_i \in \{0, 1\}$ for agent $i$'s vote on the policies and write $v$ for the vote profile. The policy will be selected using *majority rule*, i.e., the policy that derives more votes wins. Specifically, let $y \in \{0, 1\}$ denote the index of the policy that is elected, and $y = h(v)$ where $h(v) = 1$ if and only if $\sum_i v_i > \frac{n}{2}$. The agents have heterogeneous preferences on an elected policy: they prefer a policy that is closer to their own ideological bias as will be explained later.

However, at the beginning each agent does not observe the value of the policies $p_0, p_1$. Each agent can vote by herself and choose whether to learn the policies. Alternatively, the agent can choose to delegate her vote to another agent and her vote

[1] We set $n$ to be odd to have a single median and to guarantee existence of coalition-proof pure strategy Nash equilibrium of the form in Theorem 14. In the low learning cost scenario (Section 4.4.2), under some settings of if the middle two voters are sufficiently far away from each other such that each prefers random voting than other's choice in expectation, one of the middle two voters being the dictator can be not be a cpNE.

will be the same as the delegated agent's vote. We provide the timeline and specifics of the game as below.

**Timeline.**

The game proceeds in three stages $0, 1, 2$. At stage 0, Nature draws two policies that are unobservable to the agents. At stage 1, each agent submits her action, consisting of a tuple $(d_i, \ell_i)$, where $d_i \in \mathcal{N}$ names the voter that $i$ delegates to, and $\ell_i \in \{0, 1\}$ represents $i$'s learning decision, i.e., if $\ell_i = 1$, voter $i$ commits to perfectly observe both of the policies at a learning cost in stage 2. At stage 2, each agent $i$ with $\ell_i = 1$ observes $p_0, p_1$ (privately) at a learning cost. Each agent who has not delegated casts her vote on the policies. All of these settings are common knowledge.

**Priors on the policies.**

We assume that the agents share a common prior on the policies; in particular, they agree that the policies are independently drawn according to a probability density function $f$, i.e.,

$$p_0, p_1 \sim f. \tag{4.1}$$

We assume that $f$ is symmetric around 0.[2]

**Delegation network.**

At stage 1, each agent $i$ simultaneously chooses whether to delegate her vote to another agent (or herself). If she selects an agent $j$, her vote will be the same as the delegated agent $j$'s vote, i.e., $v_i = v_j$. We denote agent $i$'s delegation decision as $d_i \in \mathcal{N}$ and $d_i = i$ when agent $i$ decides to vote by herself. Therefore, a *delegation network* $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is induced by the agents' delegation decisions ($(i, j) \in \mathcal{E}$ if and only if $d_i = j$). The emergent delegation network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is observed by all the agents.

Note that the delegation exhibits *transitivity*: If agent $i$ delegates to agent $j$ and

[2]We will briefly discuss in Section 4.4.1 on how the form of prior influence the incentive structures of the current model, and how the analysis techniques for deriving cpNEs generalize.

agent $j$ delegates to $k$, then $v_j = v_k$ and then $v_i = v_j$.[3] Without loss of generality, we consider delegation profiles where $d_i$ represents the end of delegation chain starting from $i$; this is because all voters know any other voters' actions, and can predict where his delegation ends, and thus he always has the ability to directly delegate in one step.

**Learning, voting, and utility functions**

At stage 2, agents simultaneously cast their votes. If an agent delegated her vote to another, then her vote is determined via the delegation network $\mathcal{G}$ as explained. If an agent did not delegate, she makes her decision on the policies. A policy $y = h(v)$ is then elected based on majority rule. Agent $i$'s utility derived from her learning decision and the election is given by

$$u_i(p_0, p_1, v, b_i) = -(p_{h(v)} - b_i)^2 - c\mathbb{1}\{\ell_i = 1\}, \tag{4.2}$$

where $b_i \in \mathbb{R}$ represents agent $i$'s *ideological bias*. Agent $i$ prefers a policy that is closer to her ideological bias. Agents' ideological biases are common knowledge. We sometimes refer to the set of voters directly by the set of their biases, i.e., $\mathcal{B} = \{b_1, \cdots, b_n\}$.

## 4.3 Preliminary

We will solve the delegation game using solution concepts coalition-proof Nash equilibrium [18] and Strong Nash equilibrium [3]. While Nash equilibrium concerns with unilateral deviations, both of these concepts concern with deviations by subsets of players, rendering them especially suitable for studying voting games. Due to the nature of most voting schemes, deviation by one agent rarely changes the outcome of the game, making many arbitrary and unrealistic voting profiles Nash equilibria. Allowing deviations by groups of agents provides an important way for equilibrium refinement. Therefore, we focus on deriving coalition-proof Nash equilirbriums for

[3]When a delegation cycle forms, it is accounted as no one in the cycle casts a vote. We generally do not need to worry about this in our model, because voters know any other voter's action.

liquid democracy game,

A Strong Nash equilibrium is a Nash equilibrium where there does not exist a subset of players $J$ accompanied with a deviating strategy $s_J$ who find it beneficial to deviate (everyone in that coalition becomes strictly better off). Formally, for $n$-player game with strategy sets $\{S^j\}_{j=1}^n$ and payoff functions $\{u^j : \prod_{i=1}^n S^i \to \mathbb{R}\}_{j=1}^n$, Strong Nash equilibrium is defined as follows.

**Definition 16** (Strong Nash Equilibrium [3, 18])**.** *$s^* \in \prod_{j=1}^n S^j$ is a Strong Nash equilibrium if and only if for all $J \subseteq \{1, \cdots, n\}$, and for all $s_J \in \prod_{j\in J}^n S^j$, there exists an agent $i \in J$ such that $u_i(s^*) \geq u_i(s_J, s^*_{-J})$, where $s^*_{-J} := \{s^*_j\}_{j\notin J}$.*

In contrast, a coalition-proof Nash equilibrium is a slightly weaker notion that is defined against deviations where no sub-coalitions have incentive to further deviate. A coalition-proof Nash equilibrium is a Nash equilibrium where no subset of players benefit from deviating in a self-enforcing way, i.e., such that no sub-coalition benefit from further deviating.

Formally, consider an $n$-player game $\Gamma = [\{u^i\}_{i=1}^n, \{S^i\}_{i=1}^n]$, where $S^i$ is player $i$'s strategy set and $u^i : \prod_{j=1}^n S^j \to \mathbb{R}$ is player $i$'s payoff function. Let $\mathcal{J}$ be the set of proper subsets of $\{1, \cdots, n\}$, and denote an element of $\mathcal{J}$ (a coalition) as $J \in \mathcal{J}$. Let $S^J := \prod_{i\in J} S^i$. Let $-J$ denote the complement of $J$ in $\{1, \cdots, n\}$. Finally, for each $s'_{-J} \in S^{-J}$, let $\Gamma \setminus s'_{-J}$ be the game induced on subset $J$ by the actions $s'_{-J}$ of coalition $-J$, i.e.,

$$\Gamma \setminus s'_{-J} := \left[\{\tilde{u}^i\}_{i\in J}, \{S^i\}_{i\in J}\right], \tag{4.3}$$

where $\tilde{u}^i(s_J) := u^i(s_J, s'_{-J})$ for all $i \in J$ and $s_J \in S^J$.

**Definition 17** (Coalition-proof Nash Equilibrium [18])**.** *(i) In a single player game $\Gamma$, $s^* \in S$ is a Coalition-Proof Nash equilibrium if and only if $s^*$ maximizes $u^1(s)$.*

*(ii) Let $n > 1$ and assume that Coalition-Proof Nash equilibrium has been defined for games with fewer than $n$ players. Then,*

(a) *For any game* $\Gamma$ *with* $n$ *players,* $s^* \in S$ *is self-enforcing if, for all* $J \in \mathcal{J}$, $s^*_J$ *is a Coalition-Proof Nash equilibrium in the game* $\Gamma \setminus s^*_{-J}$

(b) *For any game* $\Gamma$ *with* $n$ *players,* $s^* \in S$ *is a Coalition-Proof Nash equilibrium if it is self-enforcing and if there does not exist another self-enforcing strategy vector* $s \in S$ *such that* $u^i(s) > u^i(s^*)$ *for all* $i = 1, \cdots, n$.

In other words, an agreement is coalition-proof if it is efficient within the class of self-enforcing agreements, where self-enforceability requires that no coalition (proper subset) can benefit by deviating in a self-enforcing way. Specifically, a Strong Nash Equilibrium is always a coalition-proof Nash equilibrium, because it is immune to all deviations and is efficient within the set of all strategy profiles. We will leverage this simple fact to prove the equilibrium in Theorem 14 is coalition-proof by proving it to be Strong. Note that the equilibriums in Theorem 16 is a cpNE (not Strong NE). We will elaborate in remarks in Section 4.4.3 on which non-self-enforcing deviation exist for the equilibrium strategy profile in Theorem 16.

### 4.3.1 Terminologies for liquid democracy game

Each liquid democracy game with voters $\mathcal{N}$ with preference peaks $\{b_1, \cdots, b_n\}$, learning cost $c$ and prior $\mathbb{P}$, denote the median vote as $m$. Each equilibrium strategy profile $s$ induces a *delegation graph* $\mathcal{D}_s$, a directed graph with directed edges corresponding to delegations. As all voters are aware of any other voter's action in a Nash equilibrium, we can without loss of generality focus on delegation graphs with only direct delegations. (The diameter of the graph is 1).

Each equilibrium strategy profile also induces a *decision rule*, a function (deterministic or random, depending on whether there exist voters who conceive mixed strategy in the equilibrium profile $s$) that maps each pair of $p_0, p_1$ to the outcome that the equilibrium produces. We denote it as $f_s : \mathbb{R}^2 \to \mathbb{R}$.

For example, the strategy profile where every voter always picks policy 0 induces a constant decision rule $f(x, y) = x$. The strategy profile where all voters delegate to a voter with bias $b_i$ induces a decision rule $f(x, y) = \text{argmin}_{p \in \{x,y\}}(p - b_i)^2$. The

strategy profile where all voters delegate to a voter that votes uniformly at random induces a decision rule,

$$f(x,y) = \begin{cases} x & \text{w.p. } 1/2 \\ y & \text{w.p. } 1/2 \end{cases} \tag{4.4}$$

Obviously, the first two example rules are deterministic (for every input pair $(x, y)$, the output is deterministic), the last example rule is a random function. Note again that the expected utility from the first rule (arbitrary voting) and the third rule (random voting) are the same for every voter.

We define *incentive of learning* according to an individual rationality constraint. Individual rationality means that if the game consists of a voter by herself, she prefers learning about the policies to voting randomly.

**Definition 18** (Individual rationality)**.** *We say that a voter $i$ satisfies individual rationality (IR) if her expected gain from learning and picking the policy that she favors compared to being assigned a policy candidate randomly equals or exceeds the learning cost $c$, i.e.,*

$$\mathbb{E}\left[\max_{p\in\{p_0,p_1\}} u_i(p)\right] - \mathbb{E}u_i(p_0) \geq c, \tag{IR}$$

*where $p_0, p_1 \sim \mathbb{P}$, independently.*

IR is a necessary condition for any voter to submit $\ell_i = 1$ in the game. This is because if an agent does not satisfy (IR), even when she gains all delegations and is the dictator of the election, she still prefers voting randomly to learning, thus $\ell_i = 1$ is strictly dominated.

In addition, we naturally define the quantity on the left-hand side of (IR), $\left[\max_{p\in\{p_0,p_1\}} u_i(p)\right] - \mathbb{E}u_i(p_0)$ as voter $i$'s *incentive of learning*.

**Definition 19** (Incentive of learning $\mathcal{I}_i$)**.** *Voter $i$'s incentive of learning is defined as $\mathbb{E}\left[\max_{p\in\{p_0,p_1\}} u_i(p)\right] - \mathbb{E}u_i(p_0)$, denoted as $\mathcal{I}_i$.*

In particular, plugging in the quadratic utility as defined in our model, we slightly abuse notation and write that a voter with bias $b$ has incentive of learning $\mathcal{I}(b) =$

$\mathbb{E}\left[\max_{p\in\{p_0,p_1\}} -(b-p)^2\right] - \mathbb{E}\left[-(b-p_0)^2\right]$.

At times we would need to evaluate a voter's expected untility under another voter's decision rule. We define three relevant concepts as follows.

**Definition 20** (Choice function $f_i$)**.** *The choice function of voter $i$ maps policy candidates to the policy that maximizes $i$'s utility, i.e., chooses the policy closer to the voter $i$'s bias $b_i$: $f_i(p_0,p_1) = \text{argmin}_{p\in\{p_0,p_1\}}|p-b_i|$.*

Note that for any utility function that is essentially a distance metric, $u_i(y) \cong d(y,b_i)$ (in our model, the distance is Euclidean distance), choice function is of this form.

**Definition 21** ($\mathbb{E}u_i(f_j)$)**.** *Voter $i$'s expected utility under voter $j$'s choice function is defined as $\mathbb{E}u_i(f_j) = \mathbb{E}\left[-\left(b_i - \text{argmin}_{p\in\{p_0,p_1\}}|p-b_m|\right)^2\right]$*

For example, in a strategy profile where all voters delegate to the median voter $m$, a voter $i$ with bias $b_i$'s expected utility is $u_i(f_m) = \mathbb{E}\left[-\left(b_i - \text{argmin}_{p\in\{p_0,p_1\}}|b_m-p|\right)^2\right]$.

**Definition 22** (Indifferent ball $B_i$)**.** *We say that a voter $j$'s bias $b_j$ is in the indifferent ball of voter $i$ with bias $b_i$, $b_j \in B_i$, if $i$ weakly prefers delegating to $j$ to voting himself, i.e.,$u_i(f_i) - c \le u_i(f_j)$, or elaborately:*

$$\mathbb{E}\left[\max_{p\in\{p_0,p_1\}} -(b_i-p)^2\right] - c \le \mathbb{E}\left[-\left(b_i - \operatorname*{argmin}_{p\in\{p_0,p_1\}}|b_j-p|\right)^2\right]. \tag{4.5}$$

Notice that $B_i$ becomes larger as $c$ increases. With quadratic-formed utility, $B_i$ in general is not symmetric around $b_i$.

### 4.3.2 Discussion of model constructs

**$n$ is odd**

We require $n$ to be odd in order to have one unique median. Even values of $n$ create equilibriums very sensitive to specific settings of learning cost $c$ through the possibility of tie-breaking, which is distracting from the main form of equilibrium we would like

to illustrate (i.e., under regimes of low learning cost, delegation aggregate inward, and under regimes of relatively high learning cost, central spectrum delegate outward). For example, Theorem 14 would not hold for even $n$. For many specific combination of learning cost and voter distribution, one of the middle voter votes and everyone delegating to him is a cpNE. However, observe that for any set of voters, if we make learning cost $c$ small enough, this is no longer a cpNE, while both middle voters learn and vote, remaining voters delegate to the one closer to them is a cpNE (because each of them prefers random tie-breaking when they disagree to defaulting to the other's preferred policy). This type of tie-breaking behavior and resulting multitude of equilibria highly sensitive to specific values of $c$ seems more distracting than illuminating of the core forces at work of the game, so we restrict to odd values of $n$ each learning cost regime has a unified form of equilibrium.

**Functional form of utility function** $u(\cdot)$

The current analysis uses the specific quadratic form of utility function in two places. One is in the proof of monotonicity of the incentive structure (i.e., extreme voters have more incentive to learn than the middle voters), Proposition 8. The second is in the proof for convexity (i.e., the utility function of any voter with bias between voter $A$ and voter $B$ can be expressed as a convex combination of $u_A(\cdot)$ and $u_B(\cdot)$ plus a constant).

The argument of Prop 8 heavily relies on the quadratic form of the utility function by expanding out the quadratic and makes an optimization-style argument.

The exact claim of convexity is not true for general utility functions since general utilities may contain higher order terms of the bias parameter, but more general form of convexity for general utility function may hold. For example, for any form of single-peaked utility that can be written as a distance metric (i.e., $u_i(y) \cong d(y, b_i)$ for some metric $d$, when $A$ and $B$ prefer a decision rule $f$ to $g$, anyone with bias between $A$ and $B$ also does, circumventing this exact convex combination argument for general distance functions.

This been said, it is a open question which other single-peaked and symmetric

(around the peak) utility functions Prop 8 extends to. We conjecture that concavity may be a sufficient condition. However, for any other single-peaked and symmetric utility form to which Prop 8 does extend, Theorems 14,15,16,17 extend immediately.

**Prior, separate priors, and prior's form** In general, the probability from which policies $p_1$ and $p_2$ are generated from plays an important role on the incentive structure of voters. Generally speaking, when the two policies share the same prior, the position of the peak of this prior can change which voter has the most incentive to learn (therefore, this shifts the incentive structure). When the two policies are drawn from different priors, voters possess prior information on how the two policies differ. In this case, monotonicity property of the incentives can be partially reverted (i.e., on the tail of the spectrum, the incentive of learning decays with $|b|$, more extreme voters have lower incentives to learn). We will expand on this in Section 4.4.1.

## 4.4 Results on Pure-strategy Coalition-proof Nash Equilibria

In this section, we present our results on pure-strategy Nash equilibrium of liquid democracy voting game. We will show that the form of pure-strategy delegation profile varies with the learning cost parameter $c$. In particular, when learning cost is low enough, all voters delegating to the median voter is a coalition-proof Nash equilibrium. As learning cost increases, moderate voters lose incentive to learn, someone less moderate but still retain IR becomes the dictator. Interestingly, all other voters (including those from the opposite extreme of the political spectrum) delegating to this voter forms a coalition-proof Nash equilibrium.

### 4.4.1 Incentive structure

First, as the main building block of the equilibrium analysis, we study the incentive structure in this game. We observe that in this model, with any unbiased and

symmetric prior distribution, voters' learning incentives increase with their biases' distance from 0.

**Proposition 8** (Monotonicity of incentive of learning $\mathcal{I}(b)$)**.** *Let $p_0, p_1$ be drawn independently from any continuous prior distribution $\mathbb{P}$ that is symmetric around 0, then incentive of learning, $\mathcal{I}(b) = \mathbb{E}\left[\max_{p\in\{p_0,p_1\}} -(b-p)^2\right] - \mathbb{E}\left[-(b-p_0)^2\right]$, monotonically increases with voter's bias $|b|$.*

This monotonicity may be shown through explicit calculation when the functional form of the prior distribution $\mathbb{P}$ is known and may even be extended to more general forms of utility functions that induce distances among $b_i$'s. If the functional form of the prior is not known or intractable, this monotonicity claim is still true due to the following general argument. Note that the argument we provide here is dependent on the quadratic functional form of the utility.

*Proof.* We will start by proving a voter with a positive bias $b_i$ has higher incentive to learn than the voter with bias 0. We will then prove through a similar argument that an voter with bias $b+\varepsilon$ has higher incentive to learn than the voter with bias $b > 0$, for any $b > 0, \varepsilon > 0$. By symmetry, an voter with a negative $b_i$ has higher incentive to learn than the voter with peak 0, and voter with $b_i - \varepsilon > 0$ has more incentive to learn than $b_i < 0$, for any $\epsilon > 0$. These together complete the proof that incentive of learning $\mathcal{I}(b)$ monotonically increases with $|b|$.

First, we compare incentive of learning $\mathcal{I}(b)$ to $\mathcal{I}(0)$, for any $b > 0$. We want to show $\mathcal{I}(b) - \mathcal{I}(0) > 0$. The first term $\mathcal{I}(b) = \mathbb{E}\left[\max_{q\in\{p_0,p_1\}} -(b-q)^2\right] - \mathbb{E}\left[-(b-p)^2\right]$, and the second term $\mathcal{I}(0) = \mathbb{E}\left[\max_{q\in\{p_0,p_1\}} -q^2\right] - \mathbb{E}\left[-p^2\right]$. Substract the second from the first and simplify,

$$\mathcal{I}(b) - \mathcal{I}(0) = \mathbb{E}\left[\max_{q\in\{p_0,p_1\}} -(b-q)^2\right] - \mathbb{E}\left[-(b-p)^2\right] - \mathbb{E}\left[\max_{q\in\{p_0,p_1\}} -q^2\right] + \mathbb{E}\left[-p^2\right] \tag{4.6}$$

$$= -\mathbb{E}\min_{q\in\{p_0,p_1\}}(b-q)^2 + \mathbb{E}\min_{q\in\{p_0,p_1\}} q^2 + \mathbb{E}b^2 - \mathbb{E}2bp \tag{4.7}$$

$$= -\mathbb{E}\min_{q\in\{p_0,p_1\}}(b-q)^2 + \mathbb{E}\min_{q\in\{p_0,p_1\}} q^2 + b^2 \tag{4.8}$$

$$= -\mathbb{E}\min\{p_0^2 - 2bp_0, p_1^2 - 2bp_1\} + \mathbb{E}\min\{p_0^2, p_1^2\}. \tag{4.9}$$

We can in fact prove Eq.(4.9)$>0$ through an argument on the incidence level. Facing any incidence of two policies $\{p_0, p_1\}$, the optimization problem of the voter with bias $= 0$ is equivalent to $\min\{p_0^2, p_1^2\}$, while voter with a positive bias $b$ solves $\min\{p_0^2 - 2bp_0, p_1^2 - 2bp_1\}$, $b > 0$. Let $\bar{p}$ be the policy that voter with bias 0 prefers, $\bar{p} := \text{argmin}\{p_0^2, p_1^2\}$. Notice that a feasible strategy for voter $b > 0$ is to copy and also always choose $\bar{p}$. Therefore, $\min\{p_0^2 - 2bp_0, p_1^2 - 2bp_1\} \leq \bar{p}^2 - 2b\bar{p}$, for any realization of $p_0, p_1$, thereby, $\mathbb{E}\min\{p_0^2 - 2bp_0, p_1^2 - 2bp_1\} \leq \mathbb{E}\left[\bar{p}^2 - 2b\bar{p}\right]$.

Now, find any set of incidences with positive measure where voter $b$ has a better strategy than following $\bar{p}$. For example, anytime $p_0, p_1$ both fall in $[0, b]$, which occurs with probability $\mathbb{P}\left(p \in [0, b]\right)^2 > 0$, voter $b$ can choose the policy closer to $b$ instead of $\bar{p}$, positively improving his utility. Therefore, the inequality holds strictly,

$$\mathbb{E}\min\{p_0^2 - 2bp_0, p_1^2 - 2bp_1\} < \mathbb{E}\left[\bar{p}^2 - 2b\bar{p}\right] = \mathbb{E}\bar{p}^2. \tag{4.10}$$

The last equality follows from $\mathbb{E}\bar{p} = 0$. With $p_0, p_1 \sim \mathbb{P}$ drawn from prior distribution $\mathbb{P}$ symmetric around 0, the expectation of the optimal policy $\bar{p}$ is 0. ($\bar{p}(p_0, p_1) + \bar{p}(-p_0, -p_1) = 0$, thus $\mathbb{E}\bar{p} = 0$.) The quantity in Equation (4.9) is strictly positive.

In fact, this type of argument extends to show that $b + \varepsilon$ has more incentive to learn than $b$, for any $b > 0, \varepsilon > 0$. Analogous to the term in Eq (4.9), we would like to show that the difference of learning incentive between voter $b + \varepsilon$ and voter $b$ is

greater than 0.

$$\mathcal{I}(b+\varepsilon) - \mathcal{I}(b) \tag{4.11}$$

$$= \mathbb{E}\left[\max_{p_0,p_1} -(b+\varepsilon-p)^2\right] - \mathbb{E}\left[-(b+\varepsilon-p)^2\right] - \left(\mathbb{E}\left[-\max_{p_0,p_1}(b-p)^2\right] - \mathbb{E}\left[-(b-p)^2\right]\right) \tag{4.12}$$

$$= -\mathbb{E}\min_{p_0,p_1}\{p_0^2 - 2bp_0 - 2\varepsilon p_0, p_1^2 - 2bp_1 - 2\varepsilon p_1\} + \mathbb{E}\min_{p_0,p_1}\{p_0^2 - 2bp_1, p_1^2 - 2bp_1\}, \tag{4.13}$$

Again, facing any incidence of two policies $\{p_0, p_1\}$, let $\tilde{p}$ be the policy that voter with bias $b$ prefers, i.e., $\tilde{p} = \text{argmin}\{p_0^2 - 2bp_2, p_2^2 - 2bp_1\}$. Copying voter $b$'s choice is a feasible strategy for voter $b+\varepsilon$. Therefore,

$$\mathbb{E}\min\{p_0^2 - 2b_i p_0 - 2\varepsilon p_0, p_1^2 - 2b_i p_1 - 2\varepsilon p_1\} \tag{4.14}$$

$$\leq \mathbb{E}\left[\tilde{p}^2 - 2b_i\tilde{p} - 2\varepsilon\tilde{p}\right] \tag{4.15}$$

$$= \mathbb{E}\left[\tilde{p}^2 - 2b_i\tilde{p}\right] - \mathbb{E}\left[2\varepsilon\tilde{p}\right] \tag{4.16}$$

Note that for a voter with positive $b$, her expected optimal policy is positive, i.e., $\mathbb{E}\tilde{p} > 0$. One way to see this is that for any incidence of $\{p_0, p_1\}$, voter $b$'s optimal policy is always greater than or equal to voter 0's optimal policy, i.e., $\tilde{p} \geq \bar{p}$, always. Furthermore, with positive probability (at least $\mathbb{P}(p_0, p_1 \in [0,b]) = \mathbb{P}(p_0 \in [0,b])^2 > 0$, since 0 and $b$ disagree for these realizations), $\tilde{p} > \bar{p}$. Therefore, $\mathbb{E}\tilde{p} > \mathbb{E}\bar{p} = 0$.

From Eq (4.16), $\mathbb{E}\tilde{p} > 0$ implies that $\mathbb{E}\min\{p_0^2 - 2bp_0 - 2\varepsilon p_0, p_1^2 - 2bp_1 - 2\varepsilon p_1\} < \mathbb{E}\left[\tilde{p}^2 - 2b\tilde{p}\right]$, i.e., the difference of incentive of learning between voters $\mathcal{I}(b+\varepsilon) - \mathcal{I}(b)$ in Eq (4.13) is positive. □

Proposition 8 is an important building block that underlies all main analysis and results including Theorems 14,16,17 in this paper. It reveals the incentive structure of the liquid democracy game with a single prior: more extreme voters have higher

incentives to learn than voters in the middle of the political spectrum.

As a result, as learning cost varies, while voters on the more extreme opinion stance still satisfy individual rationality, voters in the middle start to lose incentives to learn. In fact, as we will show in Section 4.4.2-4.4.4, learning cost is the key parameter that dictates the form of equilibrium delegation. More explicitly, when learning cost is low, all voters delegating inward to median is a coalition-proof NE. When learning cost is intermediate, one of the most moderate incentivised learner attracts all delegations (i.e., extreme voters delegate inward, and moderate voters delegate outward) forms a coalition-proof Nash equilibrium. When learning cost exceeds a certain threshold, voting arbitrarily (say always choose $p_0$) remains as the only pure-strategy Nash equilibrium.

In general, the probability from which policies $p_0$ and $p_1$ are generated from plays an important role on the incentive structure of voters. Generally speaking, when the two policies share the same prior, the position of the peak of this prior can change which voter has the least incentive to learn (therefore, this shifts the incentive structure). When the two policies are drawn from different priors, voters possess prior information on how the two policies differ. In this case, monotonicity property of the incentives can be partially reverted (i.e., on the tail of the spectrum, the incentive of learning decays with $|b|$, more extreme voters have lower incentives to learn).

It is noteworthy that upon changing of forms of priors, as long as the incentive structure of the game is classified, results similar to that in Theorems 14,16,17 can still be claimed based on the new incentive structure, through identifying the least extreme voters still with incentives to learn. When median satisfy IR, then equilibrium form in Theorem 14,15 hold. When median do not satisfy IR, then equilibrium form in Theorem 16 holds.

### 4.4.2 Results on low learning cost

We start by presenting equilibrium results for low learning cost where all voters satisfy individual rationality (IR).

**Theorem 14.** *Suppose the learning cost $c$ is low enough such that $\mathcal{I}_i \geq c$ for all voter $i$. Assume there is no other voter in median voter's indifferent ball, then median voter $m$ learns and votes truthfully, everyone delegates to $m$ is a Strong Nash equilibrium (therefore also a coalition-proof NE).*

*Proof of Theorem 14.* Assume towards contradiction that there exists a set of voters who find it beneficial to deviate. The size of this set must be $> N/2$ (with strict inequality), because otherwise the voting result will not change. This means someone from the left of the median and someone from the right of the median are both in this coalition. However, this indicates that median $m$ will also benefit from this deviation.

This is because if voter $A$ and $B$ both prefer a decision rule $f$ to $g$, then anyone, $C$, between $A$ and $B$ also prefers rule $f$ to $g$ (by writing $u_C(x) = au_A(x) + bu_B(x) + c$, point-wise for any outcome $x$. Note that this $c$ is a constant independent of $x$. Therefore $\mathbb{E}_{x\sim\mathbb{P}} u_C(x) = a\mathbb{E}_{x\sim\mathbb{P}} u_A(x) + b\mathbb{E}_{x\sim\mathbb{P}} u_B(x) + c$ , for $\mathbb{P}$ induced by any decision rule.) This means that there exists a decision rule that median strictly prefers than the decision rule induced by herself being the dictator. Such decision rule does not exist. Contradiction. □

In fact, an approximate converse to Theorem 14 exists. It states that any coalition-proof Nash equilibrium must be close to the decision rule induced by median being the dictator.

**Lemma 9.** *Suppose the learning cost $c$ is low enough such that all voters satisfy* (IR)*. Let $s$ be a coalition-proof NE of this game, and $u_m(s)$ be median voter's utility induced by this strategy profile. Then the utility that median voter derived from $s$ is at most $c$ away from the $\mathbb{E}\max_{p_1,p_2}\{u_m(p_1), u_m(p_2)\} - c \leq u_m(s) \leq \mathbb{E}\max_{p_1,p_2}\{u_m(p_1), u_m(p_2)\}$.*

In words, any cpNE yields an expected utility for the median voter at least the same as the expected utility of median voter where he owns all voting power and votes truthfully, $\mathbb{E}u_m(f_m) - c$. I.e., let $f$ be induced by any cpNE, let $f_m$ be induced by $m$ being the dictator. Then $\mathbb{E}u_m(f_m) - c \leq \mathbb{E}u_m(f) \leq \mathbb{E}u_m(f_m)$.

*Proof.* Assume towards contradiction that there exists such cpNE $s$ that gives less expected utility to the median voter than median dictator - $c$. This means if the median voter deviates to learn and vote truthfully, there does not exist a set of $(N-1)/2$ (remember that $N$ is odd) voters who find it beneficial to deviate with the median. This means that at least one voter to the left of median and one voter to the right of median find strategy profile $s$ strictly better than $s'$ (median learns and everyone delegate to median). By the same argument as the proof for Theorem 14, median also strictly prefers the decision rule induced by this cpNE to the decision rule induced by himself being the dictator. This contradicts the assumption that this cpNE yields less expected utility for the median compared to median having all the voting power. □

To further understand the form of coalition-proof Nash equilibriums in this game, we show that in fact, only one learner exists in any pure-strategy Nash equilibrium.

**Lemma 10.** *In a Nash Equilibrium, if the voting outcome is non-stochastic, then there exists at most one voter who learns,* $\sum_{i\in\mathcal{N}} \ell_i = 1$.

*Proof.* Suppose there exist $K \geq 2$ voters who learn and vote in a Nash equilibrium. Order these voters from left to right according to their ideological biases. In a Nash equilibrium, any learner votes according to her utility function (otherwise she influence the outcome with zero probability and she should not have learned in the first place).

Note that any voting function induced by liquid democracy is a weighted majority function. In any non-stochastic weighted majority function with $K$ inputs, let $j$ be the index such that $\sum_{i=1}^{j} w_i > N/2$, $\sum_{i=1}^{j-1} w_i < N/2$. Then 0 is chosen iff $v_j = 0$.

Other voters can delegate to this pivotal voter $j$ without changing the result of the election (because each such voter realizes when she agrees with the pivotal voter, she wins, when she disagrees with the pivotal voter, she loses. Delegating to the pivotal voter does not change the voting result, so her utility increases $c$). Therefore, the currently considered profile with $K \geq 2$ learners is not a Nash equilibrium. Contradiction. $K$ can only be 1. □

Note that Lemma 10 holds generally for this game, regardless of learning cost $c$.

Though Theorem 14 has a condition where median's indifferent ball contains no other voters $B_m = m$, putting results in Lemma 9 and Lemma 10 together, we can immediately derive that when there are voters in the $B_m$, any pure-strategy coalition-proof NE, if exists, has to be of the form of a unique voter in median's indifferent ball being the dictator.

**Theorem 15.** *Suppose the learning cost $c$ is low enough such that all voters satisfy* (IR)*. If a pure-strategy cpNE exists, it is of the following form: a voter in median voter's indifferent ball $B_m$ learns and votes according to her preference, and over $N/2$ voters delegating to her.*

*Proof of Theorem 15.* According to Lemma 10, if a pure-strategy cpNE exists, only one voter learns in this pure-strategy cpNE. According to Lemma 9, median voter's utility under this pure-strategy cpNE is greater than or equal to $\mathbb{E}\max\{u_m(p_0), u_m(p_1)\} - c$, which is the definition of the indifferent ball of voter $m$. Thus the unique learner resides in $B_m$. □

Finally, we separate the property of convexity that are useful for proving both Theorem 14 and Lemma 9.

**Claim 14** (Convexity of preference)**.** *When voter $A$ and $B$ both prefer a decision rule $f$ to $g$, then anyone, $C$, with bias between $A$ and $B$'s biases (i.e. $b_C = \lambda b_A + (1-\lambda) b_B$, for some $\lambda \in [0,1]$) also prefers rule $f$ to $g$.*

*Proof.* When voter $C$'s preference peak is between voter $A$ and $B$, i.e., $b_C = \lambda b_A + (1-\lambda) b_B$, we can write $u_C(x) = \lambda u_A(x) + (1-\lambda) u_B(x) + c$ with the same $\lambda$. This is because:

$$\begin{aligned}\lambda u_A(x) + (1-\lambda)u_B(x) &= -\lambda(x-b_A)^2 - (1-\lambda)(x-b_B)^2 \\ &= -x^2 - 2b_C x - \lambda b_A^2 - (1-\lambda)b_B^2 \\ &= -(x-b_C)^2 - c \\ &= u_C(x) - c,\end{aligned}$$

where $c = -b_C^2 + \lambda b_A^2 + (1-\lambda)b_B^2$, a constant independent of $x$. Each decision rule induces a probability distribution of the chosen policy, denoting $\mathbb{P}_f$. By linearity of expectation, for any decision rule $f$, voter $C$'s expected utility is a convex combination of voter $A$ and $B$'s (plus a constant):

$$\mathbb{E}_{x\sim\mathbb{P}_f} u_C(x) = \lambda \mathbb{E}_{x\sim\mathbb{P}_f} u_A(x) + (1-\lambda)\mathbb{E}_{x\sim\mathbb{P}_f} u_B(x) + c,$$

Therefore, when both $A$ and $B$ prefers a decision rule $f$ to $g$, $C$ also prefers $f$ to $g$, because $\mathbb{E}_{x\sim\mathbb{P}_f} u_A(x) > \mathbb{E}_{x\sim\mathbb{P}_g} u_A(x)$ and $\mathbb{E}_{x\sim\mathbb{P}_f} u_B(x) > \mathbb{E}_{x\sim\mathbb{P}_g} u_B(x)$, implies that $\mathbb{E}_{x\sim\mathbb{P}_f} u_C(x) > \mathbb{E}_{x\sim\mathbb{P}_g} u_C(x)$. □

When both voter $A$ and $B$ prefer a decision rule $f$ to $g$, any voter with bias between $A$ and $B$ also does. Therefore, a deviating coalition can always be extended to include any voter whose bias is a convex combination of members in the coalition.

### 4.4.3 Intermediate learning cost

When learning cost $c$ is high enough, the median voter $m$ may lose incentive to learn even when the game contains only herself. Specifically, for any cost $c$, there exists a radius $r$ such that voters with biases outside of $(-r, r)$ satisfy (IR), and others do not, a natural result of voters' learning incentives increasing with distance from the center (Proposition 8). In the following, we prove that when voters in the center lose incentives to learn, depending on the parameters (learning cost $c$ and positions of voters $\mathcal{N}$), (i) one voter learns and all voters delegating to him is a coalition-proof

Nash equilibrium, or (ii) all voters vote arbitrarily is the only coalition-proof Nash equilibrium.

**Lemma 11.** *Suppose the learning cost $c$ is such that there exist voters on both sides of 0 that satisfy IR. Let $m^-$ be the first voter to the left of the 0 that satisfies IR, denote her bias as $b^-$. Let $m^+$ be the first voter to the right of the 0 that satisfies IR, denote her bias as $b^+$. Then, any voter $i \in \mathcal{N}$ prefers delegating to of one of $m^-, m^+$ to voting randomly (equivalent to arbitrarily). Then, at least one of $m^-, m^+$ is preferred by $> N/2$ voters as a delegate to voting randomly.*

If one of $m^-, m^+$ is preferred by $> N/2$ voters as a delegate to voting arbitrarily, let it be $m^*$; if both satisfy (i), let the one preferred by more voters be $m^*$ (break tie arbitrarily).

**Theorem 16.** *Suppose the learning cost $c$ is such that there exist voters on both sides of 0 that satisfy IR. Then voter $m^*$ learns and votes according to her preference, all other voters delegate to her (directly or indirectly) is a cpNE.*

To unpack the claims here, Lemma 11 states as long as there exists incentivised voters on both side of 0, any voters between the two incentivised voters always prefers delegating to one of them to voting randomly. Theorem 16 ensures the existence of cpNE and describes the form of one cpNE under such learning cost. In the following, we first prove Lemma 11, and then present proof of Theorem 16.

*Proof of Lemma 11.* We will show that any voter $i \in \mathcal{N}$ prefers either $m^-$ or $m^+$ to randomly casting a vote, and the lemma follows.

First, any voter $\ell$ to the left of $m^-$ prefers delegating to $m^-$ to arbitrarily casting a vote. To this voter $\ell$, randomly casting a vote (regardless of the probability assigned to $p_0, p_1$) is dominated by delegating to the voter with peak 0, which is further dominated by delegating to $m^-$. Compared to voting arbitrarily, $\ell's$ expected loss from $\frac{p_0+p_1}{2} \in [b_\ell, 0]$ is outweighed by the expected gain from $\frac{p_0+p_1}{2} \in [0, -b_\ell]$. By monotonicity, delegating to $m^-$, someone with closer peak, increases $\ell$'s utility compared to delegating to 0. Thus to $\ell$, $m^- \succ 0 \succ$ vote randomly. Similarly, any voter to the right of $m^+$ prefers delegating to $m^+$ to voting arbitrarily.

Next, for any voter $i$ between $m^-$ and $m^+$, randomly casting a vote is strictly dominated by randomly delegating to $m^-$ and $m^+$ with probability $1/2$. To see this, in the incidences where $m^-$ and $m^+$ disagree, $i$ gets the same expected utility as random voting. In the incidences where $m^-$ and $m^+$ agree, $i$ in fact would also agree (by convexity), and gets strictly higher expected utility than random voting. Finally, for general voters, mixing between delegation to $m^-$ and $m^+$ is dominated by delegating to the one yielding higher expected utility. Therefore, any voter $i \in [m^-, m^+]$ prefers delegating to one of $m^-, m^+$ to random voting.

Therefore, any voter in $\mathcal{N}$ prefers delegating to one of $m^-, m^+$ to random voting. When $N$ is odd, at least one of $m^-$ and $m^+$ is preferred by $> N/2$ voters to randomly casting a vote. □

Now, we are ready to prove Theorem 16.

*Proof of Theorem 16.* WLOG, let $m^*$ be $m^-$. In the following, we first check that $m^-$ learns and votes truthfully, and all other voters delegate to her is cpNE.

First, we only need to consider coalitions with $> N/2$ members. Note that by the same token as the previous proofs, the coalition must be entirely on one side of $m^-$. There are fewer than $N/2$ voters to the left of $m^-$, so assume towards contradiction that a coalition $C$ exists to the right of $m^-$ with a self-enforcing deviation $s_C$. There exists at least one voter who learns in this deviation $s_C$. Otherwise, the coalition votes (individually or collectively) randomly, contradicting the the fact that $> N/2$ voters prefer delegating to $v^-$ to random voting.

In a self-enforcing deviating strategy, any voter who learns must vote truthfully (otherwise, either he forms a singleton deviating sub-coalition, or his vote has zero influence for all incidences of $p_0, p_1$ thus learning is not incentive compatible). Denote the set of voters who learn and vote in $C$ by $L(C)$. By individual rationality, the voter(s) in $L(C)$ have peaks equal to or to the right of $v^+$. However, over $N/2$ voters prefer $v^-$ to $v^+$, thus also prefer $v^-$ to the the weighted majority rule induced by $L(C)$. Coalition $C$ does not contain this set of voters, so is of size $< N/2$. Contradiction.

In fact, since we did not restrict the coalition's size to be $< N$, we have already

checked that it is efficient within all self-enforcing agreements (equivalent to it not having a self-enforcing deviation of size $N$). □

**Remark 8.** *Note that although the strategy profile $\{m^-$ learns, voters who prefer random voting votes randomly$\}$ is a Nash equilibrium, and results in the same voting outcome as Theorem 16, it is in fact not a cpNE. To see this, some $k$ voters in the right-end of the spectrum vote collectively/individually randomly while the rest delegates to a learner $m^-$. This effectively truncates the size of the voting population to $N-k$. Among these $N-k$ voters (who either vote or delegate), there may exist a voter $m'$ to the left of $m^-$ who finds it beneficial to deviate to learn and vote. Then all voters to the left of $m'$ will deviate to delegating to $m'$, making the overall voting outcome more left-leaning than the cpNE in Theorem 16.*

As a result, interestingly, although extreme right-wing voters may not like $m^-$'s decision rule, they may find it necessary to delegate to $m^-$ in order to rule out a more left-leaning coalition.

**Remark 9.** *Unlike the equilibrium in Theorem 14, the equilibrium in Theorem 16 is in fact not a Strong Nash equilibrium. This is because there exists non-self-enforcing deviation from this equilibrium. Suppose $m^* = m^-$, consider the following coalition: find a voter $r$ (to the right of $m^+$) with bias $b_r$ large enough such that $\mathbb{E}u_r(f_m) > \mathbb{E}u_r(f_{m^-})$, ane let her claim to vote according to median $m$'s utility. Then all voters to the right of $m$, including $m$ form a coalition that everyone strictly benefit from deviating. Since the equilibrium in Theorem 16 does not protect against this coalition, it is not Strong. However, notice that as we would expect, this deviation is not self-enforcing. Voter $r$ herself forms a deviating singleton sub-coalition whose utility increase from further deviating to vote according to her own bias.*

### 4.4.4 High learning cost

When learning cost is high so that no voter satisfies IR, the only pure-strategy NE is everyone voting arbitrarily. In fact, depending on the form of the prior distribution $\mathbb{P}$, learning cost does not need to be so high that no voter satisfies IR. There exist

cases where one side of the political spectrum contains incentivised learners yet no one learns forms the only NE. This is due to the opposite-wing voters forming a randomly voting coalition. We will show it specifically through construction. These results are encapsulated in Theorem 17.

**Theorem 17.** *Let $c^*$ be the minimum cost such that $\mathcal{I}_i \leq c^*$, for all $i$. There exists constant $c_h \leq c^*$ such that for any learning cost $c \geq c_h$, in any NE (thus cpNE), every voter votes arbitrarily (e.g., chooses $p_0$).*

*Proof of Theorem 17.* For general cases, $c_h = c^*$ suffices. No voter satisfies IR, thus the only pure-strategy NE is everyone voting arbitrarily.

Next, we show that there exists prior $\mathbb{P}$ and $\mathcal{N}$ such that $c_h < c^*$, i.e., no voters learn despite some satisfy IR. By Theorem 16, this necessarily means that all incentivised voters are to the same side of 0. We show existence by construction.

It suffices to show that there exists $\ell, r \in \mathbb{R}, |\ell| < |r|$ such that voter with bias $r$ prefers random (arbitrary voting) to delegating to $\ell$.

Take any integrable and square-integrable distribution $\mathbb{P}$ with cdf $F$. Fix any $r > 0$. Consider the decision rule $f^*$ that $\min\{u_r(p_0), u_r(p_1)\}$ when $\frac{p_0+p_1}{2} \leq r$ and $\max\{u_r(p_0), u_r(p_1)\}$ when $\frac{p_0+p_1}{2} > r$. $f^*$ is dominated by random (arbitrary) voting. Denote $r$'s expected utility induced by this $f^*$ as $u_r(f^*)$, and $r$'s expected utility from delegating to $\ell$ as $u_r(\ell)$.

Observe that two voters with biases $\ell$ and $r$ agree when $\frac{p_0+p_1}{2} \in [\ell, r]$, and disagree otherwise. Therefore, delegating to $\ell$ is same as choosing $\operatorname{argmin} u_r(p_0, p_1)$ when $\ell \leq \frac{p_0+p_1}{2} \leq r$, and choosing $\operatorname{argmax} u_r(p_0, p_1)$ when $\frac{p_0+p_1}{2} > r$ or $\frac{p_0+p_1}{2} < \ell$. Therefore, the sequence of $u_\ell$ as $\ell \to -\infty$ forms a decreasing sequence with lower bound $u_r(f^*)$. By monotone convergence theorem, $\{u_\ell\}$ converges. In fact, the limit is necessarily $u_r(f^*)$ since the decision rule differs from $f^*$ by a decreasing amount $F(\ell)$. $\{u_r(\ell)\}$ converges to $u_r(f^*) < u_{rand}$. As a result, there exists $\ell$ such that $u_\ell < u_{rand}$. In words, for any $r > 0$, there exists some $\ell < 0$ such that a voter with peak $r$ prefers random voting to delegating to any voter with a peak $\leq \ell$.

Now, fix $r > 0$, find $\ell$ such that $u_\ell < u_{rand}$. If $|r| > |\ell|$, let $r' = |\ell| + \epsilon$, for some small

$\epsilon > 0$, otherwise, let $r' = r$. Note that $r'$ also prefers random voting to $\ell$; otherwise, $r$ prefers $\ell$'s decision to random voting by convexity. Construct the following population and learning cost $c$. Let $\mathcal{N} = L \cup R$ where $L = \{N/2 - 1$ voters with peaks equal $\ell\}$ and $R = \{N/2 + 1$ voters with peaks equal $r'\}$, and let $c \in (\mathcal{I}_\ell, \mathcal{I}_{r'})$ so that voters in $L$ satisfy IR, while the voters in $R$ do not. Voters in $R$ prefer arbitrary voting to $L$'s decision rule. Therefore, everyone in $R$ votes arbitrarily is a dominant strategy, and $L$ best respond by voting arbitrarily. The only Nash equilibrium is everyone voting arbitrarily.

□

## 4.5 Discussion and future directions

An important general remark for the results in this paper is that though they describe the final result of the delegation (the final effective voters), any delegation graph that leads to the delegation outcome suffices as a cpNE. For example, local delegation that lead to median and everyone directly delegating to median both map to the same cpNE, and thus are equivalent as far as the theorems presented here concern.

Therefore, further equilibrium refinements seems both feasible and reflecting reality more closely. For example, trembling-hand equilibrium may be a suitable solution concept for justifying chains of local delegation that lead to the median over everyone directly delegating to the median, as the chain may probabilistically terminate with a voter decides to learn before the chain reaches median.

Another important element of the game worth further exploration is information in this game. Suppose voters only have local information about others' preferences, it seems the delegation is still locally computable or approximatable, especially if the quantile of the voter are provided as auxiliary information to the nodes. This also connects with the rest of liquid democracy literature with models naturally residing on networks. For example, consider the following concrete formulation: given a graph $\mathcal{G} = (V, E)$, publicly known to all voters in the society. Each voter $v_i \in V$ is assigned a preference peak $\mu_i \in \mathbb{R}$, known to himself and all neighbors. A locally computable

voting strategy profile would be $s_i = f(\mathcal{G}, \mu_j)$ with all $j \in \mathcal{N}(v_i)$.

The fundamental forces at play in the current model are incentive and relative position in the population. In fact, in models with learning noise, it could also be realistic for the forces to play reversely: accumulating enough voting power may work as an incentive to motivate voters to pay higher learning cost for a higher accuracy. This is beyond the realm of the current paper, but may be modelled by imposing a concave function between learning accuracy and and learning effort, and analyze the equilibrium delegation and effort investment that arise. For example, let learning cost increase linearly with learning effort, and learning quality be concave in learning effort. Consider a two-step game: in the first step, agents form the delegation graph; in the second step, voters observe final voting power distribution, bids a learning investment and start voting. This formulation models liquid democracy as a voting process where both the voting weights and efforts are endogenously constructed.

# Chapter 5

# Conclusion

In this thesis, we studied three mathematical models for informational interactions in social contexts: corruption detection on networks, multi-dimensional opinion dynamics and elections that allow delegation. Some of these models provide insight on the structure of networks that allows good surveillance in face of corrupt agents. Some models further this theme by demonstrating mechanisms that allow good aggregation of information even when all agents have heterogeneous preferences and behave strategically.

In some of these models or in certain parametric regimes, undesirable social outcomes occur: polarization as an unintended outcome of biased assimilation (a natural tendency of cognition) and information outlet's natural attempt (and need) to persuade; polarization and abandoning of votes as an outcome of liquid democracy when learning costs are high. In others, desirable social outcomes are ensured by robust designs: on networks with good expanding properties, one can always identify at least one truthful node with only local reports; liquid democracy achieves outcome nearly consistent or even better than direct voting when cost is low to intermediate.

From the point of view of mathematics and its application, two general syntheses that arise for robustness are expander graphs and its applications in networks, and median and its applications in information aggregation and statistics. We were able to show the the strength of expander graphs' even for guaranteeing robustness for highly stylized and dynamic agent-based model. We were also able to extend the

the power of "median is robust" to a social choice problem with a game structure. Median's surprising reappearance in delegative voting, resembling the form of median voter theorem in economics, also sheds some light on seeing median voter theorem and natural strategic outcome of voting not only as a game theoretic outcome, but also a robust statistical procedure that aggregates data, which are voters' preferences, and a even possible robust protection against corruption in elections.

Another reappearing theme in this thesis is local computation. Liquid democracy, identity reconstruction with local reports, and opinion update are all social functions that can be locally computed. This is not only interesting from a social perspective because local computation is feasible and the status quo socially, but also from a computational perspective because local computation is efficient computationally.

To bear in mind are the limitations of theoretical deviations. Models start with behavior assumptions, the model itself may exhibits several variations, and the results also often specify various parameters. We look forward to experiments and small-scale applications of many of the mechanisms, models, and theoretical results that appeared in this thesis. Whether the empirical results confirm or contradict the theoretical predictions, we are hopeful that they will further shed light on both the understanding of human and social behaviors, and the design of computationally efficient and incentive-compatible mechanisms in an ever-expanding social world.

# Appendix A

# Supplementary Material for Chapter 2

## A.1 Omitted Results

We give an NP-hardness result for computing $\min_k S_G(k) + k$ exactly. Note that this is insufficient to say anything about corruption detection, as $\min_k S_G(k) + k$ only gives a 2-approximation to the critical number $m(G)$, but we include this observation here as it may be of independent interest.

**Theorem 18.** *It is NP hard to compute* $\min_k S_G(k) + k$ *exactly.*

*Proof.* It is known that finding $k$-vertex separator for a graph is NP hard [63]. We present a reduction of the problem of computing $\min_k S_G(k) + k$ to the $k$-vertex separator problem.

Assume towards contradiction that there is a polynomial-time algorithm $\mathcal{A}$ for finding $\min_k S_G(k) + k$. Then for any graph $G$ and any $M < |V|$, the minimal $M$-vertex separator of the graph $G = (V, E)$ can be found in the following way. Construct a graph $G' = (V', E')$, where

$$G' = G \cup \{n^2 \text{ disjoint M-cliques}\},$$

with $n \gg N := |V|$. Construct a second auxiliary graph $G'' = (V'', E'')$, such that

$$G'' = G' \cup \{kn + N \text{ disjoint } (n-1)\text{-cliques appended to each vertex of V'}\}.$$

Each $(n-1)$-clique is appended to a vertex of $G'$ in the sense that each node of the clique is connected to the vertex in $G'$ with an edge. The idea is to make each vertex in $G'$ "$n$ times larger".

Run the polynomial-time algorithm $\mathcal{A}$ for finding $\min_k S_{G''}(k) + k$ on graph $G''$. The algorithm outputs a vertex set $S'' \subseteq V''$, which divides $G''$ into connected components of with maximal size $k''$.

**Lemma 12.** *Let $G''$ be as constructed above, $k''$ and $S''$ be the output given by an algorithm that computes $\min_k S_{G''}(k) + k$. Then $k'' = nM$, and without loss of generality, the subset $S''$ contains only vertices from the original graph $G$. In other words, finding $\min_k S(k) + k$ of $G''$ is equivalent to finding the $M$-vertex separator of $G$. i.e.,*

$$\arg\min_k S_{G''}(k) + k = nM,$$

$$\min_k S_{G''}(k) + k = S_G(M) + nM.$$

*Proof of Lemma 12.* Let $f_{G''}(k) := S_{G''}(k) + k$, and let $f^*_{G''} := \min_k f_{G''}(k)$. Note there exists following upper bound for $f^*_{G''}$.

$$f^*_{G''} \leq S_G(M) + nM$$

This is achieved by removing the $M$-vertex separator of $G$ from $G''$ and divide $G''_{V'' \setminus S_G(M)}$ into connected components with size at most $nM$.

Now we prove that $f^*_{G''}$ has to be exactly $S_G(M) + nM$ by showing that $f_{G''}(k) > f^*_{G''}$ for $k > nM$, and for $k < nM$.

1. $f_{G''}(k) > f^*_{G''}$ for all $k < nM$.

For $k < nM$:

$$f_{G''}(k) \geq n^2 + k > S_G(M) + nM,$$

because the separator has to include at least one vertex from each of the $n^2$ disjoint $nM$-cliques in $G''$. This value $f_{G''}(k)$ is clearly larger than $S_G(M)+nM$ when $n \gg N > M$.

2. $f_{G''}(k) > f^*_{G''}$ for all $k > nM$.

**Claim 15.** *We claim that it suffices to only consider $k$ in the form of $k = nM + n\alpha$, where $\alpha \in \mathbb{Z}_+$. i.e. for any $k > nM$, $f_{G''}(k) \geq f_{G''}(nM + n\alpha)$ for some $\alpha \in \mathbb{Z}_+$.*

*Proof of Claim 15.* Call the nodes in $G$ to which each of the $n$-clique is appended to (while constructing $G''$) the **center** of the $n$-clique in $G''$. If $k$ cannot be expressed in the form of $nM + n\alpha$, this means the corresponding separator $S$ contain some non-center nodes of the $n$-cliques in $G''$.

If the center $\notin S$, while some other node(s) of the clique $\in S$, there exists another $S^*$, $|S^*| < |S|$ that includes the center instead of the other node(s), and suffice to be a $k$-vertex separator. This is because after the removal of the center node, the rest of the clique can be of size at most $(n-1)$, and $k > nM > n-1$.

Suppose the center $\in S$, while some of the other node(s) of the clique also $\in S$, in order to obtain a $k$-vertex separator. Then $S^*$ that only contains center will suffice to be $k$-vertex separator, because $k > n$. □

By Claim 15, for any $k > nM$, $f_{G''}(k) \geq f_{G''}(nM + n\alpha)$ for some $\alpha \in \mathbb{Z}_+$. In words, there is never any incentive to include any non-center nodes of an $n$-cliques in separator $S$. Without loss of generality, $S \subseteq V_G$, and $k = nM + n\alpha \geq nM + n$.

$$f_{G''}(nM + n\alpha) > nM + n > S_G(M) + nM$$

when $n \gg N$.

Summarizing 1 and 2, we conclude that

$$f^*_{G''} = f_{G''}(nM) = S_G(M) + nM$$

□

This gives us a polynomial time algorithm to find any $M$-vertex separator for any graph $G$, and any value $M$. This contradicts the fact that computing $M$-vertex separator is NP-hard. Therefore, there does not exist polynomial time algorithm for computing $\min S_G(k) + k$. □

# Appendix B

# Supplementary Material for Chapter 3

## B.1 Proof of Claim 7

Let us embed our underlying space $\mathbb{R}^2$ in $\mathbb{R}^3$ by setting the last coordinate to zero. Letting $\times$ denote the cross product, we have

$$u \times u' = (0, 0, \sin \alpha_t) \, , \qquad f(u, v) \times f(u', v) = (0, 0, \sin \widehat{\alpha}) \, .$$

Since the case $\alpha_t \in \{0, \pi\}$ is easily handled by noticing that $\widehat{\alpha} = \alpha_t$, we can assume that $0 < \alpha_t < \pi$. In that case, it is enough that we prove

$$\langle u \times u', f(u, v) \times f(u', v) \rangle = \sin \alpha_t \sin \widehat{\alpha} \geq 0 \, . \tag{B.1}$$

Setting $C(w) := \sqrt{1 + (2\eta + \eta^2)\langle w, v\rangle^2}$, we apply (3.2) and bilinearity of cross product to compute

$$\begin{aligned} f(u,v) \times f(u',v) &= \frac{1}{C(u)C(u')}\Big(u \times u' + \eta(\langle u, v\rangle(v \times u') + \langle u', v\rangle(u \times v))\Big) \\ &= \frac{1}{C(u)C(u')}\Big(u \times u' + \eta(u \times u' + (\langle u, v\rangle v - u) \times (u' - \langle u', v\rangle v))\Big) \qquad \text{(B.2)} \\ &= \frac{1+\eta}{C(u)C(u')} \cdot u \times u' \ , \qquad \text{(B.3)} \end{aligned}$$

where in (B.2) we used the identity $a \times b + c \times d = a \times d + c \times b + (a - c) \times (b - d)$, and in (B.3) we used that both $\langle u, v\rangle v - u$ and $u' - \langle u', v\rangle v$ are projections of vectors onto the line orthogonal to $v$, and therefore they are parallel and their cross product vanishes.

Consequently, we conclude that $f(u,v) \times f(u',v)$ is parallel to $u \times u'$ with a positive proportionality constant, which implies (B.1) and concludes the proof. □

## B.2 Example with two advertisers

For another slightly more involved example, suppose there are two advertisers marketing their products. Agents' opinions now have five dimensions ($d = 5$) with the fourth and fifth coordinates corresponding to the opinions on these two products. Initially, 500 opinions on the first three coordinates are distributed randomly and uniformly on a three-dimensional sphere, and the last two coordinates are equal to zero:

$$u_i = (u_{i,1}, u_{i,2}, u_{i,3}, 0, 0) \qquad \text{subject to} \qquad u_{i,1}^2 + u_{i,2}^2 + u_{i,3}^2 = 1 \ .$$

Suppose the two advertisers apply interventions $v_1$ and $v_2$ in an alternating fashion. We take $v_1$ and $v_2$ to be orthogonal, letting

$$v_1 = (\beta, 0, 0, \alpha, 0) \ , \qquad v_2 = (0, \beta, 0, 0, \alpha) \ , \qquad \alpha = \frac{3}{4}, \beta = \sqrt{1 - \alpha^2} \ .$$

We proceed to apply $v_1$ and $v_2$ in an alternating fashion. In Figure B-1 we illustrate the agents' opinions after each advertiser applied their intervention two, four and six times (so the total of, respectively, four, eight and twelve interventions have been applied). A pattern of polarization on the fourth and fifth coordinates can be observed. At the same time, the pattern on the first three coordinates is more complicated. The opinions on these dimensions are scattered around a circle on the plane spanned by the first two coordinates. This is a somewhat special behavior that arises because vectors $v_1$ and $v_2$ are orthogonal. It is connected to the difference between Theorem 12 and Proposition 4 discussed in Section 3.6.

## B.3 Proof of Proposition 2

Recall that the two-agent intervention maximizes $\min(\widetilde{u}_{1,d}, \widetilde{u}_{2,d})$. Due to symmetry, we will consider wlog the one-agent intervention that maximizes $\widetilde{u}_{1,d}$. Substituting into (3.2), we get that applying an intervention $v$ results in

$$\widetilde{u}_{i,d} = \frac{\langle u_i, v\rangle \cdot v_d}{\sqrt{1 + 3\langle u_i, v\rangle^2}} \,. \tag{B.4}$$

Recalling (3.4), we can apply any unitary transformation on the opinions without changing the correlations, and hence assume that

$$u_1 := (\sin\alpha, \cos\alpha, 0, \ldots, 0) \,, \qquad u_2 := (-\sin\alpha, \cos\alpha, 0, \ldots, 0) \tag{B.5}$$

for some $0 \le \alpha \le \pi/2$ and accordingly, $c = \cos^2\alpha - \sin^2\alpha = \cos(2\alpha)$. In particular, $\alpha = 0$ means that the agents are in full agreement, $\alpha = \pi/4$ corresponds to the case of orthogonal opinions and $\alpha = \pi/2$ is the case where the opinions are antipodal.

Assuming (B.5), once we fix the first two coordinates of the intervention $v_1$ and $v_2$, also the values of $\langle u_1, v\rangle$ and $\langle u_2, v\rangle$ become fixed. Therefore, due to (B.4), the values of $\widetilde{u}_{i,d}$ depend only on $v_d$ in a linear fashion. Accordingly, the influencer should place as much weight as possible on the last coordinate and we can conclude that both two- and one-agent interventions have $v_j = 0$ for $2 < j < d$. Hence, in the following we

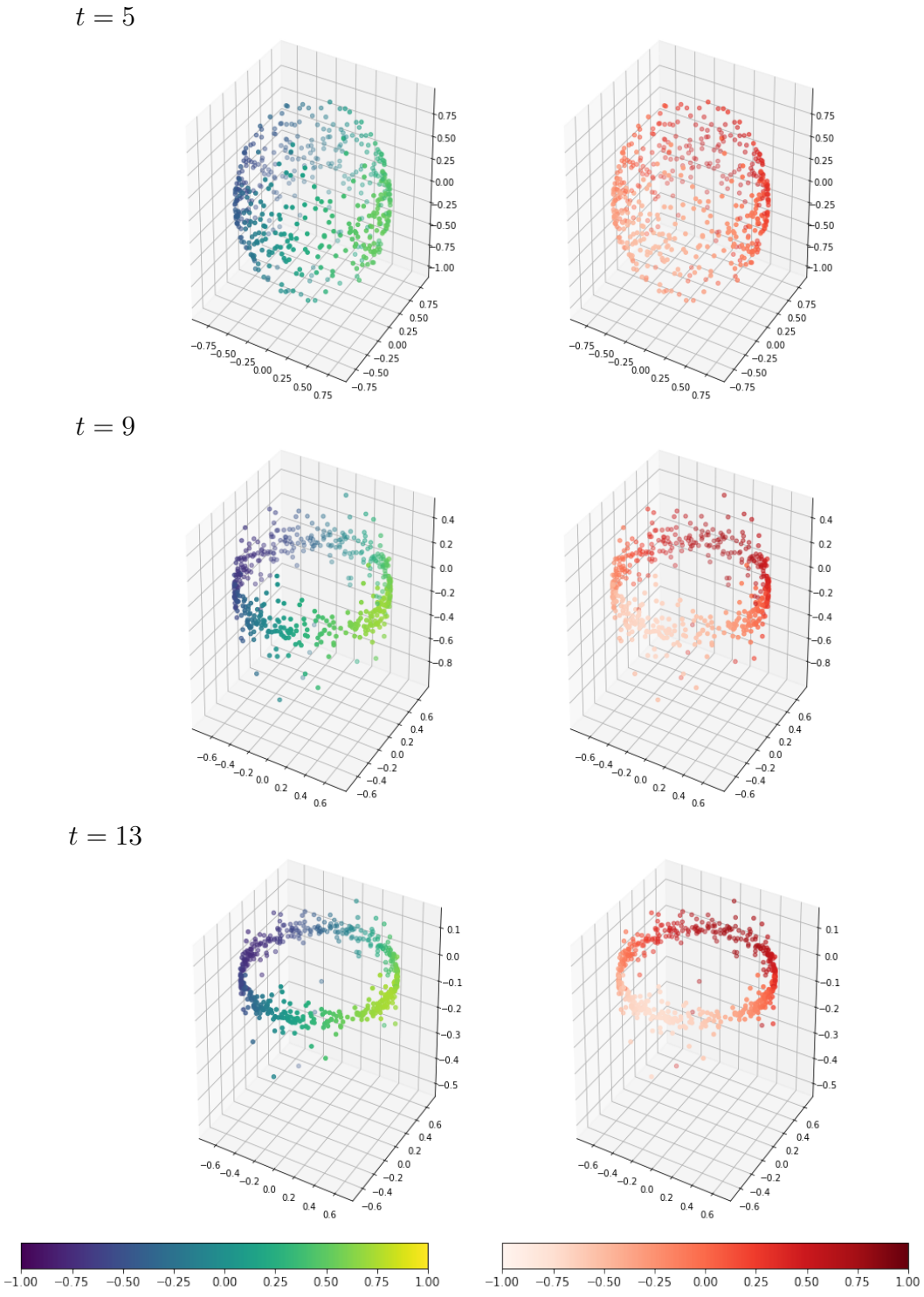


Figure B-1: Illustration of the process described in Appendix B.2. This time we need to visualize five dimensions. This is done with spatial positions for the first three dimensions $j = 1, 2, 3$ and two different color scales for $j = 4, 5$. Accordingly, two figures are displayed for each time step $t = 5, 9, 13$. In each pair of figures the points in the left figure have the same spatial positions as in the right figure and the colors illustrate dimensions $j = 4$ (on the left) and $j = 5$ (on the right).

will assume wlog that $d = 3$, $u_1 = (\sin\alpha, \cos\alpha, 0)$ and $u_2 = (-\sin\alpha, \cos\alpha, 0)$ (see Figure B-2).

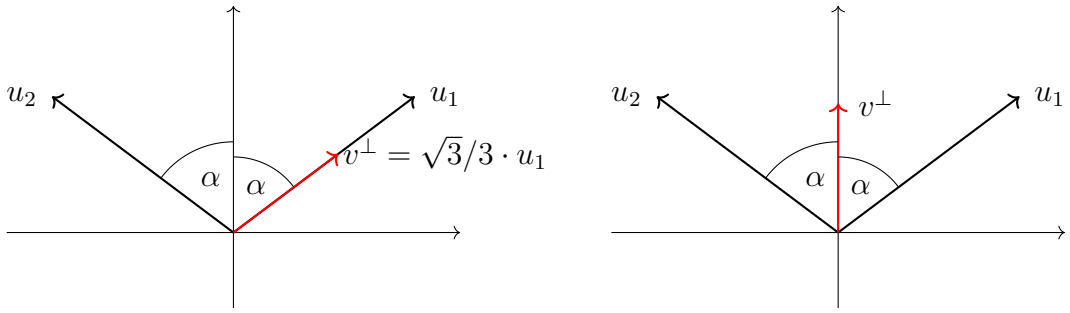


Figure B-2: The projection of one-agent (left) and two-agent (right) interventions onto the first two dimensions.

First, consider the one-agent intervention maximizing $\widetilde{u}_{1,3}$. Clearly, the intervention should be of the form

$$v_{\text{one}} = \cos\beta \cdot u_1 + \sin\beta \cdot (0, 0, 1)$$

for some $0 \le \beta \le \pi/2$. Substituting in (B.4), we compute

$$(\widetilde{u}_{1,3})^2 = \frac{\cos^2\beta \sin^2\beta}{1 + 3\cos^2\beta} \,. \tag{B.6}$$

Maximizing (B.6), we get the maximum at $\cos\beta = \sqrt{3}/3$ and

$$v_{\text{one}} = \frac{\sqrt{3}}{3} \cdot u_1 + \frac{\sqrt{6}}{3} \cdot (0, 0, 1) \,,$$

resulting in $\widetilde{u}_{1,3} = 1/3$. The value $1/3$ is the benchmark for what can be achieved by one intervention. It is a maximum value for $\widetilde{u}_{1,3}$ attainable provided that initially $u_{1,3} = 0$.

What is the effect of this intervention on the other opinion $u_2$? Since $\langle u_2, v_{\text{one}} \rangle = \sqrt{3}c/3$, substituting into (B.4) we get

$$\widetilde{u}_{2,3} = \frac{\sqrt{3}c/3 \cdot \sqrt{6}/3}{\sqrt{1 + c^2}} = \frac{c\sqrt{2}}{3\sqrt{1 + c^2}} \,.$$

The value of $\widetilde{u}_{2,3}$ as a function of the correlation $c \in [-1, 1]$ is shown in blue in Figure 3-4. In particular, it increases from $-1/3$ to $1/3$, passing through 0 for $c = 0$.

Moving to the two-agent intervention, in this case it is not difficult to see (cf. Figure B-2) that the intervention vector should be of the form

$$v_{\text{two}} = (0, \cos\beta, \sin\beta)$$

for some $0 \leq \beta \leq \pi/2$. A computation in a computer algebra system (CAS) establishes that $\widetilde{u}_{1,3} = \widetilde{u}_{2,3}$ is maximized for

$$\cos^2\beta = \frac{\sqrt{2}(\sqrt{3c+5} - \sqrt{2})}{3(c+1)} \,,$$

yielding an expression

$$\widetilde{u}_{1,3} = \widetilde{u}_{2,3} = \sqrt{\frac{3c + 7 - 2\sqrt{6c+10}}{9(c+1)}} \,.$$

This function is depicted in Figure 3-4 in red. In particular, for $c \in [-1, 1]$, it increases from 0 to $1/3$ and its value at $c = 0$ is approximately 0.27. Furthermore, its growth close to $c = -1$ is of the square-root type.

Turning to the new correlation values $c_{\text{one}}$ and $c_{\text{two}}$, another CAS computation using the formulas for $v_{\text{one}}$ and $v_{\text{two}}$ gives

$$c_{\text{one}} = \frac{c\sqrt{2}}{\sqrt{c^2+1}} \,, \qquad\qquad c_{\text{two}} = 1 - \frac{\sqrt{2}(1-c)}{\sqrt{3c+5}} \,,$$

establishing (3.10). To conclude the proof we need another elementary calculation showing that $c_{\text{two}} \geq c_{\text{one}}$ always holds. We omit the details, referring to Figure 3-3 and noting that in the critical region for $c = 1 - \varepsilon$ we have

$$c_{\text{two}} = 1 - \frac{1}{2}\varepsilon - \frac{3}{32}\varepsilon^2 + O(\varepsilon^3) \geq c_{\text{one}} = 1 - \frac{1}{2}\varepsilon - \frac{3}{8}\varepsilon^2 + O(\varepsilon^3) \,. \quad \square$$

## B.4 Proof of Proposition 3

Let us write a generic intervention vector as

$$v = (\cos\beta \cdot v^*, \sin\beta) \ ,$$

where $0 \leq \beta \leq \pi/2, v^* \in \mathbb{R}^{d-1}$ and $\|v^*\| = 1$. If $v$ is applied to an opinion vector $u_i = (u_i^*, 0)$ and we let $c_i := \langle u_i^*, v^* \rangle$, substituting into (3.2) we can compute

$$u_i + \langle u_i, v \rangle \cdot v = (u_i^*, 0) + c_i \cos\beta(\cos\beta \cdot v^*, \sin\beta) = (u_i^* + c_i \cos^2\beta \cdot v^*, c_i \cos\beta \sin\beta) \ ,$$

and therefore, using (3.3),

$$\widetilde{u}_{i,d} = \frac{c_i \cos\beta \sin\beta}{\sqrt{1 + 3c_i^2 \cos^2\beta}} = \frac{c_i z\sqrt{1-z^2}}{\sqrt{1+3c_i^2 z^2}} \ , \tag{B.7}$$

where we let $z := \cos\beta$.

Consider a fixed unit vector $v^* \in \mathbb{R}^{d-1}$. In order to maximize $\widetilde{u}_{i,d}$ for an opinion $u_i$ with $\langle u_i^*, v^* \rangle = c_i$, we need to optimize over $z$ in (B.7), resulting in $z = \sqrt{\sqrt{1+3c_i^2} - 1}/(\sqrt{3}c_i)$ and, substituting,

$$\widetilde{u}_{i,d} = \frac{\sqrt{1+3c_i^2} - 1}{3c_i} \ . \tag{B.8}$$

The right-hand side of (B.7) is easily seen to be increasing in $c_i > 0$ for a fixed $z$. Therefore, in order to maximize the number of points with $\widetilde{u}_{i,d} > T$ for a fixed $v^*$, we solve the equation $T = \frac{\sqrt{1+3c^2}-1}{3c}$ for $c$, resulting in $c = \frac{2T}{1-3T^2}$ and apply the intervention

$$v = (\cos\beta \cdot v^*, \sin\beta) \ ,$$

just as claimed in (3.11). This intervention results in $\widetilde{u}_{i,d} > T$ for all opinions satisfying $\langle u_i^*, v^* \rangle > c$. In other words, the objective $\widetilde{u}_{i,d} > T$ is achieved for exactly those opinions contained in the spherical cap $\{x \in \mathbb{R}^{d-1} : \langle x, v^* \rangle > c\}$. Maximizing over all directions $v^* \in \mathbb{R}^{d-1}$ completes the proof. $\square$

# Bibliography


[1] Noga Alon, Elchanan Mossel, and Robin Pemantle. Corruption detection on networks. *CoRR*, abs/1505.05637, 2015.

[2] Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1–2):237–260, 1998.

[3] Robert Aumann. *Acceptable points in general cooperative n-person games, in "Contributions to the Theory of Games IV*. Princeton University Press, Princeton, N.J., 1959.

[4] Per Austrin, Toniann Pitassi, and Yu Wu. Inapproximability of treewidth, one-shot pebbling, and related layout problems. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 13–24, 2012.

[5] Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, 41(2):203–226, 1997.

[6] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[7] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.

[8] Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *The Review of Economic Studies*, 65(3):595–621, 1998.

[9] Delia Baldassarri. *Crosscutting Social Spheres? Political Polarization and the Social Roots of Pluralism*. PhD thesis, Columbia University, 2007.

[10] Delia Baldassarri and Peter Bearman. Dynamics of political polarization. *American Sociological Review*, 72(5):784–811, 2007.

[11] Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.

[12] BBC. Trump signs executive order targeting Twitter after fact-checking row, 2020. 29 May 2020. `https://www.bbc.com/news/technology-52843986`.

[13] Jan Behrens, Axel Kistner, and Andreas Nitsche. *The Principles of LiquidFeedback*. Interaktive Demokratie e. V., 2014.

[14] Walid Ben-Ameur, Mohamed-Ahmed Mohamed-Sidi, and José Neto. The k - separator problem: polyhedra, complexity and approximation results. *J. Comb. Optim.*, 29(1):276–307, 2015.

[15] Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

[16] Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. The computational complexity of densest region detection. *Journal of Computer and System Sciences*, 64(1):22–47, 2002.

[17] Shai Ben-David and Hans-Ulrich Simon. Efficient learning of linear perceptrons. In *Advances in Neural Information Processing Systems (NIPS)*, pages 189–195, 2000.

[18] B.Douglas Bernheim, Bezalel Peleg, and Michael D Whinston. Coalition-proof nash equilibria i. concepts. *Journal of Economic Theory*, 42(1):1 – 12, 1987.

[19] Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34, 1948.

[20] Jarosław Błasiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan. General strong polarization. In *Symposium on Theory of Computing (STOC)*, pages 485–492, 2018.

[21] Daan Bloembergen, Davide Grossi, and Martin Lackner. On rational delegations in liquid democracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1796–1803, Jul. 2019.

[22] Christian Blum and Christina Isabel Zuber. Liquid democracy: Potentials, problems, and perspectives. *Journal of Political Philosophy*, 24(2):162–182, 2016.

[23] Hans L. Bodlaender, John R. Gilbert, Hjálmtyr Hafsteinsson, and Ton Kloks. Approximating treewidth, pathwidth, frontsize, and shortest elimination tree. *J. Algorithms*, 18(2):238–255, 1995.

[24] Nader H. Bshouty and Lynn Burroughs. Maximizing agreements and coagnostic learning. *Theoretical Computer Science*, 350(1):24–39, 2006.

[25] Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on Twitter. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 89–96, 2011.

[26] Anton T. Dahbura and Gerald M. Masson. An $O(n^{2.5})$ fault identification algorithm for diagnosable systems. *IEEE Trans. Computers*, 33(6):486–492, 1984.

[27] Pranav Dandekar, Ashish Goel, and David T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.

[28] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. Modeling confirmation bias and polarization. *Scientific Reports*, 7:40391, 2017.

[29] Valerio Dotti. *Multidimensional voting models: theory and applications*. PhD thesis, London, England, 2016.

[30] Anthony Downs. An economic theory of political action in a democracy. *Journal of Political Economy*, 65(2):135–150, Apr 1957.

[31] Timothy Feddersen and Alvaro Sandroni. A theory of participation in elections. *American Economic Review*, 96(4):1271–1282, Aug 2006.

[32] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Symposium on Foundations of Computer Science (FOCS)*, pages 563–574, 2006.

[33] Morris P. Fiorina and Samuel J. Abrams. Political polarization in the American public. *Annual Reviev of Political Science*, 11:563–588, 2008.

[34] Morris P. Fiorina, Samuel J. Abrams, and Jeremy C. Pope. *Culture War? The Myth of a Polarized America*. Pearson-Longman, 2005.

[35] Odd-Helge Fjeldstad. Fighting fiscal corruption: lessons from the tanzania revenue authority. *Public Administration and Development: The International Journal of Management Research and Practice*, 23(2):165–175, 2003.

[36] Bryan Ford. A liquid perspective on democratic choice. March 2020.

[37] Jacey Fortin. A list of the companies cutting ties with the N.R.A., 2018. The New York Times website, 24 February 2018. `https://www.nytimes.com/2018/02/24/business/nra-companies-boycott.html`.

[38] Jason Gaitonde, Jon Kleinberg, and Éva Tardos. Polarization in geometric opinion dynamics. To appear in ACM Conference on Economics and Computation (EC), 2021.

[39] Kiran Garimella. *Polarization on Social Media*. PhD thesis, Aalto University, 2018. 20/2018.

[40] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. Balancing information exposure in social networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4663–4671, 2017.

[41] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *International Conference on Web Search and Data Mining (WSDM)*, pages 81–90. ACM, 2017.

[42] Benny Geys. 'Rational' theories of voter turnout: A review. *Political Studies Review*, 4(1):16–35, 2006.

[43] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[44] Paul GÖlz, Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. The fluid mechanics of liquid democracy. *ACM Transactions on Economics and Computation*, 9(4):1–39, Oct 2021.

[45] S. Louis Hakimi and A. T. Amin. Characterization of connection assignment of diagnosable systems. *IEEE Trans. Computers*, 23(1):86–88, 1974.

[46] Daniel Halpern, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, and Manon Revel. In defense of liquid democracy, 2021.

[47] Steve Hardt and Lia C. R. Lopes. *Google Votes: A Liquid Democracy Experiment on a Corporate Social Network*. Jun 2015.

[48] Jan Hązła, Yan Jin, Elchanan Mossel, and Govind Ramnarayan. A geometric model of opinion polarization. *arXiv preprint arXiv:1910.05274*, 2019.

[49] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence. models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[50] Harold Hotelling. Stability in competition. *The Economic Journal*, 39(153):41–57, 1929.

[51] Shanto Iyengar and Sean J. Westwood. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707, 2015.

[52] Ali Jadbabaie, Elchanan Mossel, and Mohammed Amin Rahimian. Bayesian group decisions: Algorithms and complexity. arXiv:1705.04770, 2017.

[53] Ali Jadbabaie, Nader Motee, and Mauricio Barahona. On the stability of the Kuramoto model of coupled nonlinear oscillators. In *American Control Conference (ACC)*, volume 5, pages 4296–4301, 2004.

[54] Anson Kahng, Simon Mackenzie, and Ariel Procaccia. Liquid democracy: An algorithmic perspective. *J. Artif. Int. Res.*, 70:1223–1252, may 2021.

[55] Tiko Kameda, S Toida, and FJ Allan. A diagnosing algorithm for networks. *Information and Control*, 29(2):141–148, 1975.

[56] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. *Theory of Computing*, 11:105–147, 2015.

[57] Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the 17th Annual IEEE Conference on Computational Complexity, Montréal, Québec, Canada, May 21-24, 2002*, page 25, 2002.

[58] Christoph Kling, Jérôme Kunegis, Heinrich Hartmann, Markus Strohmaier, and Steffen Staab. Voting behaviour and power in online democracy: A study of liquidfeedback in germany's pirate party. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):208–217, Aug 2021.

[59] Stefan Krasa and Mattias K. Polborn. Political competition in legislative elections. *American Political Science Review*, 112(4):809–825, 2018.

[60] Jon G Kuhl and Sudhakar M Reddy. Distributed fault-tolerance for large multiprocessor systems. In *Proceedings of the 7th annual symposium on Computer Architecture*, pages 23–30. ACM, 1980.

[61] Ilyana Kuziemko and Ebonya Washington. Why did the democrats lose the south? Bringing new data to an old debate. *American Economic Review*, 108(10):2830–67, 2018.

[62] Howard Lavine, Eugene Borgida, and John L. Sullivan. On the relationship between attitude involvement and attitude accessibility: Toward a cognitive-motivational model of political information processing. *Political Psychology*, 21(1):81–106, 2000.

[63] Euiwoong Lee. Partitioning a graph into small pieces with applications to path transversal. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1546–1558, 2017.

[64] Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979.

[65] Michael W. Macy, James A. Kitts, Andreas Flache, and Steve Benard. Polarization in dynamic networks: A Hopfield model of emergent structure. *Dynamic Social Network Modeling and Analysis*, pages 162–173, 2003.

[66] Shachindra N. Maheshwari and S. Louis Hakimi. On models for diagnosable systems and probabilistic fault diagnosis. *IEEE Trans. Computers*, 25(3):228–236, 1976.

[67] James C. Miller. A program for direct and proxy voting in the legislative process. *Public Choice*, 7-7(1):107–113, 1969.

[68] Elchanan Mossel and Grant Schoenebeck. Reaching consensus on social networks. In *Innovations in Computer Science (ITCS)*, 2010.

[69] Thebeth Rufaro Mukwembi and Simon Mukwembi. Corruption and its detection: a graph-theoretic approach. *Computational and Mathematical Organization Theory*, 23(2):293–300, Jun 2017.

[70] Sendhil Mullainathan and Andrei Shleifer. The market for news. *American Economic Review*, 95(4):1031–1053, 2005.

[71] Richard P Nielsen. Corruption networks and implications for ethical corruption reform. *Journal of Business ethics*, 42(2):125–149, 2003.

[72] Mark Noah. Beyond individual differences: Social differentiation from first principles. *American Sociological Review*, 63(3):309, 1998.

[73] Andrzej Nowak, Jacek Szamrej, and Bibb Latané. From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, 97(3):362, 1990.

[74] Maarten Oosten, Jeroen HGC Rutten, and Frits CR Spieksma. Disconnecting graphs by removing vertices: a polyhedral approach. *Statistica Neerlandica*, 61(1):35–60, 2007.

[75] Eli Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, New York, 2011.

[76] Sergey E. Parsegov, Anton V. Proskurnikov, Roberto Tempo, and Noah E. Friedkin. Novel multidimensional models of opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285, May 2017.

[77] Alois Paulin. An overview of ten years of liquid democracy research. *The 21st Annual International Conference on Digital Government Research*, 2020.

[78] Jacopo Perego and Sevgi Yuksel. Media competition and social disagreement, 2018. Working Paper.

[79] Thomas F. Pettigrew and Linda R. Tropp. A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90(5):751–783, 2006.

[80] Pew Research Center. Political polarization in the American public: How increasing ideological uniformity and partisan antipathy affect politics, compromise and everyday life, 2014. `https://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/`.

[81] Franco P. Preparata, Gernot Metze, and Robert T. Chien. On the connection assignment problem of diagnosable systems. *IEEE Trans. Electronic Computers*, 16(6):848–854, 1967.

[82] Markus Prior. Media and political polarization. *Annual Review of Political Science*, 16:101–127, 2013.

[83] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 755–764, 2010.

[84] Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *Proceedings of the 27th Conference on Computational Complexity, CCC 2012, Porto, Portugal, June 26-29, 2012*, pages 64–73, 2012.

[85] Manon Revel, Daniel Halpern, Adam Berinsky, and Ali Jadbabaie. Liquid democracy in practice: An empirical analysis of its epistemic performance. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

[86] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. On the inevitability of online echo chambers. arXiv:1905.03919, 2019.

[87] Elizabeth A. Saylor, Katherine A. Vittes, and Susan B. Sorenson. Firearm advertising: Product depiction in consumer gun magazines. *Evaluation Review*, 28(5):420–433, 2004.

[88] Arthur Schram and Joep Sonnemans. Voter turnout as a participation game: An experimental investigation. *International Journal of Game Theory*, 25(3):385–406, Sep 1996.

[89] Guodong Shi, Alexandre Proutiere, Mikael Johansson, John S. Baras, and Karl H. Johansson. The evolution of beliefs over signed social networks. *Operations Research*, 64(3):585–604, 2016.

[90] John Sides and Daniel J. Hopkins. *Political polarization in American politics*. Bloomsbury Publishing USA, 2015.

[91] Brendan Snyder. LGBT advertising: How brands are taking a stance on issues. Think with Google, 2015.

[92] Gregory F. Sullivan. A polynomial time algorithm for fault diagnosability. In *FOCS*, pages 148–156. IEEE Computer Society, 1984.

[93] Donald E. Vinson, Jerome E. Scott, and Lawrence M. Lamont. The role of personal values in marketing and consumer behavior. *Journal of Marketing*, 41(2):44–50, 1977.

[94] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[95] Hywel T.P. Williams, James R. McMurray, Tim Kurz, and F. Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, 2015.

[96] Jie Xu and Shi-ze Huang. Sequentially t-diagnosable systems: A characterization and its applications. *IEEE Trans. Computers*, 44(2):340–345, 1995.

[97] John R. Zaller. *The Nature and Origins of Mass Opinion*. Cambridge University Press, 1992.

[98] Yuzhe Zhang and Davide Grossi. Power in liquid democracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5822–5830, May 2021.