Reasoning Together: Network Methods
for Political Talk and Normative Reasoning

by Sarah Shugars

B.A. in Physics, Clark University
M.A. in Integrated Marketing Communications, Emerson College

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

April 10, 2020

Dissertation directed by

Nicholas Beauchamp
Assistant Professor of Political Science

David Lazer
University Distinguished Professor of Political Science and Computer and
Information Science
Dedication

For those who were told that they couldn’t.
And for all those who believed that was true.
Almost exactly twenty years ago, a good friend of mine died. He wasn’t the first friend I lost, and he, unfortunately, was also not the last. But I remember learning of his death so perfectly clearly. I remember that awful sinking feeling, the Hitchcock dolly zoom of the world retreating into faded obscurity under the harsh light of a stark reality. Nothing would ever be the same.

My friend had been missing for a few days. He’d gotten himself into a bit of trouble and, I expected, was laying low until things blew over. That is how these things went, after all. One day’s drama was subsumed by the next day’s laugh track, a real-life sitcom in which even the worst truths could be erased through a smile and a shrug. No matter how bad things seemed, it would always work out in the end.

In retrospect, my optimism for this outcome seems foolish even then. At the time, I was a quiet and lonely sixteen-year-old who hadn’t quite put the “post” in post-traumatic stress disorder. I regularly skipped or carefully portioned meals because I couldn’t always afford to eat. I had witnessed violence up close and spent my nights clinging to consciousness, too terrified to sleep in my own home. Despite dissociation, depression, and a tendency towards bipolar disorder, as the youngest member of my household, I was also the most financially and emotionally stable. I had long since learned how to keep up appearances despite a life that, as it turns out, was quite simply not normal.

So, you would think, perhaps, that I would have known better than to believe it could all be okay. You would think that I would have been prepared for the worst. Expected it, even. Yet on that warm spring afternoon, as I blithely danced in the rain like the California hippy that I am, I was not at all prepared for the news. I was not at all prepared for my mother to tell me that they had found my friend. That they had found his body.

The world slipped away.
Just a few short weeks later was my high school graduation. Speaker after speaker stood up to share platitudes about how we were the future and how we could do anything. I was so, so angry. I still am angry, to be honest. My friend had graduated from that same high school just two years before. Just two years before he died. And these distant adults had the gall to get up there and tell me that our futures were bright? Kids like me didn’t get to have a future. We just didn’t. I knew it. They knew it. Everybody knew it. These adults were lying and thought I was too naïve to figure it out.

In retrospect, though, I think they were actually just lying to themselves.

That is the great deceit of the American dream, after all. The story that any kid can grow up to be President; that any child can become anything they want. That if you work hard, if you perfect your craft, if you persevere through some character-building setbacks, if you are a Good person who deserves Good outcomes, then you too can have it all.

What an unforgivable lie.

I write this in 2020. It feels like a short lifetime ago that Senator Elizabeth Warren suspended her campaign to become President of the United States. It has been less than two weeks. Just two weeks since a Super Tuesday showing which could politely be referred to as “disappointing.” An election in which she took third place in our mutually adopted home state of Massachusetts.

The closing of Senator Warren’s campaign was deeply painful for me. Not only because she was my candidate – though she was – but because the misogyny which plagued her candidacy was unavoidable, intractable. Any kid can grow up to be President, sure, but only if that kid isn’t a woman who tries to run for President. Kids like me don’t get to be President. We just don’t.

The anger and grief I felt in response to the end of Senator Warren’s campaign is not unlike that which I felt during my high school graduation all those years ago. They
tell us we can do anything, be anything. That if we shoot for the moon, we will land among the stars. They fill us with platitudes and pretty poetry, saying that all it takes is hard work and dedication. That our success is out there for the taking if only we have the fortitude to claim it. They peddle the empty hope and false promises of a meritocracy where all is ultimately as it should be.

What an unforgivable lie.

I am supposed to be using this space to thank the many, many people who have helped me through my doctoral studies and this dissertation – and I absolutely promise that I will get to that. But a dissertation isn’t just the accomplishment of a single person, and it isn’t merely the result of a small city of support. It is the product of a system.

A system that allows some people to succeed while blocking others from advancing. A system that is non-random in its treatment of subjects, in which the possession of certain personal characteristics is highly correlated with positive outcomes. In which success depends not only upon hard work, resilience, and intelligence but – perhaps even more fundamentally – upon luck and on privilege. No matter how qualified or tenacious you are, it is ultimately this system that determines whether or not you are allowed to succeed.

In other words, I cannot acknowledge the many people who have helped me through this process without first acknowledging the incredible privileges I’ve had along the way. My life has not always been easy, and it would certainly be inaccurate to interpret “privilege” as indicating a total lack of challenges or barriers. Rather, I use that word to acknowledge that – while some dimensions of my experience and identity have made my life harder – others have, undeniably, made it easier. Have made it possible, quite frankly.

The high school I mentioned earlier, for example – it was a good high school. One of the best in the country, to be honest. And I was never supposed to go there at all. I was never supposed to be able to get the kind of education I have been able to get, and I was never supposed to live the kind of life I have been able to live.
How do I even begin to thank people for that?

I went to school out of district, you see. Without the official permission I was supposed to get for doing so. Thanks to a lot of help from a lot of people. I went to public schools very unlike my local public schools. For Kindergarten through 8th grade, I went to a public school, magically nestled amongst the redwoods of northern California, where I was one of four students in my graduating 8th-grade class. After that, I went to an extremely privileged public high school. A school with extracurriculars, college prep classes, and new lab equipment. A school where students were safe and fed, where no one ever, ever got stabbed or shot.

It was very different from the neighborhood I grew up in.

Going to school out of district was challenging in a lot of ways. My parents would drop me off at public transit around 5 in the morning. I’d take the train and the bus – something like an hour and a half commute, if I remember correctly – then I’d sit in the hallway and do my homework while I waited for my first class to start. I was able to participate in after-school activities only thanks to the several teachers who quietly agreed to drive me home. Most teachers could only afford to live on my side of the hill anyway. And, of course, this whole arrangement was only possible thanks to my parents. They would regularly pick me up at some absurdly late hour – after theater, or dance, or choir rehearsal – despite needing to rise by 4am in order to get back to work.

I didn’t like high school at all. I liked my teachers. I liked my classes. But I...well, I frankly didn’t like the community. I just never fit in there. A poor, dirty girl wearing tattered hand-me-downs. I had more in common with the housekeepers who shared my early morning commute than I did with the children who lived in those houses. It was painful. Isolating. Exhausting. Constantly pretending to be something I was not.

But here’s the thing: I could pretend.

Despite how painfully uncomfortable I found high school to be, no one ever questioned
my right to be there. As intensely as I felt that I didn’t belong...no one ever suspected that I did not, in fact, belong. I could pretend to fit in because I looked like I fit in.

Every year, the Computing Research Association (CRA) conducts the Taulbee Survey¹, tracking the enrollment, graduation, and employment of Ph.D.s in computer science and related fields. According to the most recent data – and many thanks to Rediet Abebe for pointing me towards this survey – 1,521 people were awarded degrees in computer science in 2018. 19 were African American, and a mere 8 were African American women. 4 were Hispanic American women. 1 was a Native American woman, and 1 was a multiracial American woman.

Those numbers are bleak.

I grew up in East Oakland. I grew up in a majority Black neighborhood. My family was literally the only white family around. But the high school I secretly commuted to was around 75% white. In my senior year, the school newspaper published a front-page article featuring the faces of the 9 – 9 – Black students who were among the over-1000-member student body. The headline was some slightly more eloquent version of, “Hey, did you know there are Black students here?!”

Yeah, turns out people had noticed them.

So I don’t think it was an accident that I was able to fly by unnoticed. I don’t think it was luck that allowed me to get the education I was able to have. I don’t think it was just hard work that got me to where I am.

It is not coincidental to my success that I am white.

It’s important to acknowledge this. It’s important to acknowledge that the system is not fair, that who succeeds is not merely a matter of merit. Our whole lives are subtly shaped by the systems and institutions around us. These institutions are systemically unjust, built for a model “normal” which, as Kenji Yoshino would argue, not a single one of us completely conforms to.

¹https://cra.org/resources/taulbee-survey/
If we fail to acknowledge these systemic disparities, if we continue to spread lies about how any kid can grow up – then we risk ignoring all the work left to do.

My dissertation is rooted in civic studies. It is rooted in the belief that systems shape our lives and that our lives shape those systems. The challenges we collectively face are huge, daunting, and complex. We will not tomorrow, and, frankly, may not ever achieve some idyllic form of The Good Society. But this is not cause for nihilism or despair. It is not reason to throw up our hands and declare all efforts useless. We are not merely wayward waifs, caught in the current of forces beyond our control.

No. We are independent agents with the power to repair the world around us.

Every one of us.

But that power is not equally distributed. Those with more power, those most well-positioned to bring about systemic change, are also more likely to have benefited from the unequal systems currently in place. And because this privilege works by subtly removing barriers, by making difficult tasks just a little more tractable, those with power aren’t always aware of the lines of privilege which have supported their journey.

So I think it’s important to acknowledge this privilege. To acknowledge this power.

I am a first-generation to college, queer woman with chronic pain and significant mental health issues. But I still fall well within the bounds of a socially acceptable “normal.” Or at least I am both willing and able to pretend like I do. I am queer, but not that queer. My chronic pain is not typically incapacitating. And my mental health – well, I suppose I am fortunate to be broken in a very high-functioning sort of way.

My life has been difficult, but I have still been incredibly privileged. Across all my dimensions, I fall close enough to that fictional “normal” that I am allowed to maintain voice within existing power systems. The dimensions where I don’t fit are typically hidden: I can draw within the lines without disrupting the whole system.
I have had to fight hard for everything I have, but my privilege is what has made those battles winnable.

In moving towards more traditional acknowledgments, it feels appropriate to start by thanking the many therapists, councilors, and health care providers who have helped me over the years. I’d particularly like to thank Dr. Spottswood, who was, perhaps, the first therapist I ever really trusted. I started seeing her after my father passed away in 2012. That was another life-shattering event. An experience which destroyed the ersatz coping mechanisms I had accrued over the years and forced me to really figure out the kind of person I wanted to be. The loss of my father broke me completely. Giving me no choice but to rebuild myself. Better. Stronger. Faster.

Dr. Spottswood helped me come to terms with the strange, inherent tension of mental health. Appreciating who I am, who I’ve become, and who I am still becoming means appreciating, in a way, the myriad experiences that brought me here. I will fight like hell to protect others from many of the things I have gone through, but I cannot simply remove that part of my life. My trauma is an indelible part of who I am. Dr. Spottswood helped me appreciate that both those things can be true at once.

I don’t think I ever would have started a Ph.D. program without her.

More recently, Dr. Golkar saw me through the entire job market process as well as many of my years in graduate school. I literally don’t think I could have done this without her. There are many others, of course: therapists, councilors, workers who process my payments, schedule my appoints, and clean the offices where I’ve attended sessions. I owe them thanks as well.

I also owe a great deal of thanks to my committee.

Peter Levine, whom I’ve known for nearly 12 years, has been an incredibly positive force in my life. He is, I think, the first person who ever really believed in me. Or, perhaps, he was the first person who I ever really believed.

I have, as you may have guessed, very little patience for platitudes or empty optimism.
I am not interested in interpretations of democracy that are all sunshine and rainbows, where someday we’ll just roll the dice into a better society.

When Peter Levine says that every person’s voice matters, he truly means that every person’s voice matters. It is no platitude. He believes it in his bones, in his blood, in the very way in which he interacts with the world.

And he made me believe it, too.

When I first met Peter, I was deeply skeptical of deliberative democracy. Not because I thought it was too idyllic or too impractical, but, as I came to realize, for a much simpler reason. I could not possibly believe in a society in which every person’s voice mattered when I was so deeply convinced that my own voice was worthless.

My mother, an avid genealogist, once told me that no one in our family had ever done anything important and that no one in our family would ever do anything important. She loved tracking our family’s history because no one else would ever care.

Incidentally, she did tell me more recently that no one, in the entire history of our family, had ever received a PhD. Just a few generations ago, most people in my family were illiterate. “So when you get that degree,” she said, “it won’t be just for you. You’ll be representing all the generations which came before you.”

Peter Levine made me believe that I mattered.

And in doing so, he made me recognize that everyday people do shape the world around them. By making an immeasurable difference in my life, he made me appreciate that I had the power to make a difference in other people’s lives, too.

I believe deeply in deliberative democracy. I believe it in my bones, in my blood, in the very way I live my life. I am unrelenting in my conviction that we all benefit, individually and collectively, when everyone’s voice is heard. I believe that every single person - every single one of us - has something valuable to contribute. We each, then, have a moral obligation to not only contribute to public life, but to ensure that
all others are able to contribute as well.

I know that I am flawed. We are all of us flawed. There is no magic utopia where all our collective challenges will be resolved. But I stand firmly by the Levinian conviction that by working together, by talking together, by reasoning together – we can leave this world better than how we found it. We are not hopelessly bound by the unjust systems we inherit.

No. Those systems were built by people like us, and they can be changed by people like us.

Peter and I have been known to disagree. Yet, if this dissertation has an intellectual grandparent, it is Peter Levine. He made me a deliberative democrat and made me truly understand “democracy as a way of life.”

David Lazer is an equally important scholarly grandparent to this work. As my co-chair, the graduate studies director, and as the senior social scientist at the Network Science Institute, David has taught me so much about what it means to be a scholar, a scientist, and a computational social scientist. He has taught me how to be both skeptical and intellectually compassionate, to be both transparent about limitations and rigorous about interpretations.

He taught me that science is a slow, incremental endeavor. That our job is to move the needle in the gradual accumulation of human knowledge. That any single paper or study is flawed and incomplete, but by being intellectually honest and by bringing those tiny pieces together, we can learn great things. I keep a note of something he once said in class: “Science is an endeavor built flawed brick by flawed brick. Yet we can still build palaces.”

David has also taught me a lot about what good leadership looks like. What deliberative leadership looks like. As a student in the second-ever NetSI cohort and a founding member of our Graduate Student Council, I’ve had a lot of, ahem, constructive suggestions to share over the years. This past year, David closed out the
semester with a student town hall, a sort of Festivus airing-of-grievances, that went on for no less than *two and a half hours*. I could barely sit through it, to be honest.

David always listens seriously to others’ perspectives. He is clear in justifying his own positions and is open to being persuaded when he is wrong. Indeed, he is a paragon of the deliberative ideal – ardently standing by his convictions while simultaneously aiming to learn and be convinced by others.

He has also been a truly good friend and source of support throughout my PhD. Particularly over this past year, as I experienced the adventures of the academic job market, David was always there. I have sent him so many panicked, anxiety-induced emails, complaining about the injustices of the patriarchy, and doubting my own ability to survive and thrive. He has always been so kind, compassionate, and generous with his time. I have really needed his emotional, intellectual, and moral support throughout this process, and I can hardly thank him enough for everything he’s done for me.

David is one of the most thoughtful, moral, and genuinely good people I have ever met. I can only hope to be the kind of scholar and academic he has taught me to be.

I also owe a great deal of thanks to my co-chair and coauthor, Nick Beauchamp. One chapter of this dissertation is coauthored with Nick, and another builds off work we coauthored together. Nick has taught me a lot about how to develop a research idea and shape a research agenda. He’s been a consistent collaborator and has played an important role in helping me get from research ideas to published papers. I am very fortunate to have an advisor who shares so many of my research interests, and who has given me so much freedom during my doctoral studies. It is because of his mentorship that I am coming to the end of this Ph.D. with several published papers in my chosen research area.

I also very much want to thank the fourth and final member of my dissertation committee: Lu Wang. Lu has been a tremendous coauthor and mentor. She has taught me so much about Natural Language Processing and has given me valuable
feedback on this work. I very much appreciate all her contributions to this and our other projects.

Additionally, there are a number of other people from my academic life to whom I also owe thanks.

First, everyone at the Network Science Institute. NetSI has been such an amazing scholarly community and family. I want to thank Mark Giannini, the Ph.D. Program Coordinator, who has done so much to make the logistics of this program work. Also, Janette Briceno, James Stanfill, and Kate Coronges, who all do so much to help the Institute and its labs run. All my instructors: Lászó Barabási, Nick Beauchamp, Matteo Chinazzi, Sean Cornelius, Cody Dunne, David Lazer, Gabor Lippner, Chris Riedl, Roberta Sinatra, David Smith, Alex Vespignani, and Qian Zhang.

I’d particularly like to thank Chris Riedl. Chapter 2 of this dissertation started as a project for his Network Economics class. While the final model and paper have developed dramatically since then, his initial feedback and guidance were invaluable. Chris helped me think through what an Agent-Based Model of deliberation might look like, as well as the steps needed to incrementally develop such a model. Additionally, Chris’ office is directly behind me at the Institute, and, while it seems like a small thing – I also appreciate that he took the time to add padding to his cabinet doors. I have a terrible jump reflex, you see. It’s one of the most noticeable, though still subtle signs of my PTSD. I don’t think Chris ever knew why the sound of a closing cabinet made me involuntarily jump, but he noticed that I did, and he took steps to fix it. It was, really, incredibly kind.

Gabor Lippner, also, has been an incredible friend and colleague. I am so glad that through this program, I’ve had an opportunity to get to know both him and his family.

And, of course, a special thanks needs to go to Alex Vespignani – who taught me so much about how to consider, interpret, and model complex systems. Furthermore, I am writing these acknowledgments in the midst of the global COVID-19 pandemic.
Schools and businesses are closed, everyone’s working from home for the foreseeable future, and it feels like the worst is still ahead of us. It is a scary and uncertain time. I don’t think I’ve ever been so thankful for all of Alex’s work.

I never had the formal pleasure of taking a class with Brooke Foucault Welles, Tina Eliassi-Rad, or Sam Scarpino, but I have very much appreciated all their contributions to the Institute and community. They are all outstanding mentors and advocates, and I have benefited immensely from their support and from having them in my circle.

Similarly, I have very much appreciated the amazing scholars I have met while they were postdocs at NetSI. Particularly, Lisa Friedland, Nir Grinberg, Kenny Joseph, Evelyn Panagakou, Sebastian Ruf, Briony Swire-Thomson, Emma Towolson, and Stefan Wojcik.

Additionally, I cannot thank my NetSI classmates enough. We truly are a family, and they have been there for me every step of the way. My cohort Brennan Klein, Xindi Wang, Syed Arefinul Haque, and Matt Simonson have, in particular, been in the trenches with me since the beginning. I certainly could not have made it through this program without them. The current and past leadership of our Graduate Student Council: Ryan Gallagher, Leo Torres, and Carolina Mattsson, have also all played an important role in helping to grow and nurture this community.

Additionally, Jessica Davis, Ieke De Vries, Tim LaRock, Charles Levine, Stefan McCabe, Dina Mistry, Ronald Robertson, Timothy Sakharov, Sara Williams, and Chia-Hung Yang have all provided much-needed advice, support, and friendship throughout my graduate studies. I am also so glad to be able to count among my colleagues: Lucas Almeida, Sara Benedetti, Ayan Chatterjee, Hanyu Chwe, Mike Foley, Zach Fulker, Harrison Hartle, Wan He, Yanchen Liu, Ben Miller, Kunpeng Mu, Alexi Quintana Mathé, Ivan Viotalov, Xinyue Xiong, and Jingyi (Joy) Xu. Truly, NetSI is a tremendously welcoming, supportive, inclusive community in which we all bring out each other’s best. I am so lucky to have been a part of it.
Also at Northeastern, I’d like to thank Kirsten Rodine Hardy, who has been so kind and supportive over the last few years. She has listened to my practice talks, given me tremendous feedback, helped me navigate academia, is always there when I need to vent, and sends me pictures of her dog, Chip, whenever I request them. Also, Sarah Connell and the NULab for Texts, Maps, and Networks: Sarah has been a great friend, and the NULab has been an amazing resource for community and financial resources. Cara Marta Messina and Laura Johnson have also put in a great deal of work in helping to organize many NULab events and trainings. I’d also like to thank Alexia Ferracuti and the Center for Advancing Teaching and Learning Through Research, which has been an excellent resource for me. Additionally, several other Northeastern faculty have aided me in my journey: Meryl Alper, Aleszu Bajak, Colin Brown, Donghee Jo, Elizabeth Maddock Dillon, Julia Flanders, Laura Nelson, Costas Panagopoulos, and John Wihbey.

I am also incredibly thankful for the many current and former graduate workers at Northeastern who have taken the lead in union organizing efforts. This includes Alex Ahmed, Tim LaRock, Sam Maron, Liz Polcha, and James Robinson.

Beyond NetSI, there are so many others I have yet to thank.

Jennifer Victor, an amazing scholar and mentor, who has done so much to make me feel welcomed and supported. Mirya Holman, infamous founder of the Feminist Mafia, who is such an outspoken inspiration. Samara Klar, who is an absolute delight and is the founder of Women Also Know Stuff, an invaluable force within the field of Political Science. Rebecca Gill, who has done so much to make this discipline better. Amanda Bittner, who taught me that a group of female otters is properly referred to as a “raft of bitches.” Anand Sokhey, who, among other things, taught me that the Velveteen Rabbit had no hind legs. Lori Bougher, who is the best text-buddy a girl on the academic job market can have. Haley Norris, who is going to do such great things in this discipline. Stella Biderman, who helped remind me who I am. Felicia Sullivan, the first person to tell me that I could succeed in a doctoral program. Michael Neblo, who has been so, so kind and encouraging. Betsy Sinclair, who is
nothing short of a superhero. Rebekah Tromble, who was my first PolMeth discussant and is a veritable force to be reckoned with. Arthur Spirling, who I am so excited to work with following my PhD. Zander Furnas, an outstanding scholar, who also wrote my favorite tweet of all time (“Fly me to the moon / and launch me into the sun”). And, of course, my personal raft/meerkat mob: Del Bharath, Devon Cantwell, Siobhan Kirkland, Hannah Lebovits, Maricruz Osorio, Kate Petrich, Natalie Rohas, Rosalie Rubio, Rachel Torres, Gabi Vitela, and Rachel Winter.

Additionally, I owe thanks to the many, many other faculty, graduate students, and otherwise academically-adjacent folks who have supported me along the way. This includes: Terese Anders, Lisa Argyle, Julia Azari, Aeshena Badruzzaman, Mike Bailey, Kevin Banda, Sarah Bernt, Jodi Beneson, Rachel Blum, Sarah Bouchchat, Amber Boydstun, Jan Box-Steppensmeier, Anna Broido, Jessica Byrnes, Ben Campbell, Kelsey Campbell, Phil Chodrow, Dino Christenson, Olga Chyzh, Aaron Clauset, Mike Colaresi, Mia Costa, Cesi Cruz, Matt Denny, Bruce Desmarais, Amanda Sahar d’Urso, Laurel Eckhouse, Ryan Enos, Justin Esarey, Emily Farris, Mike Fix, Sandra González-Bailón, Lilly Goren, Justin Grimmer, Justin Gross, Alex Hanna, Michael Heaney, Dan Hopkins, Sarah Jackson, Lorien Jasny, Nathan Kalmoe, Kris Kanthak, Kei Kawashima-Ginsberg, Elsa Talat Khwaja, Abby Kiesa, Lorraine Klimowich, Rebecca Kreitzer, Chrissy Ladam, Hannah Lamarre, Juniper Lovato, Will Lowe, Janica Magat, David Masad, Shannon McGregor, Rachael McLellan, Burt Monroe, Laura Moses, Allison Morgan, Lucia Motolinia, Joshua Miller, Amanda Murdie, Hans Noel, Ioana Panatiu, Periloux Peay, Matt Pietryka, Cara Pina, Eric Plutzer, Jessica Preece, Tara Rigs, Molly Roberts, Pedro Luis Rodriguez, Luke Sanford, Kyle Saunders, Jessica Schoenherr, Alice Schwarze, Ian Shapiro, Yotam Shmargad, Noah Smith, Kelsey Shoub, Alexandra Siegel, Jane Sumner, Catriona Standfield, Nikki Stevens, Michelle Torres, Josh Tucker, Kim Turner, Annie Waldherr, Mike Wagner, Laila Wahedi, Joe Wasserman, Julie Wronski, Bi Zhao, and Justin Zimmerman.

On more personal notes, I am very thankful to Janet Kelley and all the wonderful members of my book club: Julia Africa, Allison Hastings, Courtney Hering, and Rebeca Salguero Palacios. Additionally, a few other dear friends who have done so
much for me over the years: Jessica Byrnes, Nancy Wilson, Jamie Barret-Riley.

I am also incredibly thankful to The Training Room, my wonderful, locally-owned gym, which has done so much to keep me sane these last few years. I’d especially like to thank owner Maren Kravitz, and trainers Chris Mullins, Wes Luce, and Alex Tanskey. And an especially strong thank you to Rob Colameta. Rob has been a driving force in increasing my strength and helping me achieve the other big goal I accomplished this year: doing a single-arm, overhead, strict press with 28 kg (62 lbs). Looking forward to crushing 30 kg next.

Finally, I would like to thank my family.

Especially my incredibly supportive partner, Peter Jehlen. I’m pretty sure I would have starved to death long ago if it weren’t for him. He is funny, kind, incredibly caring, and has supported my choices to really throw myself – and continue to throw myself – into this work. I don’t know what I would do without him. He gets all the points.

Also, to my sister, Rebekka Shugas, who has supported me through so many difficult times. And to my mother, Suzanne Todd, who has worked so hard to give me the opportunities I’ve had.

Finally, I want to end by thanking my father, who I miss so much, every day. If he were here to see this, I have no doubt of what he’d say:

*I hope you know that this will go down on your PERMANENT RECORD.*
Abstract of Dissertation

Democracy is fundamentally an interactive endeavor that relies upon the collaborative reasoning of its citizens. In formal governing bodies, online interactions, and informal conversations, people engage in political talk in which they use reasons to convince, persuade, or justify. These interactions form a rich deliberative system which spans every level of political life, providing myriad venues for both everyday citizens and elites to clarify their thinking, share relevant information, persuade and be persuaded, and have a voice in determining what topics are important and how those topics should be discussed. At its best, such a deliberative system allows every citizen the power to both develop an informed opinion and play a role in developing the opinions of others. Of course, there is good reason to believe the ideal deliberative system has not been achieved — systemic power disparities, limits to human cognitive capacity, and tendencies towards blind partisanship all make a functional deliberative system appear desperately far away. However, if we are to accept as an ideal that citizens in a democracy should have their voices heard — that they should talk, reason, and work with others around matters of common concern — then we have an obligation to work towards that democratic ideal. This dissertation, therefore, aims to better understand several crucial elements of the deliberative system. Can a process of sharing reasons and ideas bring a group to better decisions? Do everyday citizens express themselves in individually distinctive ways or simply repeat party talking points? What drives people to engage in ongoing, repeated interactions around political topics? How can we measure phenomena related to the deliberative system in order to design and test possible interventions? This dissertation aims to address all these questions through a focus on three elements of the deliberative system: the exchange of arguments, the structure of individual-level reasoning, and the dynamics of ongoing conversation. Each of these elements has unique networked dynamics — they can be understood most fully by considering interconnections along a range of dimensions. Social networks influence the people and ideas we come into contact with, semantic networks influence the very words which are exchanged, and conversation itself has a network structure as participants respond to each
other’s comments. Indeed, the very notion of a “deliberative system” implies a richly interconnected symphony of dialogue. In short, the deliberative system cannot be thoroughly interrogated without network methods capable of considering the interconnectedness of reasoning and engagement. Therefore, each chapter further aims to build methods and models capable of strengthening our understanding of these richly interconnected systems. Specifically, this dissertation examines the effects of cognitive capacity and bias through an agent-based model, finding that agents can discover good solutions if they are willing to accept good information from others. It further explores the semantic and cognitive connections inherent in expressing a political view, finding that individuals express themselves in distinctive ways that are correlated with known personality traits and not merely a result of ideological talking points. Finally, this dissertation delves into the drivers of engagement in real-world, online conversations, finding that these conversations are surprisingly predictable, often driven by emotional and topical salience, and occur over a range of ideological divides. Taken together, this work suggests that a functional deliberative system may indeed be obtainable, and it provides conceptual frameworks and methodological tools for moving towards such an ideal.
# Table of Contents

Dedication .................................................................................................................. 2

Acknowledgments ......................................................................................................... 3

Abstract of Dissertation ............................................................................................... 18

Table of Contents .......................................................................................................... 20

List of Figures ................................................................................................................ 23

List of Tables .................................................................................................................. 24

Chapter 1 Introduction .................................................................................................. 25
1.1 Reaching for the democratic ideal ............................................................................ 28
1.2 Network methods for a deliberative system ............................................................... 31
1.3 Reasoning together .................................................................................................... 33

Chapter 2 Collaborative Reasoning on Value-Laden Topics: A Game of Giving and Asking for Reasons .................................................................................................................. 35
2.1 Introduction ................................................................................................................ 35
2.2 Related Work .............................................................................................................. 39
2.3 Methods .................................................................................................................... 41
   2.3.1 Ground truth solution space ............................................................................. 42
   2.3.2 Individual beliefs ............................................................................................. 47
   2.3.3 Reasoned exchange ......................................................................................... 51
   2.3.4 Experiments and measures ............................................................................. 53
2.4 Results ...................................................................................................................... 58
   2.4.1 Effects of cognitive capacity and ideological skew ........................................... 58
   2.4.2 Effect of cognitive capacity and open-mindedness .......................................... 61
2.5 Discussion ................................................................................................................. 62

Chapter 3 The Structure of Reasoning: Measuring Justification and Preferences in Text .................................................................................................................. 66
3.1 Introduction ................................................................................................................ 66
3.2 Related Work .............................................................................................................. 69
3.3 Data .......................................................................................................................... 74
   3.3.1 Amazon’s Mechanical Turk ............................................................................. 75
6.1.2  Simulation Steps ........................................ 164
6.1.3  Explored Parameters and Findings .......................... 165
6.1.4  Static Parameters and Limitations ............................ 177
6.2  Chapter 3 Appendix ........................................... 180
6.2.1  Prompts .................................................. 180
6.2.2  Survey Questions ............................................ 181
6.2.3  The Big Five Inventory ....................................... 185
List of Figures

Figure 2.1: Example influence patterns with $N = 4, K = 2$ ............... 43
Figure 2.2: Example policy platform calculation .......................................... 45
Figure 2.3: Example policy vote .............................................................. 56
Figure 2.4: Deliberative outcomes by reasoning type. ............................. 59
Figure 2.5: Deliberative outcomes portion of intellectuals ......................... 60
Figure 2.6: Deliberative outcomes by open-mindedness. ............................ 62

Figure 3.1: Example of the grammatical parse of a sentence ..................... 82
Figure 3.2: Comparison of network statistics across three stylized example networks. Each network has $N = 16$ and $E = 24$ .......... 88
Figure 3.3: Correlations between network statistics and latent personality and demographic measures ...................................................... 90
Figure 3.4: Distribution of network similarity for authentic respondents. “Self” captures similarities between a single respondent’s liberal and conservative text, while “Ideological” captures between-subject similarity for a fixed topic and ideology. A similarity of 0 indicates that networks have identical portraits. ....................... 94
Figure 3.5: Distributions of common quality metrics between authentic and inauthentic respondents ....................................................... 95

Figure 4.1: Distributions of length and users by thread ............................. 108
Figure 4.2: Cumulative distribution of the time taken to reply to a tweet for those tweets with replies. ................................. 109
Figure 4.3: Hand-coding of tweet pairs by sentiment and agreement. Each plot represents 100 pairs randomly sampled from the closest (a) or furthest (b) tweet pairs. ................................. 130
Figure 4.4: Correlation between tweet $t$ and response ......................... 133

Figure 6.1: Uninformed: Comparing the Number of Agents ..................... 166
Figure 6.2: Intellectuals: Comparing the Number of Agents ..................... 167
Figure 6.3: Ideologues: Comparing the Number of Agents ...................... 168
Figure 6.4: Uninformed: Comparing Different Openness Levels (10 Agents) 170
Figure 6.5: Intellectuals: Comparing Different Openness Levels (10 Agents) 171
Figure 6.6: Ideologues: Comparing Different Openness Levels (10 Agents) 172
Figure 6.7: Comparing Mix of Uninformed and Intellectual Agents .......... 175
Figure 6.8: Comparing Mix of Ideologue and Intellectual Agents ............ 176
List of Tables

Table 2.1: Example influence values for $N = 4$ policies . . . . . . . . . . . . . 44
Table 3.1: Measures of network structure. . . . . . . . . . . . . . . . . . . . . 87
Table 3.2: Results of logistic regressions. . . . . . . . . . . . . . . . . . . . . 98
Table 4.1: Example conversation flow between participants $A, B$ and $C$. . . . 111
Table 4.2: Top 10 words for each topic. . . . . . . . . . . . . . . . . . . . . . 118
Table 4.3: Response predictors: Candidate Respondent . . . . . . . . . . . . . 122
Table 4.4: Response predictors: Candidate respondent’s previous tweet . . . 124
Table 4.5: Response predictors: Conversation features . . . . . . . . . . . . . 125
Table 4.6: Response predictors: Current tweet author . . . . . . . . . . . . . 127
Table 4.7: Response predictors: Current tweet . . . . . . . . . . . . . . . . . 128
Chapter 1

Introduction

Collaborative human reasoning lies at the core of democratic life. Democracy is fundamentally an interactive endeavor in which people use reasons to convince, persuade, or justify. Such interactions take place within the hallowed halls of formal governing bodies, among elites with the ear of the public, around family dinner tables, and in virtual, online spaces. Every level of political life involves human interaction and is couched within a framework of ostensibly reasoned debate. Of course, it could well be the case that many of the stated “reasons” are simply post-hoc rationalizations, strategic political maneuvers, or recapitulations of elite talking points. And it is perhaps equally likely that people engage in such “collaborative” discussions without a genuine openness to learning from others. Yet, neither the collaboration nor the reasoning must meet some Platonic democratic ideal in order for the functional description of collaborative reasoning to hold: people within a shared sphere articulating reasons regarding matters of common concern.

It is this observed phenomenon – people talking about and talking together about politics – which comprises the focus of my dissertation. Such political talk is fre-
quently taken to be a core tenant of democracy; serving to generate public opinion (Cohen, 1989; Habermas, 1984) and build democratic legitimacy (Dryzek, 2009) by helping citizens better understand their own views, exchange factual and normative information, and reason together about matters of common concern (Mansbridge, 1999; Rawls, 1971). Political talk amongst everyday people is of particular interest and is widely taken to form the very fabric of democracy itself. As Dewey (1937) argues, democracy is far more than a mere system of government: democratic life is generated through interpersonal relations and is characterized by the valuing of human interaction both mundane and sublime. Democracy, at its core, is the art of associated living. We are not one person by day and citizens by night, we are whole people whose experiences and perspectives permeate every aspect of our lives.

In this sense, civic life is deeply characterized by the problem of embeddedness (Granovetter, 1985). Citizens are both independent agents capable of making their own choices and constrained by the accumulated culture of past generations. That is, existing cultural norms and institutions provide the context in which citizens act, but simultaneously, the actions of citizens can reshape those norms and institutions. In other words, our political culture is continuously generated through the collaborative reasoning processes which result from cohabiting the world. We are each individually shaped by our own experiences and perspectives, but we collectively create the political context around us as we handle workplace politics, deal with family members’ differing needs, and navigate through cities teeming with different people moving towards different places at different speeds.

This embeddedness highlights the need for a rich deliberative system (Mansbridge, 1999) in which both formal and informal talk cultivates relevant ideas and – if the system works – filters for the “best” ideas. Such a system would include elite discourse, communication between elected officials and constituents, as well as both public and private conversations between everyday citizens. Some of this political talk may
meet, or come close to meeting, the ideals of deliberation (Mansbridge, 2015), but much of it inevitably will not. Regardless of whether this talk is ideal or could even be properly labeled as “deliberative,” the beauty of a deliberative system is that it provides a venue for information to be generated, considered, and aggregated beyond the contexts in which individuals are embedded.

For example, what counts as “political” is itself a question negotiated by both individual concerns and the larger context. As Hanisch (1969) memorably put it: “the personal is political.” When individual struggles – no matter how seemingly mundane – are interconnected with larger, value-laden issues of social justice, they are “political” insofar as they are topics worthy of public discussion (Mansbridge, 1999). More broadly, the dominant cultural context in which we are embedded frames the bounds of what topics are suitable for public discussion (Gaventa, 1982; Flyvbjerg, 1998) and how such issues should be framed (Eliasoph, 1998). These are matters of power, not handed down by nature. Counterpublics (Fraser, 1990), whose constituents bring additional cultural context beyond that experienced or accepted by the mainstream, actively seek to shift the content and bounds of public discussion. This negotiation takes place within the deliberative system: activists strategize, politicians posture, people talk in ways formal and informal – and that mere talk has the potential to change the bounds and norms of the public sphere. In this sense, political talk isn’t simply the articulation of political preferences, but can more properly be understood as the venue for ongoing cultural debate. It is the mechanism through which citizens and elites alike can clarify their thinking, share relevant information, persuade and be persuaded, and have a voice in determining what topics are important and how those topics should be discussed.

This dissertation examines several key elements of the deliberative system. Stemming from work in deliberative democracy, this dissertation imagines citizens with the ability to be reasoned, collaborative agents capable of co-creating a shared world. Yet
this work also recognizes that humans are imperfect and that the deliberative ideal is not always – perhaps is even rarely – met. It aims to build an understanding of what real-world political talk looks like, examine how openness and cognitive capacity interact with the deliberative ideal, and to develop new methods capable of measuring individual-level reasoning structures. That is, it builds upon the deliberative literature in believing that political talk is necessary for a functional democracy and aims to develop tools and insights which could help us move towards this democratic ideal.

1.1 Reaching for the democratic ideal

Deliberative democracy is arguably an overly romantic notion. Everyday citizens engaging in thoughtful, respectful dialogue may seem like the epitome of democratic virtue, but, detractors argue, it is so far from reality as to be a poor model for any sort of real-world governance. We might hope for government truly by the people, but as a practical matter must ultimately settle for institutions that govern for the people.

In many ways, this concern about the practicality of deliberative democracy is reasonable: people are observably imperfect and known to be prone to numerous cognitive and social biases. Arguably, the great mass of men are simply unfit to govern: they are too easily led astray by false beliefs and too easily manipulated by those with power (Lippmann, 1922). Or, perhaps, they are too conflict-averse to engage properly in deliberative life; preferring to side with the perceived majority (Sunstein, 2009), going out of their way to avoid possible disagreement all together (Eliasoph, 1998), or becoming hopelessly disengaged upon their first discovery of political friction (Mutz, 2006). Perhaps, even, the average person dislikes engaging with political life, and would far prefer to delegate the task, if only their representatives were appropriately trustworthy custodians (Hibbing and Theiss-Morse, 2002). A great deal of scholarly
attention has further focused on the many failings of human reasoning - the mistakes we make, individually and collectively, when we act too quickly, when we rationalize reflexive reactions as reasoned beliefs, or when we are too easily swayed by social pressure (Festinger, 1954; Graham et al., 2011; Haidt and Joseph, 2008; Kahneman, 2011; Sunstein, 2009). Taken together, these bodies of work suggest that the average citizen simply does not have the time, interest, or expertise required to develop coherent and informed political views. Given this dire view of citizen interest and capability, one might reasonably abandon the vision of people actively shaping their shared world and instead re-envision democracy as a transactional guardian designed to protect the people from both the state and themselves.

Yet if we believe in the democratic value of political talk – if we believe, regardless of practicality, that citizen engagement truly is the ideal – then we should not be so quick to settle for such a cynical alternative. There is compelling empirical evidence to suggest that given thoughtful design and appeals to participants’ democratic sensibilities, deliberative discussions can lead to better democratic outcomes, (Fishkin, 2014; Knobloch et al., 2013; Neblo et al., 2010). Furthermore, rather than a liability likely to shut down conversation, differences of opinion and perspective can be crucial to developing better ideas and solutions (Burt, 2004; Hong et al., 2004; Page, 2008). Arguably, then, what citizens need is not freedom from the duties of civic engagement, but more opportunities to participate thoughtfully and meaningfully (Levine, 2015; Neblo, 2015; Neblo et al., 2018). In this sense, the cynical approach – while intended to protect democracy from a fickle and flawed public – may simply serve to exacerbate the problem; training citizens to internalize an image of themselves as consumers who cannot be trusted with the awesome power of governance.

Far from being rosy-eyed optimists, deliberative democrats envision a perpetual system of negotiated growth and change (McKenna, 2001). They do not argue that people are perfect nor believe that all stated reasons are truly rational. Rather, they
believe that when citizens feel the full weight of civic obligation upon them, they are capable of doing their imperfect best to work towards a better world. That is, the deliberative vision assumes that citizens want to live in a good society, even as the very notion of “good” is itself fraught with debate.

The full deliberative system provides the venues for this perpetual negotiation. Citizens form their opinions through personal reflection (Goodin, 2000; Neblo, 2015) and “everyday talk” with peers (Mansbridge, 1999). These constant, informal conversations aggregate to become public opinion (Cohen, 1989; Habermas, 1984) which in turn builds democratic legitimacy (Dryzek, 2009) as formal governing bodies incorporate the collective views of their constituents into their own deliberative processes. Ultimately, this system forms a virtuous cycle as citizens, inspired by the democratic outcomes of their deliberative practice, continue to dedicate their time and energy to deep engagement in public life and imbue future generations with similar civic norms and habits (Levine, 2015; Neblo, 2015; Neblo et al., 2018; McKenna, 2001).

Chapter 2 interrogates the functional bounds of such a deliberative system by measuring the degree to which several canonical failures of deliberation may harm group decision making. Can citizens with poor cognitive capacity possibly come to good decisions? What role does ideology and polarization play in mediating flawed extremes? How accepting of others’ views should citizens be in order for groups to arrive at good decisions? In addressing those questions, this chapter presents an agent-based model of collaborative reasoning with tunable parameters related to agents’ capacity for reasoning, ideology, and openness. In a deliberative game of “giving and asking for reasons” Neblo (2015), agents share beliefs around possible policy initiatives and attempt to enact “good” policies through a process of mutual exchange and consideration. Agents hold unique normative views as to what constitutes “good” and aim to convince and be convinced (Mercier and Landemore, 2012) as to the best course of action. This model reveals that polarized groups do surprisingly well
at identifying optimal policy solutions, even when their agents have low cognitive performance. These findings bolster the vision of a functional deliberative system, suggesting that positive deliberative outcomes can be achieved by imperfect agents who individually fall short of a lofty, deliberative ideal.

The model in Chapter 2 assumes an interconnected policy landscape – e.g., a system in which the implementation or non-implementation of one policy may influence the value of implementing or not implementing another policy. However, the whole deliberative system could more broadly be considered as networked along multiple dimensions. Social networks influence the people and ideas we come into contact with, semantic networks influence the very words which are exchanged, conversation itself has a network structure as participants respond to each other’s comments, and the very notion of a “deliberative system” implies a richly interconnected symphony of dialogue.

In short, the deliberative system cannot be fully interrogated without network methods capable of considering the interconnectedness of reasoning and engagement. Chapters 3 and 4, therefore, delve into this challenge; each developing new network methods appropriate for distinct elements of the deliberative system.

1.2 Network methods for a deliberative system

Arguably, the deliberative system starts not with interpersonal discussion, but with an individual process of “deliberation within” (Goodin, 2000; Neblo, 2015). This personal reflection is inherently networked: citizens cognitively retrieve information based on perceived interconnections (Collins and Loftus, 1975; Quillian, 1967; Shavelson, 1974) and weigh the pros and cons of those connected ideas in order to reach their final judgments (Axelrod, 1976; Carley, 1993). These individual-level conceptual network
structures could then play a role in interpersonal deliberation, as discussants draw on these same cognitive networks in order to articulate and justify their views to others (Toulmin, 1958; Walton, 1996).

For decades, scholars have argued for the value of measuring the network structure of individuals' justification for their preferences (Axelrod, 1976; Carley, 1993). The ability to measure and interrogate individuals' expressions of political reasoning holds the potential to shed new light on the dynamics of public opinion and political behavior (Lane, 1962; Axelrod, 1976; Campbell, 1960). Questions of persuasion, ideological fracturing, and conversation quality all rely upon understanding individual styles of political expression. These dynamics are driven not just by what someone says but, indeed, by how they say it.

Despite the long-standing interest in this topic, the task has only recently become tractable with the emergence of modern computational methods. Chapter 3, therefore, takes advantage of new computational techniques in order to infer the conceptual network structure of individuals' political reasoning. This text-based method identifies the key concepts a person raises and examines the implicit connections between those concepts. This approach provides insight into the quality of a person's reasoning and reveals meaningful individual variation which is correlated with known behavioral traits and reflective of more than mere ideological talking points.

Finally, no examination of the deliberative system is complete without studying the messy reality of informal, real-world, everyday political talk. Such settings serve as the primary venue through which average citizens engage in the deliberative system (Mansbridge, 1999) and are therefore essential to the overall functioning of the deliberative system itself. While conventional wisdom regularly bemoans the polarization and toxicity of these conversations, very little scholarly work has aimed to deeply understand them or their dynamics. Why do people engage in extended
political conversations? What brings them back to repeated interactions and what role do these network dimensions play in determining the quality and content of a conversation?

Chapter 4 answers these questions through a new model of argument engagement and a collected corpus of Twitter conversations about President Trump. The model incorporates numerous user, tweet, and thread-level features to predict user participation in conversations with over 98% accuracy. This work finds that users are likely to argue over wide ideological divides, and are increasingly likely to engage with those who are different from themselves. Additionally, the emotional content of a tweet has important implications for user engagement, with negative and unpleasant tweets more likely to spark sustained participation. Though often negative, these extended discussions can bridge political differences and reduce information bubbles. This suggests a public appetite for engaging in prolonged political discussions that are more than just partisan potshots or trolling, and the results suggest a variety of strategies for extending and enriching these interactions.

1.3 Reasoning together

Democracy is a collaborative enterprise that demands the engagement of its citizens. Such engagement need not be flawless or perfectly reasoned – rather, a functional democracy should seek to cultivate a deliberative system in which ideas are generated, considered, and appropriately filtered and in which every citizen has a meaningful voice. Such a system is inherently networked along a range of dimensions: self-reflection, everyday talk, and formal deliberation are all indelibly intertwined, policies may have effects on each other, words and ideas have complex connections, and conversations have observable patterns of engagement and disengagement.
These interconnections form the core motivation for this dissertation, which aims to develop network methods to support a better understanding of elements of the deliberative system. Chapter 2 examines the effects of cognitive capacity and bias on a deliberative system through an agent-based model. Chapter 3 explores the semantic and cognitive connections inherent in expressing a political view, and chapter 4 delves into the drivers of conversational engagement. Taken together, this work presents a rich picture of deliberative life which is dynamic, complex, and imperfect. It suggests new theoretical and empirical frameworks for engaging with the deliberative system and builds tools for diagnosing the real-world state of deliberative life.
Chapter 2

Collaborative Reasoning on Value-Laden Topics: A Game of Giving and Asking for Reasons

2.1 Introduction

Collaborative reasoning - whether successful or unsuccessful - is a core facet of human society. In business, politics, and everyday life, individuals with varying opinions, experience, and information attempt to collaborate and make decisions. However, the ability of these groups to come to “good” decisions may be hindered by both group dynamics and individual failings. Such challenges are a core concern among detractors of deliberative democracy who argue that it is overly idealistic to imagine everyday citizens capable of successfully reasoning together about matters of common concern. Such efforts are arguably doomed to fail due to individuals’ limited cognitive

\[^1\text{Replication materials for this chapter can be found at https://github.com/sshugars/collaborative_reasoning.}\]
capacity (Lippmann, 1922; Shah and Oppenheimer, 2008; Oswald and Grosjean, 2004; Tversky and Kahneman, 1981), the inability of divided factions to agree (Dworkin, 2006; Madison, 1787), and tendencies for individuals to accept or reject information on inaccurate grounds (Sunstein, 2002; Festinger, 1954; Janis, 1972; Nickerson, 1998).

These concerns about group decision making are heightened in political settings where citizens may hold subjective, value-laden perspectives, and where the “truth” is hard or impossible to know. Such settings are importantly different from deliberation under factual circumstances where individuals may be limited in their knowledge but would fully agree if given access to the same information. In these factual settings, collaborative reasoning can be well-modeled under an explore/exploit framework where agents attempt to find a global optimum given local information (Lazer and Friedman, 2007; Mason and Watts, 2012). However, if the topic is value-laden, if agents hold their own subjective opinions of the solution space, existing models can neither explain nor predict the process of collaborative reasoning.

However, a great deal of real-world problems rely upon agents coming to normative judgments. In the political realm, for example, a policy solution is only optimal if it results in outcomes an agent would qualify as “good.” Political polarization, in this sense, does not necessarily represent an inability to discover and share relevant information, but rather a deeper disagreement as to the value of that information. Despite skepticism to the contrary, numerous empirical studies demonstrate that people are able to productively discuss value-laden matters (Fishkin, 2014; Knobloch et al., 2013; Neblo et al., 2010, 2018), suggesting a growing need to understand the conditions under which these conversations succeed.

This paper presents an agent-based model of collaborative reasoning on value-laden topics, drawing upon literature on group problem-solving (Lazer and Friedman, 2007; Mason and Watts, 2012; March, 1991) and belief convergence (DeGroot, 1974;
Friedkin et al., 2016). In a deliberative game of “giving and asking for reasons” (Neblo, 2015), agents share beliefs around possible policy initiatives and attempt to enact “good” policies through a process of mutual exchange and consideration. The policy landscape is taken to be a complex system in which the implementation or non-implementation of individual policies has non-linear effects on the value of an overall policy platform. In order to assess the ground truth “value” of deliberative outcomes, this model considers settings in which members agree as to the best overall outcome but disagree as to the specific policies which come closest to achieving that outcome. For example, we might imagine citizens who all want to live in healthy, safe, communities with access to good education, but might also imagine those same citizens disagreeing on both the degree to which those goals are obtainable and the set of policies which come closest to bringing about those goals. In such a setting, the ground truth solution would describe the optimum set of policies needed to achieve the best possible outcome.

The model, described in detail in Section 2.3, considers a solution space in which the implementation or non-implementation of policies affects the value of other policies. Specifically, the model leverages the $NK$ solution space in order to describe a set of $N$ policies whose values are each mediated by $K$ other policies. Initially conceived of in the context of adaptive evolution (Kauffman and Weinberger, 1989; Kauffman and Levin, 1987), $NK$ models have been widely applied to organizational and group problem-solving tasks (Levinthal, 1997; Lazer and Friedman, 2007; Geisendorf, 2010; Herrmann et al., 2014; Shore et al., 2015).

Furthermore, this paper considers deliberative outcomes for imperfect agents who are prone to several individual and group failures. The mere existence of such failures is often taken to imply that deliberative success is an unrealistic and unachievable goal (Lippmann, 1922; Dworkin, 2006; Sunstein, 2002). However, the extent to which any one of these failures dooms deliberation or fits within a broader deliberative
system (Mansbridge, 1999) has remained largely unaddressed. Specifically, this paper determines the deliberative impacts of three canonical failures of deliberation: limited cognitive capacity, group factions, and tendencies to make poor judgments when accepting or rejecting other’s views. As described in Section 2.3, uninformed agents reflect concerns that individuals have limited cognitive capacity and are therefore unlikely to hold views closely reflective of any underlying ground truth (Lippmann, 1922; Shah and Oppenheimer, 2008; Oswald and Grosjean, 2004; Tversky and Kahneman, 1981). Ideologues reflect concerns that polarized factions will remain entrenched and divided with conflicting viewpoints (Dworkin, 2006; Madison, 1787). Finally, a tunable open-mindedness parameter captures both the effects of groupthink and confirmation bias, as individuals are either too quick to accept group views (Sunstein, 2002; Janis, 1972; Festinger, 1954) or, alternatively, reject any view which does not closely conform to their priors (Nickerson, 1998). This paper, of course, does not address all concerns about deliberation. Most notably, this model assumes that all agents participate openly and equally and therefore does not consider that people may generally shy away from conflict (Mutz, 2006; Eliasoph, 1998), may have little interested in engaging (Hibbing and Theiss-Morse, 2002), or may be shut out of deliberative discussions by those with more power (Sanders, 1997; Gaventa, 1982). Nevertheless, this model allows us to isolate and examine several core concerns about the practicality of deliberation.

Running this model with small groups of deliberators, I find that ideologues – whose skewed beliefs reflect concerns about factions and polarization – do surprising well at identifying optimal policy solutions. Indeed, while concerns about individuals’ cognitive capacity appear to be well-founded, factions of oppositely-skewed peers can overcome this limitation in settings where they are willing to listen to each other. This finding suggests that factions – while ostensibly contentious – may actually balance each other out and come to an optimal middle ground.
2.2 Related Work

A long line of work has leveraged agent-based models to examine the dynamics of group consensus and dissensus. Capable of interrogating and isolating the myriad elements which influence group decision making, these stylized models have proven to be a fruitful supplement to human subject experiments. This work was primarily pioneered by DeGroot (1974), who imagined small teams tasked with reaching consensus about the ground truth value of some parameter, $\theta$. In this model, each agent holds a unique probability distribution as to the value of $\theta$ and also assigns a value to the opinion of each other agent. Trust is a significant variable of interest in this model, and, indeed, DeGroot (1974) finds that the group’s opinion will converge if a single agent’s opinion is positively valued by all players.

March (1991) more explicitly examines the organizational context, presenting a model in which individuals are socialized to an organization’s beliefs while the organization simultaneously learns from its members. March (1991) frames this as an explore/exploit trade-off where individuals and the organizational code may either converge quickly and exploit sub-optimal solutions, or converge slowly and explore better solutions. March finds that the inefficiency introduced by the presence of ‘slow’ learners helps the system converge optimally rather than settling on a local peak. This finding is echoed in more recent work on solution convergence in group problem-solving tasks. In work that also uses the $NK$ model as a solution space, Lazer and Friedman (2007) find a notable “trade-off between maintaining the diversity necessary for obtaining high performance in a system and the rapid dissemination of high-performance strategies.” While the human-subject experiments of Mason and Watts (2012) provide partial validation for the results of Lazer and Friedman (2007), they notably find that human subjects were able to benefit simultaneously from a diversity of opinions and rapid dissemination. This suggests that while rapid dissemination allowed subjects to adopt others’ approaches, users were set enough
in their ways that this mechanism did not automatically result in convergence on sub-optimal solutions as simulations predicted.

Models of consensus and belief systems are particularly relevant for deliberative theory, where decision making is often taken to be a required outcome (Mansbridge, 2015) of an exchange of reasons (Gutmann and Thompson, 1998; Neblo, 2015; Mansbridge, 2015). In this sense, political deliberation can be taken as a group problem-solving task, where participants come in with private information and search for the optimum of a complex solution space through an iterative process of search and knowledge exchange. In this task, deliberative agents face a similar explore/exploit trade-off: good faith discussants ought to “convince others and be convinced when appropriate” (Mercier and Landemore, 2012). That is, a person engaging in political discourse ought to aim to believe true things – but doing so requires a careful balance between intellectual humility and self-confidence. When encountering another person’s viewpoint, a good-faith deliberator ought to carefully consider that view, assess its accuracy to the best of their ability, and then either adopt this alternative view or provide a reasoned explanation as to why their interlocutor is wrong. While this may not be the most common form of political discourse, it is worth noting that such reasoned exchange can occur (Fishkin, 2014; Knobloch et al., 2013; Neblo et al., 2010). Indeed, such productively collaborative exchange is the backbone of academic discourse.

Explicitly bringing belief systems to the deliberative domain, Altafini (2013) considers the case of “agreed upon dissensus” or “bipartite consensus” where antagonistic agents converge on the same values but with different signs. In political talk, this constitutes the case where deliberators “agree to disagree” - each finding the other’s argument to be reasonable, rational, and wrong. Choi and Robertson (2013) further model belief systems in collaborative governance. Comparing the outcomes of deliberation across various decision rules (unanimity, supermajority, and dominant coalition), they find that deliberation is more important than voting rules in building consensus and
enhancing decision quality. Friedkin et al. (2016) study belief systems with logic constraints. Examining the convergence of opinion when arguments are rationally linked, they find that a strongly linked logic structure can shift people away from their initial beliefs, though the presence of unrelenting critics can mitigate this effect.

2.3 Methods

This paper imagines small groups of citizens deliberating about a set of interconnected policy solutions that all influence an overarching social issue. Agents are assumed to agree on the ideal overall outcome but disagree about the specific policies which best move towards that ideal. For example, a group discussing public education would share a desire for students to get a good education but might disagree on what makes an education “good” or on which specific policies meet the needs of most students. A group discussing healthcare would agree that, ideally, everyone would have access to affordable healthcare, but they would disagree as to feasible solutions for providing the best healthcare to the most people. Similarly, a group discussing crime might all hope to see a reduction in crime while also disagreeing as to what policies are most likely to bring about such a reduction.

While agents share an overarching goal and consider a common set of possible policies, they hold differing views as to the value and implication of each policy under consideration. These individual-level views, which may be driven by normative or practical considerations, lead agents to begin deliberation with different beliefs about the best set of policies to implement. During deliberation, agents take turns sharing and considering reasons for why a policy should or should not be implemented. These reasons, which will be described in detail in Section 2.3.3, can be intuitively understood as the broader implications of implementing a policy, beyond the value of that policy itself. After each exchange, agents give every policy an individual
up or down vote, resulting in an implemented policy platform. The value of the implemented platform is then judged against a ground truth solution.

This model, then, requires three essential components which will be described in detail below: (1) a ground truth solution space against which we can benchmark deliberative performance, (2) individually-initiated belief spaces representing the views of each deliberating agent, and (3) a process of reasoned exchange in which agents share views and decide whether to accept the views of others. Within this framework, the paper evaluates the impact of three canonical deliberative failures. Concerns about the limits of humanity’s cognitive capacity (Lippmann, 1922; Tversky and Kahneman, 1981) are tested by manipulating the initial accuracy of individuals’ beliefs. Fears about the inability of polarized groups to productively collaborate (Dworkin, 2006) are examined by inducing coalitions which are divided in their initial beliefs. Finally, the implications of deliberators being too easily swayed by peer information (Sunstein, 2002; Janis, 1972) or too closed to alternative views (Nickerson, 1998) is determined by tuning a parameter governing agents’ acceptance or non-acceptance of information. These parameters and experiments are described in detail in the following sections.

2.3.1 Ground truth solution space

A core assumption of this model is that policies cannot be fully understood in isolation; that they are deeply interconnected. In other words, the value of any single policy is mediated by whether or not other policies are implemented. For example, in public education, a group might consider policies targeting student testing, teacher evaluation, or free and reduced lunch programs. However, the success of any single policy may depend on the implementation – or non-implementation – of the other policies under consideration. The reliability of student testing may be influenced by the existence of a free and reduced lunch program, while the value of teacher
evaluations may be influenced by practices for student testing. Within the healthcare domain, a group may consider a mandate for health insurance separately from the implementation of a single-payer system, but the success of either policy may rely upon the implementation of the other. Similarly, a group aiming to reduce crime will not only have to grapple with questions of police officer training, patrol strategies, and incarceration policies but also consider how those policies interact with each other. For example, increasing patrols may require additional training in order to be effective.

![Policy A](Policy A) ![Policy B](Policy B) ![Policy C](Policy C) ![Policy D](Policy D)

**Figure 2.1**: Example influence patterns with $N = 4, K = 2$

In order to capture the interconnected implications of policy implementation, the model leverages the $NK$ solution space for both the ground-truth solution and individuals’ beliefs. Initially developed for adaptive evolution (Kauffman and Weinberger, 1989) and widely adopted for group problem solving (Levinthal, 1997; Lazer and Friedman, 2007; Geisendorf, 2010; Herrmann et al., 2014; Shore et al., 2015), the $NK$ model allows us to capture both positive and negative influences between policies. Each of $N$ policies can exist in one of two states: either implemented (state = 1) or not implemented (state = 0). A system with $N = 4$ policies under consideration would, therefore, have $2^N = 16$ possible policy platforms – unique combinations
of policies solutions – ranging from 0000 (no policies implemented) to 1111 (all policies implemented). The complexity of this landscape is then moderated by the parameter $K$, which indicates the number of policies whose state influences the value of implementing or not implementing a considered policy. In other words, the overall value of the policy platform 1000 differs from the value of policy platform 1001 not only by the ‘raw’ value of the 4th position policy but by the influence that policy has on the value of other policies.

Figure 2.1 shows example influence patterns for an $NK$ model in which $N = 4$ and $K = 2$. The contributing value of any given policy is determined by $K = 2$ other policies. However, this relationship is unidirectional – in Figure 2.1, for example, the contributing value of policy D is influenced by policies A and B, while policy D influences the contributing value of policies A, B, and C. Thus, some policies may have an outsized effect on the overall policy platform, while other policies may exert little influence.

Given these influence patterns, the contributing value of a single policy to the overall policy platform is then determined by the state of that policy (e.g., whether it is implemented or not implemented) as well as the states of the $K$ influencing policies. Table 2.1 shows example influence values, given the influence patterns described in Figure 2.1. For example,

<table>
<thead>
<tr>
<th>Policy</th>
<th>State</th>
<th>Influencer</th>
<th>Influencer State</th>
<th>Influence Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>-3.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>-4.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0</td>
<td>5.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>7.22</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>8.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>7.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0</td>
<td>7.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>8.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>5.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>9.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0</td>
<td>5.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>8.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>0</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>-4.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>9.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0</td>
<td>-7.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>-5.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>0</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0</td>
<td>7.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>8.03</td>
</tr>
</tbody>
</table>

Table 2.1: Example influence values for $N = 4$ policies
the contributing value of policy A is determined by the states of policies B and D. Thus, if policy A is implemented (state = 1) and policy B is not implemented (state = 0), the B \rightarrow A relation has a value of 8.87. Similarly, if policy A is implemented (state = 1) and policy D is also implemented (state = 1), the D \rightarrow A relation has a value of 7.22. Averaged together, this means that the total value of having policy A implemented in this policy platform is 8.045. Note that in this example, whether or not policy C is implemented will not affect the contributing value of policy A.

The results presented in this paper take the strength and valence of each policy interaction as a random draw from the uniform distribution \textit{unif}(-10, 10) though these results are robust across arbitrary bounds on this distribution. Negative values are assumed to be detrimental to the overall goal (e.g., weaken public education), while positive values are taken to indicate a policy is supportive of that goal (e.g., strengthen public education).

Together, these interactions assign a value to each of the \(2^N\) possible policy platforms. The contributing value of any given policy is determined by its own state (whether
it is implemented or not implemented) as well as the state of $K$ other policies. The total value of a given policy platform can then be calculated across all contributing policies. An example of this calculation is shown in Figure 2.2.

Using the $NK$ framework, the model begins by initiating ground truth parameters of influence patterns (which policies influence each other, as in the Figure 2.1 example) and influence values (the strength and valence of that influence, as in the Table 2.1 example). Together, these influence patterns and values provide ground truth measures for every possible combination of policies. For all simulations presented in this paper, we take $N = 4$ and $K = 2$. These parameters describe a relatively simple terrain which captures the inherent complexity of policy decisions while allowing for a sharper focus on the conditions under which political discussion may be productive.

Conceptually, the ground truth here reflects the true set of policies that best achieve the overarching social issue being discussed – i.e., the set of policies that best support public education or do the most to reduce crime. Note that the need to benchmark deliberative success against a ground truth solution limits the model to only considering topics where agents broadly agree on the ideal outcome. That is, agents are assumed to generally want the same things – for public education to be good or for crime to be reduced. However, these agents also hold differing normative and practical views as to which individual policies are best poised to achieve that shared goal. Agents themselves do not have direct access to this ground truth: such a solution may exist but be difficult to know or may be best considered as the outcome of some prior social process.

Indeed, this is one of the great challenges of assessing the quality of real-world political conversation: neither researchers nor deliberators typically have ground truth knowledge as to what solutions truly are best. One of the benefits of agent-based models, then, is that we can arbitrarily define the “ground truth” and define
agents’ behavior in relation to that truth. This approach assumes that there is some meaningful benchmark against which to compare agent performance, but it is agnostic as to the social or natural processes which determined that benchmark.

2.3.2 Individual beliefs

Individual agents are assumed to share an overarching goal, but begin deliberation with differing normative and practical beliefs about the best policies for achieving that goal. In the language of the NK model, all agents consider the same N policies and share the same set of ground truth interaction patterns (as in Figure 2.1). However, agents perceive different values for these policy interactions (as in Table 2.1) and thus favor different policy platforms.

These individual-level beliefs can be understood as noisy interpretations of ground truth signals. That is, while agents do not have direct access to the true, underlying influence values between policies, agents’ initial beliefs are assumed to have been shaped by information – such as scholarly research, policy briefs, or direct experience – which may contain a signal of the ground truth value. The closeness of an agent’s beliefs to the ground truth can, therefore, be interpreted as reflecting, to some degree, that agent’s cognitive capacity – e.g., their ability to synthesize an accurate view from the cacophony of information around them. While operationally, agent beliefs are generated by adding noise to the ground truth influence values, as described below, the assumption here is that agents who are good at processing complex information will hold more accurate (less noisy) beliefs. Those with limited cognitive capacity, on the other hand, will have less accurate (more noisy) beliefs.

It is worth noting that the model overall assumes some baseline cognitive capacity. That is, all agents are assumed to hold or believe they hold reasoned views. Whether an agent’s initialized beliefs are indeed reasoned judgments or simply post-hoc
rationalizations, the model assumes that agents will defend their beliefs on rational
grounds when participating in deliberation. This approach assumes that there are
some settings in which people are capable of engaging in reasoned thinking (Evans,
2003; Kahneman, 2011) and aims to understand the dynamics which lead that
reasoning to be successful or unsuccessful. This assumption does not to suggest that
well-reasoned discourse is the modal form of political engagement, merely that there
is reason to believe such discourse can and does occur (Fishkin, 2014; Knobloch
et al., 2013; Neblo et al., 2010, 2018) and it is, therefore, valuable to understand the
dynamics of these settings.

If we were to take the more pessimistic view that policy preferences were never
motivated by justifiable reasons, we could perhaps focus instead on agents’ preferred
policy platforms, using social contagion or similar models to determine which platforms
gain widespread popularity. While this may be a fruitful line of additional research,
the model here is intended to interrogate different deliberative concerns. Building
off (Lippmann, 1922) and others, the model can best be understood as examining
whether or not deliberation can be successful if citizens are too busy, distracted, or
confused to hold well-informed opinions.

Specifically, this paper examines the deliberative impacts of three canonical failures
of deliberation: limited cognitive capacity, factions, and poor reasoning regarding
the acceptance or rejection of beliefs. The effect of cognitive capacity is examined
through a noise parameter which is swept during a full simulation. Additionally,
uninformed and informed agents, described below, capture the extremes of reasoning
behavior and provide expected bounds for other types of reasoning agents. The
effect of factions, e.g., coalitions with differing views, are examined through ideologue
agents who begin deliberation with a cohort of ideological peers. The initialization of
these agents is described in more detail below. Finally, the tendency for individuals
to either acquiesce to others’ views (Sunstein, 2009; Janis, 1972) or to remain closed
to views which do not confirm their existing beliefs (Nickerson, 1998) is examined through the use of an “open-mindedness” parameter that is described in Section 2.3.3.

Uninformed agents

Uninformed agents provided a baseline condition for examining the effects of agents’ cognitive capacity. Such agents represent concerns that citizens are too inundated by information or too distracted by personal affairs to hold well-informed views (Lippmann, 1922). That is: if agents hold beliefs which are poor reflections of the true world around them, can they come to good decisions by sharing information?

Operationally, uninformed agents are initiated with a noisy version of ground truth influence values (e.g., as described in Table 2.1). These are calculated as the ground truth interaction plus a random draw from a uniform distribution determined by the noise level, a parameter which is swept over the full simulation. That is, for ground truth value $v$ and noise level $n$, an uninformed agent will hold an influence value of $v + \text{unif}(-n, n)$. Uninformed agents are, therefore, equally likely to over- or under-estimate ground truth interactions and are considered to be free of ideological skew. This mechanism captures reasoning capacity solely as a function of noise level. As the simulation sweeps the noise parameter, agents go from holding beliefs that are good reflections of the ground truth to holding beliefs that are effectively random. In other words, in low noise systems agents are assumed to have the cognitive capacity to accurately interpret signals relating to the ground truth values. In high noise systems, on the other hand, agents are assumed to be unable to form good judgments based on any signals they may be receiving.

Note that the model does assumes homogeneity across the quality of all agents’ reasoning – e.g., all agents are equally good or equally bad at estimating ground truth values. Uninformed agents are further assumed to have no pre-deliberation context; all possible beliefs are equally likely, and there are no inherent factions or coalitions.
Intellectuals

Intellectual agents are the polar extreme of uninformed agents and provide an alternative bounding condition. While uninformed agents become less accurate as noise in the system increases, intellectual agents remain closely bounded around ground truth values, independent of noise-level. These agents represent an unrealistic ideal that is reflective of concerns about the practicality of good deliberation. That is, these agents capture the argument that deliberation is too idealistic and can only work if all agents can develop accurate and well-informed views (Lippmann, 1922). These intellectual agents, therefore, provide both an estimation of ideal behavior and allow us to examine whether technocrats or other highly educated elites, can, or are need to, successfully moderate the inaccurate beliefs of other citizens.

Operationally, as the simulation sweeps the noise parameter, \( n \), an intellectual’s influence value will remain \( v + \text{unif}(-1, 1) \), for ground truth value \( v \). Thus, these agents can be considered to have high cognitive capacity and will always hold well-informed views.

Ideologues

Finally, ideologues are initialized as cohorts with intentionally skewed views. These views are equally likely to be any given absolute distance from the ground truth, but are biased in terms of the direction of that distance, e.g., either positively or negatively skewed. The result is that ideologues find themselves in polarized systems, entering a conversation with a natural cohort of peers who hold similar beliefs while simultaneously facing an opposing coalition of peers with dissimilar beliefs. Such agents represent fears that divided groups will remain entrenched in their respective positions and will, therefore, be unable to reach good decisions (Dworkin, 2006; Madison, 1787). Note that this approach still assumes homogeneity across reasoning capacity, as all ideologue agents are initialized with the same absolute skew. In other words, all ideologues are wrong; the resulting coalitions are simply wrong in different
ways.

Operationally, for ground truth value $v$ and noise level $n$, ideologues will be initialized as either \((v + (\text{unif}(n - 2, n) \times \text{sign}(v)))\) or as \((v + (\text{unif}(-n, -n + 2) \times \text{sign}(v)))\). Including the sign of $v$ in this calculation allows for non-parametric interactions that appropriately complicate agents’ selection of optimal policy platforms. Without this term, these agents could be interpreted simply as optimists (who think all interactions are good) and pessimists (who think all interactions are bad). A consensus solution in such a system would then most frequently reflect pessimists settling for the “least bad” solution rather than advocating for a policy platform they genuinely interpret as good.

### 2.3.3 Reasoned exchange

The final element of the model is a process through which agents share beliefs and assess whether or not to incorporate the beliefs of others. In this game of “giving and asking for reasons” (Neblo, 2015), good faith discussants should seek to “convince others and be convinced when appropriate” (Mercier and Landemore, 2012) – that is, they should incorporate others’ beliefs when those beliefs seem credible and reject others’ beliefs when they do not. However, people are highly prone to social bias (Festinger, 1954), and frequently go along with perceived popular views (Sunstein, 2002; Janis, 1972). Furthermore, people are also prone to confirmation bias (Nickerson, 1998) and, rather than go along with the group, may only accept views that already conform to their existing beliefs. While the deliberative ideal imagines participants aiming to carefully consider all viewpoints (Manin, 2005; Mercier and Landemore, 2012), it is unclear whether or not deliberation can succeed given these common biases. In order to examine this final deliberative failing, then, the model estimates the effect of these biases through an open-mindedness parameter which governs whether an
When sharing an opinion, an agent is more specifically sharing their personal influence values for a single policy in a single state. That is, a speaker can be interpreted as sharing a policy preference (the state of a single policy) and justifying that preference based on the influence of other policies under consideration. For example, a speaker might argue that a policy of standardized student testing should only be implemented (state = 1) if a free or reduced lunch program is also implemented. E.g., an implemented lunch program would have a positive effect on the value of student testing, while not implementing a lunch program would have a negative effect on the value of implementing student testing.

Since the value of each policy is determined by the state of $K$ other policies, this means that each agent associates $2^K$ values with a given policy in a given state. Furthermore, because we assume that all agents share the same set of underlying influence patterns, this provides listening agents a direct point of comparison. That is, speaking agents and listening agents agree on which policies are related to each other, but they hold differing evaluations of those connections. The speaking agent can, therefore, share their specific valuation of those connections while the listening agent can compare to their own evaluation of those relations. Using the example influence values from Table 2.1, a speaker might share beliefs about the value of not implementing policy A (state = 0) as the vector $[3.39, -3.16, -4.97, 3.08]$, which any listening agent would understand as proposed influence values for $[(B \text{ in state } 0), (B \text{ in state } 1), (D \text{ in state } 0), (D \text{ in state } 1)]$ respectively.

In other words, this provides a $2^K$ dimensional space within which agents can assess each other’s views. Agents can then choose to accept or reject a shared view within the context of their own beliefs. If an agent is too open to accepting shared beliefs, they are more likely to accept poor information and ultimately degrade the decision.
reached by the group (Sunstein, 2002; Janis, 1972). On the other hand, if agents prefer not to accept shared views, they fall into the trap of confirmation bias (Nickerson, 1998) and may fail to abandon their own false views.

Operationally, at every time step, a single speaking agent shares their length-$2^K$ vector of influence values for a given policy in a given state. Listening agents then compare that to their own length $2^K$ vector and move towards the speaker’s values if the cosine similarity between the two vectors is within a range determined by agents’ tunable open-mindedness parameter, described below. If an agent “moves towards” a shared view, they adopt a new belief vector which lies halfway between their old view and the shared view. If they do not accept a shared view, they maintain their original belief vector.

This paper considers three types of agent acceptance. Open agents are prone to failures of group polarization and groupthink (Sunstein, 2002; Janis, 1972) and are very likely to incorporate others’ beliefs. Skeptics are prone to the failure of confirmation bias (Nickerson, 1998) and only accept views that already conform to their own. Finally, moderates aim to accept true beliefs and abandon false beliefs (Mercier and Landemore, 2012). Operationally, since cosine similarity ranges from -1 to 1, open agents move towards a shared belief if the cosine similarity is greater than -.75. Moderate agents move towards an opinion if the cosine similarity is greater than 0, and finally, skeptical agents will only move towards beliefs if the cosine similarity is greater than +.75.

### 2.3.4 Experiments and measures

This paper examines the effects of cognitive capacity, factions, and acceptance of beliefs on deliberative outcomes. Particular focus is given to the effect of human cognitive capacity, which is one of the most serious concerns raised about the practicality of
deliberative success (Lippmann, 1922; Shah and Oppenheimer, 2008; Oswald and Grosjean, 2004; Tversky and Kahneman, 1981). This capacity is broadly captured in the “noise level” of initialized beliefs, with low-noise systems indicating high cognitive capacity and high-noise systems indicating low cognitive capacity. This noise parameter is swept from 0 - 200 during a simulation while ground truth values are always initialized between -10 and 10. Thus as we tune this parameter, we move from agents whose high cognitive capacity gives them beliefs reasonably close to the ground truth to agents whose low cognitive capacity results in beliefs that are essentially random. Agents are initialized to be either uninformed, intellectual, or ideologue, following the rules above. All agents experience the same noise level, though intellectuals, who are illustrative of high cognitive capacity, remain unaffected by the noise level of the system.

This paper tests the model particularly for small group discussions, i.e., groups between 5 - 25 agents and finds similar results across group size (see Appendix 6.1). Furthermore, the paper examines several possible group compositions. Groups of entirely uninformed agents and entirely intellectual agents provide baselines against which to compare behavior. The uninformed group, in particular, captures the effect of diminishing cognitive capacity as the noise level increases. In order to capture faction dynamics, groups composed entirely of ideologues are evenly split between agents with positive and negative skew. In groups with an odd number of agents, the positive-skew coalition is arbitrarily a single agent larger.

Additionally, to examine the possible moderating effects of intellectual agents, the paper considers groups that have a majority of intellectuals as well as groups that have a majority of ideologues. In majority intellectual groups, half of the agents are intellectuals, rounding up in systems with odd numbers of agents. The remaining agents are split evenly between ideologues with positive and negative skew. In majority ideologue groups, agents are split evenly between ideologues with positive
and negative skew, with a single agent (odd systems) or two agents (even systems) as intellectuals. The model can further be tested with arbitrary mixes of ideologue and intellectual agents, but the results presented here are illustrative of any group composition (see Appendix 6.1).

Each simulation begins by first initializing the ground truth and then determining agents’ beliefs. Once initialized, at each timestep, a randomly selected speaker shares a vector of influence values for a random policy in a randomly selected state. This can be considered as a single speech act, with the speaker making a claim about the benefit of implementing a given policy, and then justifying that claim with the influence values which inform that view. Listeners then compare these stated reasons with their existing beliefs, choosing to either move towards the shared values or maintain their existing vector of values, based on the open-mindedness level of the simulation. For simplicity, this paper only considers simulations in which all agents have the same open-mindedness rules for incorporating a speaker’s beliefs. However, the paper examines three distinct levels of open-mindedness. Open agents represent concerns around groupthink and group polarization (Sunstein, 2002; Janis, 1972; Festinger, 1954), and are likely to accept other’s views. Skeptical agents capture issues of confirmation bias (Nickerson, 1998) and only accept views which are already close to their own. Finally, moderate agents fall in between and aim to accept true beliefs and reject false beliefs (Mercier and Landemore, 2012). The paper considers this open-minded parameter for groups of ideologues, though similar behavior for uninformed and intellectual agents can be found in Appendix 6.1.

After each speech act (e.g., timestep), the model then assesses the outcome of deliberation by determining how many “good” policies would be implemented if agents held a simultaneous vote on each policy. Each agent selects their preferred policy platform, given the positive and negative trade-offs the agent sees between different policies. Every policy is then given a simultaneous up or down vote, with
agents voting for the policies indicated by their preferred platform. A policy that receives a simple majority of votes is then considered implemented, resulting in a platform of enacted policies. Note that each policy is given an independent vote with only two options (implement or do not implement), and therefore, no Condorcet-paradox issues arise from this process. An example of this voting procedure can be seen in Figure 2.3. The true value of the enacted policy platform can then be determined in the ground truth solution space.

Note that agents do not necessarily have to agree with each other’s reasoning or even have the right reasoning in order to come to good decisions. Rather, each agent maintains their own reasons for favoring or disfavoring individual policies, and agents may ultimately find that they agree on some policy implementations while disagreeing on why they hold those preferences. This process of agents sharing and considering reasons continues until the optimal policy platform is enacted or after 5000 timesteps. Each simulation is run 100 times at each noise level, with the ground truth initiated only once across all runs at a given noise value. Since this creates some spurious oscillation in deliberative outcomes by noise level, we examine the effects of noise over a rolling window of width 10.

This procedure gives three metrics by which we can assess deliberative outcomes: the

![Figure 2.3: Example policy vote](image-url)
percentage of good policies enacted, the distance of the implemented policy platform from the optimal platform, and the percentage of agents in the largest coalition. For the first measure, a policy is considered “good” if its state matches the optimal state. Thus, this value will be 1 if the enacted policies match the optimal policy platform. In the second measure, we use the ground truth influence values to determine the true value of the enacted policy platform and measure the absolute negative distance of this value from the value that would have been obtained by enacting the optimal policy platform. This number is 0 if the group identified the optimal solution and negative otherwise. Finally, we look at the percentage of agents who share the largest policy platform coalition. That is, agents are considered to share a coalition if they support the exact same policy platform – i.e., vote for and against the same policies. A value of 1 here indicates that agents have reached full consensus. Note, this does not necessarily mean that agents have identified the true optimal policy platform, merely that they have come to an agreement as to which platform to enact.

These metrics can be evaluated after any timestep in a single simulation. In order to highlight trends in deliberative outcomes, however, we focus here on the final state of simulations from a given noise level. For each noise level, we run 100 simulations that each last until the optimal policy platform is enacted or after 5000 timesteps. We then record the percent of good policies enacted at the end of the simulation, the distance of the final policy from optimal, and the size of the largest coalition. We consider the average of these values along with a bootstrapped 95% confidence interval.
2.4 Results

2.4.1 Effects of cognitive capacity and ideological skew

We first examine how different types of reasoners perform along various deliberative metrics. This includes uninformed agents, whose beliefs are evenly distributed around the ground-truth, ideologues, whose beliefs are heavily skewed either positively or negatively from the ground truth, and intellectuals, whose beliefs are closely tied to the ground-truth independent of noise value.

As we can see in Figure 2.4, groups comprised entirely of uninformed agents perform increasingly poorly as the noise level increases. This is expected, as the noise level indicates these agents are initialized with beliefs further and further from the ground truth. Since the ground truth influence values are initialized to be between -10 and 10, at even modest levels of noise, the beliefs of these uninformed agents are essentially random. Since intellectual agents are initiated independently of the noise level, we similarly see that such agents perform consistently well regardless of noise level. While in principle, these findings reinforce concerns about the influence of cognitive capacity on deliberative outcomes (Lippmann, 1922; Tversky and Kahneman, 1981), both uninformed and intellectual agents were designed to serve primarily as bounding conditions representing extremes of poor and good cognitive capacity. In other words, while these findings do confirm that ideal, “intellectual” agents would come to good solutions when deliberating, the failure of uninformed agents should not necessarily be interpreted as failure of real-world citizens. Interestingly, factioned groups entirely of ideologues perform reasonably well at all noise levels, suggesting that the process of “giving and asking for reasons” can, in fact, help groups identify good solutions even when their initial information is poor. Because ideologues are intentionally skewed away from the ground truth, but in differing directions, their competing presence appears to bring some balance to the overall group, allowing
some amount of the ground truth signal to remain. This finding suggests that even if humans are inherently poor at reasoning, it may actually be beneficial for our collective biases to be non-randomly distributed. While this finding does not alleviate all fears that factions will be unable to collaborate successfully (Dworkin, 2006), it does ameliorate this concern for settings in which people are genuinely willing to consider other views. In other words, this finding supports core deliberative arguments, illustrating that even groups whose members are biased and cognitively

Figure 2.4: Deliberative outcomes by reasoning type.
flawed may be able to come to good solutions by genuinely trying to learn from and convince each other.

We more closely examine the influence of intellectuals on deliberative outcomes in Figure 2.5. Here, we consider both groups which have a majority of intellectuals as well as groups which have a majority of ideologues. In majority intellectual groups, half of the agents are intellectuals, and the other half are split evenly between ideologues with positive and negative skew. In majority ideologue groups, only 1-2 agents are intellectuals while the bulk of agents are split in ideological skew. Additional mixtures of intellectual and ideologue agents can be found in Appendix 6.1.

Figure 2.5: Deliberative outcomes portion of intellectuals.
Overall, we find little evidence to suggest that intellectuals serve to improve deliberative outcomes. While a group comprised entirely of intellectuals quickly identifies optimal solutions – as shown in Figure 2.4 – their presence does not seem to enhance the findings of ideologue agents. One reason for this may be found in the group’s coalition behavior. As we would expect, groups entirely of ideologues form a 50% voting bloc, representing the polarization between the positively and negatively skewed ideologue groups. When a large number of intellectuals are added to the mix, the size of the largest coalition increases significantly – here, holding steady at around 80%. This finding suggests that while these groups converge to near-consensus, they do not often converge on the optimal solution. This could suggest that even a large number of intellectuals are not enough to moderate polarized spaces, or it could be an effect of the model – similar to the finding of Lazer and Friedman (2007) that efficient systems tend to converge too quickly on sub-optimal solutions. While similar results have been found in other agent-based models (March, 1991), this result has not been replicated in human experiments (Mason and Watts, 2012), where moderating behavior may emerge differently than in simulated systems.

2.4.2 Effect of cognitive capacity and open-mindedness

Finally, we examine the effects of agents’ willingness to accept others’ views, testing whether issues of group polarization (Sunstein, 2002) or confirmation bias (Nickerson, 1998) lead groups to reach worse deliberative outcomes. In Figure 2.6, we compare outcomes among groups of ideologues who are skeptical (only accept if the vectors have a cosine similarity greater than .75), who are moderate (cosine must be positive) and who are open-minded (will accept if the vectors have a cosine similarity greater than -.75). Here, we see virtually no differences between the three groups. Indeed, skeptical and open agents appear to perform equally and, if anything, slightly outperform their moderate peers. More simulations would be needed to determine whether or not this
performance is a significant effect, but it may reflect a tendency for moderates to accept just the wrong amount of information – i.e., they are convinced to change their views, but frequently change their views to wrong things.

![Graphs](image)

Figure 2.6: Deliberative outcomes by open-mindedness.

2.5 Discussion

Processes of group decision making are often complicated by normative perspectives. Indeed, a key challenge of democratic life is that – even when given access to the same information – people frequently do not agree as to the best course of action. Yet, such normative settings are common in the real world and can be very high-stakes. This paper focuses on one particularly salient setting: examining group decision-making processes around policy implementation. The model imagines a small group of deliberators considering a set of possible policies all aimed at addressing some overarching social issue such as education, healthcare, or crime. Agents will ultimately vote on which policies to implement but are given the opportunity to share information, reasons, and justification before they do.

While there are good reasons to be skeptical that such a deliberative process could meaningfully influence a vote in the real world, empirical evidence suggests that
such influence is possible (Fishkin, 2014; Knobloch et al., 2013; Neblo et al., 2010). Furthermore, even if the deliberative ideal of reasoned exchange (Mansbridge, 2015) is not frequently met, the real-world implications and repercussions of such decision-making processes obligates us to try to understand the dynamics of these processes and to assess the conditions under which they lead to better outcomes.

Agent-based models provide a promising approach to exploring the dynamics of group problem-solving in normative settings. While necessarily highly stylized, such models allow for tunable control over parameters which could never be manipulated in real life. Furthermore, in breaking complex processes down into component parts, these models help us think differently about macroscopic phenomena – probing the constituent elements and gaining insight into the role that even simple mechanisms can play. Previous agent-based models of belief systems, for example, have provided insight into processes of group consensus and dissensus (DeGroot, 1974; Friedkin et al., 2016; Altafini, 2013).

Specifically, this paper examines three canonical deliberative failures: limited cognitive capacity, group factions, and poor judgment for accepting or rejecting views. Each dimension captures some element of real-world deliberative challenges. People may be less likely to come to good decisions if humans have limited cognitive capacity, if they are too polarized to engage with people unlike themselves, or if they are too set – or too flexible – in their opinions. This paper examines these effects through groups comprised of uninformed, intellectual, or ideologue agents who operate with differing rules for adopting others’ beliefs. The model imagines small groups of deliberators attempting to engage in good-faith reasoning about topics on which they hold fundamentally differing views. Using the $NK$ model to operationalize a complex policy landscape, agents engage in a game of giving and asking for reasons – exchanging policy preferences as well as justification for those preferences.
The key finding is that polarized groups do surprisingly well at identifying optimal policy solutions. While uninformed agents perform as expected – achieving worse outcomes as cognitive capacity decreases – groups composed of oppositely-skewed ideologues appear to be resilient to the effects of declining cognition. This finding is in line with deliberative theory (Mansbridge, 1999) and suggests that heterogeneous agents can achieve good outcomes if they are able to engage in good-faith discussions.

This, of course, is a very restrictive assumption – very often, people do not engage in good-faith exchange and are more concerned with winning than in collaboratively discovering the truth. In such settings, group factions would likely lead to entrenchment rather than optimal outcomes. What this finding suggests, though, is that the presence of opposing factions may not itself be the biggest concern. Indeed, having access to a diversity of opinions has the potential to lead to better outcomes than might be achieved otherwise. This benefit can only be realized, however, if we are able to build systems, institutions, and interventions which help people truly listen to and learn from each other.

Notably, this ideological effect does not appear to be moderated by the other parameters of interest. That is, while ideologues do far better than uninformed agents, their performance is not significantly improved by the presence of highly cognitive intellectuals or by being increasingly open to, or skeptical of, others’ views. This suggests that the main factor in driving ideologues’ performance is their inherent counter-balance to each other; by constantly pushing the “other side” to be better, both groups, and ultimately the whole, can improve.

It is particularly interesting to note that the conditions under which agents accept each other’s views have little effect on observed outcomes. While we would see entrenchment if agents were wholly unwilling to consider opposing views, even a modest openness to others’ opinions can lead to better outcomes.
Taken together, these results suggest a need to further invest in studying and creating deliberative spaces. The deliberative ideal is indeed lofty – and we should not expect that every person in every conversation will fulfill the ideals of good-faith reasoned exchange (Mansbridge, 1999). But these results suggest that we do not have to. A world in which discussion and debate lead to better policy outcomes does not have to be perfect – it does not have to rely on highly cognitive people wholly free of partisan bias. It is okay for people to be imperfect and to be flawed in their thinking – but only, only, if they are modestly willing to listen to each other.
Chapter 3

The Structure of Reasoning: Measuring Justification and Preferences in Text

3.1 Introduction

The individually distinctive ways in which people express their preferences holds the potential to reveal broader variations in political behavior. A population’s agreement on a given policy position, for example, may elide deeper divisions in the motivation behind that position. Similarly, discussants with opposing preferences may, under some circumstances, be able to find ways to engage productively across their differences. While the bulk of public opinion literature has rightfully focused on the output of what people believe – i.e., individuals’ discrete policy preferences – a robust

\[^1\text{Replication materials for this chapter can be found at https://github.com/sshugars/conceptual_networks.}\]

\[^2\text{This chapter includes work coauthored with Dr. Nick Beauchamp and Dr. Peter Levine.}\]
understanding of the full deliberative system additionally requires an analysis of how people express these preferences and interpret the preferences of others. That is, while political preference models are invaluable for capturing trends in public opinion and predicting policy outcomes, they are not designed to analyze interactions between individuals’ preferences. What arguments can lead to opinion change, and under what circumstances? What factors drive a political conversation to be productive or divisive? How can a society function democratically in the face of increasing levels of affective polarization? If we hope to answer such critical questions of public opinion, we need individual-level models of political reasoning and expression.

The call for such models is not new, but the computational tools needed to develop them are. A notable line of classic public opinion research (Lane, 1962; Axelrod, 1976; Campbell, 1960) used semi-structured interviews or hand-coding of texts in order to capture individual variation in the articulation of political preferences. From a normative standpoint, studying this individual variation acknowledges democratic ideals of citizen voice and opens the door for examining strategies of moving towards this ideal. From a practical standpoint, this variation holds the potential to reflect the success and failures of elite messaging: even if average citizens primarily repeat elite talking points, it is still worth examining which talking points they find themselves repeating. While early efforts at modeling individual-level reasoning were often abandoned as too arduous and time-consuming, modern computational methods hold the promise to meaningfully revive these efforts.

This paper presents an initial step in such a revival and seeks to demonstrate that conceptual network structures (1) can be computationally inferred from text and (2) meaningfully reflect individuals’ behavioral and personality traits. This work is largely exploratory, recovering a lost tradition and shedding insight into conflicting behavioral priors. The core argument is that individuals talk about politics in subtly different and unique ways. Examining the structure of expressed reasoning, separate from
the content of those reasons, can, therefore, shed insight into variations in political behavior. This, of course, is not to argue that the content of political preferences are not themselves deeply important, merely that the structure of expressed reasons is an under-studied and meaningful dimension of political behavior as well.

In pursuit of these joint goals, this paper presents a computational, text-based approach for inferring conceptual networks from text. Through the application of this method to two distinct datasets, the paper then demonstrates the potential for these network structures to generate behavioral insights. The method proposed here, described in Section 3.4, leverages the grammatical structure of a text in order to infer the implicitly encoded connections between mentioned concepts. Furthermore, the method uses embeddings (Mikolov et al., 2013) to identify words within a text which point to the same concept.

This method is applied to two datasets, which are described in detail in Section 3.3. The first is an original dataset of 100 subjects recruited through Amazon’s Mechanical Turk. Subjects responded to two political prompts and completed an extensive survey battery, estimating Big 5 personality traits (John and Srivastava, 1999), Moral Foundations’ tendencies (Haidt and Joseph, 2008), and other politically-relevant measures (Pew Research Center, 2017; Knobloch et al., 2013; Carpini and Keeter, 1993). These personality traits are known to be correlated with political ideology (McCrae and Costa Jr, 1999; Haidt, 2012) and therefore bolster external validation for the method while simultaneously providing behavioral insights of their own. These measures and their hypothesized correlations are discussed further in Section 3.3. The second dataset is a sample of nearly 1,000 respondents recruited through YouGov for a survey designed by Daniel Hopkins (University of Pennsylvania) and Hans Noel (Georgetown University)\(^3\). In an “ideological Turing test,” subjects

\(^{3}\)I would especially like to thank Dr. Hopkins and Dr. Noel for generously sharing their data for this analysis.
generated texts for both the “liberal” and “conservative” positions on a single issue. This dataset, therefore, allows for a disaggregated examination of ideological and individual correlates beyond what is possible with the Mechanical Turk data. Furthermore, since nearly half of respondents (44%) provided ironic answers which did not meaningfully reflect a given ideological position, this dataset additionally allows for a rough examination of argument quality, e.g., of the extent to which a text reflects a reasonable representation of a given view.

After analyzing both sets of data, this paper finds that individuals do appear to structure their political expressions in individually distinctive ways that are indicative of a personal style beyond mere ideological talking points. This “reasoning fingerprint” conveys information about argument quality and holds the potential to provide new behavioral insights, especially in regards to deliberation and conversation quality. This suggests that computational approaches for inferring conceptual network structure hold the potential to be a fruitful line of research for public opinion and political behavior. If we hope to understand the deliberative system or design deliberative interventions, we must not only measure what people say but also how they say it.

### 3.2 Related Work

Cognitive processes and linguistic expression are both known to be structured phenomena (Quillian, 1967; Shavelson, 1974; Walton, 1996; Toulmin, 1958). Studies of reasoning (Axelrod, 1976; Carley, 1993; Toulmin, 1958), arguing (Toulmin, 1958; Walton, 1996), remembering (Collins and Loftus, 1975; Quillian, 1967), and learning (Shaffer et al., 2009; Shavelson, 1974) all suggest that individuals express and interpret beliefs in structured ways.

Specifically, these processes are best understood as having a network structure: people
store and retrieve information not as isolated packets of ideas, but as complex networks of interconnected concepts. When speaking with others, we raise ideas that seem related to what they said; when thinking to ourselves, we move from idea to idea via their connections; and when assessing a complex issue, we weigh the pros and cons as well as their interconnections in order to arrive at a final judgment. Network interpretations of the cognitive organization of knowledge are bolstered by behavioral observation of arguments, deliberation, written texts, and self-reports that repeatedly suggest that individuals perceive their ideas to be connected to each other in complex networks of support or contradiction.

Furthermore, cognitive and linguistic processes are inexorably linked: the conceptual networks which cognitively store information (Collins and Loftus, 1975; Dorsey et al., 1999; Quillian, 1967) cannot be directly observed and must be inferred primarily through language. This inference process has generally proceeded from two directions: a psychological approach that begins with theories of cognition and attempts to recover these structures through experimentation, observed behavior, and collaborative knowledge-building; and a linguistic approach that seeks to explain semantic patterns, meanings, and grammars using network structure. These two strains of study often converge on similar types of models, though they reflect the varied disciplines targeting this shared problem. Additionally, work in moral philosophy has aimed to normatively assess individual conceptual network structure, leaving aside issues of measuring that structure. Finally, popular behavioral approaches focus exclusively on clusters of latent traits as drivers of behavior, neglecting any network structure. This paper builds upon all these literatures, seeking to develop and validate an integrated approach for understanding individual-level conceptual network structure that can bring new insight to behavioral understandings.

Psychological models argue that human memory search is made possible by storing information as a network in which concepts, represented as nodes, are connected by
relational links to other conceptual nodes (Collins and Loftus, 1975; Quillian, 1967). In the theory of semantic memory developed by Quillian (1967), for example, a node provides a shallow understanding of a given concept and is represented by a single word or phrase. A “concept” more deeply considered, then, contains indefinitely large amounts of information and is properly expressed as the entire network accessible from a given concept node (Collins and Loftus, 1975). Such a knowledge structure allows a person to store a concept as a compressed object (node) while simultaneously allowing access to a richer understanding through the network structure (Quillian, 1967).

These psychological theories have been applied in a range of settings. Semantic network libraries such as BabelNet (Navigli and Ponzetto, 2012), ConceptNet (Speer and Havasi, 2012), and SNePS (Shapiro and Rapaport, 1987) rely on the core psychological intuition that a concept, encoded as a word, can be best described through its associated concepts, which themselves are encoded as words. Education scholars have similarly leveraged psychological theories to argue that knowledge itself has a network structure and that “learning” can, therefore, be considered as a process of developing the right knowledge structures. In other words, the skill of applying existing knowledge to new situations relies upon developing an understanding of how relevant information is interconnected (Dorsey et al., 1999; Hong et al., 2004; Shaffer et al., 2009; Shavelson, 1974). Social scientists have further argued that conceptual networks can be used to examine how individuals reason and make choices between alternatives (Axelrod, 1976; Carley, 1993). In weighing possible outcomes, a person evaluates connected concepts and consequences, exploring paths within their conceptual network in order to determine the optimal choice. Political deliberation provides a natural venue to extend such models, as participants may enter conversation with differing views and must, therefore, attempt to share structured knowledge before reaching a decision.
Notably, the exchange of knowledge is most frequently done through language, leading to a separate stream of work engaging the structure of language as a proxy for the structure of knowledge. Perhaps the most well developed such models trace their roots back to Aristotelian efforts to define the structure of argumentation (Toulmin, 1958). Such structures may be relatively simple: a major premise connected to a minor premise leads inevitably to a logical conclusion; or it may be significantly more complex, such as in the two dozen schemes described by Walton (1996) or the Context Free Grammar introduced by Mochales and Moens (2011). However, while theorists have differed in the specifics of the models they put forth, their approaches all begin with implicit acceptance of the network structure of arguments: the soundness of a conclusion rests not only upon its supporting ideas but how those supporting ideas are connected. In other words, arguments fundamentally have a coherent structure expressed through linguistic structure and defined by evidence relationships (Cohen, 1987). The search for these structures has given rise to a rich body of research known as argument mining, in which supervised and semi-supervised computational methods automate the search for the sorts of argument structures articulated by Aristotle or Toulmin (Mochales and Moens, 2011). The conceptual networks inferred via these methods tend to be more structured and hierarchical than those inferred from open-ended psychological approaches, but the basic structure of nodes and edges representing ideas and their interconnections remains.

While psychological and linguistic approaches aim to infer and examine conceptual network structure, an important line of work in philosophy has developed normative theories regarding the properties of these networks. These theories rely primarily on principles of coherence, considering a moral position valid insofar as it is coherent with other views (Christen and Ott, 2013; Dorsey, 2006; Rawls, 1993). What constitutes “coherence,” however, differs between philosophers, leading to differing topological interpretations. In Henry Sidgwick’s influential version of utilitarianism, for instance,
“the current moral rules” such as “do not lie” are used to generate most of our actual judgments (Sidgwick, 1907), leading to topologies in which some ideas serve as central gatekeepers. In particularist moral theories, by contrast, each moral judgment is only linked to others by loose and local analogies (Dancy, 1993), implying that no ideas should enjoy disproportionate centrality in a person’s network of moral ideas. McNaughton and Rawling (2000) argue that this is the flaw of particularism because some concepts really are “central” to morality. This argument suggests a hybrid approach in which core ideas are central but do not dominate the reasoning structure.

These varied definitions of “coherence” share an understanding that consistency between individual pairs of beliefs is too low of a standard for judging the validity of a moral position. On the one hand, individual beliefs may be consistent but unrelated, while on the other hand, expecting all pairs of beliefs to be directly connected is too stringent a standard since moral views range over a wide variety of topics. Several scholars have, therefore, explicitly argued for whole network approaches to coherence. Thagard (1998) proposes a theory involving literal network relations, though he overlooks many of the relevant formal features of networks. Berker (2015) posits that an individual’s beliefs should be modeled as a network to reveal its degree of coherence and begins to explore the variety of forms that a network of moral values can take.

Given the broad literatures which embrace a network understanding of human reasoning, the work here seeks to enrich existing behavioral theories of public opinion. Recent work in public opinion has examined the structure of preferences themselves but has shied away from examining the reasoning structure behind those preferences.

Finally, while this work’s focus on the expression of political reasoning runs parallel to Zaller et al. (1992)’s examination of survey response, Zaller provides a helpful framework through which to interpret the reasoning and articulation process. Zaller
et al. (1992) argue that survey responses can be modeled as a process of constrained stochastic sampling. Individuals receive information through external signals, selectively accept information that conforms to prior beliefs and then sample from those available beliefs to generate an ideal-point estimation of their preference on the fly. This process is stochastic and will result in a single individual giving varied responses over time. However, it is also heavily constrained – a subject may exhibit variability in how extreme their stated preference is, for example, but is unlikely to spontaneously flip from one end of the political spectrum to the other.

While Zaller does not consider the structure of political reasoning in his work, it is a natural extension to consider a similar process in this space. We similarly imagine that people receive and selectively accept external information. This accepted information is then stored as a latent conceptual network and represents the ideas and connections one has at their disposal. When expressing reasoning, individuals then sample from this latent network in determining the precise topics they raise.

### 3.3 Data

This study uses two distinct datasets in which subjects were asked to explain political views. The first is an original dataset of 100 subjects recruited through Amazon’s Mechanical Turk. The second is a sample of 873 respondents recruited through YouGov for a survey designed by Daniel Hopkins (University of Pennsylvania) and Hans Noel (Georgetown University). Each dataset provides different insights into the validity, usefulness, and meaning of conceptual networks as a tool for understanding political behavior.
3.3.1 Amazon’s Mechanical Turk

Initially collected for a related experiment to test the broader validity of conceptual network models (Shugars et al.), subjects in the Mechanical Turk study completed three different network elicitation mechanisms ordered at random. One activity was a simple free-response text box, while the others were specially-developed, web-based tools that allowed subjects to generate their own networks. These last two methods – an interactive network drawing program and a simulated conversation via chatbot – were inspired by previous work that engaged subjects in defining their own networks by connecting, and in some cases generating, relevant keywords (Shavelson, 1974). Subjects were randomly assigned two prompts from a pool covering issues of abortion, healthcare, and childrearing, and completed all three activities before progressing to their second assigned prompt. For each subject, we, therefore, have multiple inferred networks, spanning different issue areas and elicitation methods. Text-based responses were typically close to the imposed minimum of 100 words in length. Additionally, subjects completed an extensive survey battery covering Big 5 personality traits (John and Srivastava, 1999), Moral Foundations (Haidt and Joseph, 2008), political knowledge (Carpini and Keeter, 1993), openness to deliberation Knobloch et al. (2013), and political ideology (Pew Research Center, 2017).

By measuring both conceptual network structure and individuals’ personality traits, this dataset provides external validation for the method and illustrates the potential for substantive behavioral insights. That is, if we believe that the structure of a subject’s response is connected in some way to their personal style, we would expect to see a number of correlations between individuals’ inferred network structure and measured personality traits. Further, if we believe that such correlations are more than spurious, we would expect those correlations to cluster in meaningful ways – e.g., traits that are correlated in other settings (McCrae and Costa Jr, 1999; Haidt, 2012) should be suggestive of similar network structures here. While the specific measures of
network structure are described in detail in Section 3.4.3, their expected correlations with the specific personality traits measured in this experiment are described below. At the highest level, we would expect to see clustering between personality traits associated with liberal ideology and traits associated with conservative ideology. Moral foundations theory (Haidt and Joseph, 2008; Haidt, 2012) argues that political opinions are driven by an individual’s orientation along at least five moral dimensions, and suggests that political divides can be traced back to fundamental differences in the weighting of these moral dimensions. This would suggest that there is a “conservative” way of thinking and, separately, a “liberal” way of thinking, each of which should be reflected in the inferred network topology. That is, we would expect to see distinctive conservative network structures that meaningfully differ from liberal network structures. Specifically, we expect that respondents who score high on the moral foundation measures of purity and authority – which are associated with conservative thought (Haidt, 2012) – would produce network structures similar to those who are ideologically conservative (Pew Research Center, 2017). We would further expect this conservative thought to be reflected in heterogeneous networks with disassortative connectivity. These characteristics would indicate more hub-and-spoke like networks, where concepts differ significantly in importance (as measured by degree) and high-degree nodes tend to connect to low-degree nodes. Such networks are representative of the utilitarian view (Sidgwick, 1907), where a few core rules dictate judgments. On the other hand, we would expect respondents who score high on the moral foundations’ composite score of progressivism – characterized by the traditionally liberal traits of aversion for harm and concerns about fairness – to have networks whose structure is more in line with particularist moral theories (Dancy, 1993). Such structures would be notably interconnected, allowing for flexible, context-aware moral reasoning. This would be reflected by networks with a higher average degree and a single connected component. These networks may be more
homogenous, indicating the lack of any dominating, central ideas, or they could have densely connected cores surrounded by a sparse periphery – in which case, we would still see high standard deviation and disassortativity. These specific measures and their implications for different conceptions of moral “coherence” are discussed further in Section 3.4.3.

At a more granular level, we hypothesize that respondents with higher political knowledge (Carpini and Keeter, 1993) will produce more interconnected networks, resulting in a single component and a higher average degree. While this could potentially lead to denser networks as well, we would also expect knowledgeable subjects to produce more content overall, which might ameliorate that effect and potentially even lead to sparser networks. Furthermore, we would also expect to see similar effects when people are responding to particularly salient issues on which they have more intense or more established positions.

Respondents who are more open to deliberation (Knobloch et al., 2013) may be more likely to produce complex networks that match Dancy (1993)’s conception of coherence. These subjects would have flexible, interconnected networks which have a single component and a wide range of degrees. Again, these users could have denser, more highly connected networks, or sparser networks, if they have more content but the same number of connections on average.

Additionally, we expect that people who score higher on the Big 5 (John and Srivastava, 1999) measures of conscientiousness and neuroticism will have more connected networks which are more homogeneous. The first of these personality traits, conscientiousness, is marked by high organizational skills and a strong sense of purpose (McCrae and Costa Jr, 1999). Individuals who score high in this dimension may, therefore, be more likely to be intentional in connecting the ideas they raise, resulting in connected networks where each concept enjoys equal footing. Similarly, the
dimension of neuroticism is tied to perfectionist tendencies (McCrae and Costa Jr, 1999), which might manifest in balanced, connected networks.

Additionally, individuals who score high on the Big 5 measure of openness may be more likely to produce disconnected networks with a disassortative (degree heterogeneous) structure. Openness is characterized by diverse knowledge and interests (McCrae and Costa Jr, 1999), suggesting these users are likely to have lots of ideas, and overall more content, but may not see all these ideas as connected and may take some to be more central than others. Furthermore, the Big 5 trait of agreeableness is marked by tendencies for compliance and cooperation (McCrae and Costa Jr, 1999), suggesting that participants with this trait may be more likely to try to meet researcher expectations. This could result in more connected networks as well as more content on average.

While this paper is largely exploratory, given the relative paucity of modern work in this area, the personality measures used here were chosen explicitly in order to evaluate the potential of conceptual networks to provide meaningful behavioral insights. Many of our priors are conflicting; for example, it is unclear whether we would expect someone with more ideas (nodes) to have denser networks (more edges) or sparser networks (relatively fewer edges). Thus, in evaluating the results of these data – which will be discussed in Section 3.5 – this paper will look primarily for overall trends and clusters of network characteristics which appear to behave similarly.

### 3.3.2 YouGov Ideological Turing Test

The second dataset engaged subjects in an “ideologue Turing test,” asking them to provide two short response texts to the same prompt – one arguing the liberal position and one arguing the conservative position. Respondents were explicitly instructed to “write as if you really hold those views. Try to convince someone you don’t know that
you actually believe each position.” Each respondent was randomly assigned one of three issue areas from a pool covering topics of abortion, minimum wage, and national defense. Responses were relatively short, averaging about 17 words each, with the longest responses around 50 words. Subjects were screened with a short demographic questionnaire in which they revealed their true ideological position. Only respondents who indicated they were liberal or conservative – e.g., not moderate – were eligible to complete the Turing test.

Based on the evaluation of human coders, roughly half (56%) of respondents participated in good faith and genuinely tried to argue both sides of their assigned issue. Interestingly, the other half of respondents did not generally submit pure linguistic nonsense, but rather made inauthentic arguments which were merely caricatures of opposing views. For example: “Bomb everybody who disagrees with us,” or “It is okay to murder a fetus, as long as a gun is not involved!” In other words, many of these non-compliant participants did technically submit both liberal and conservative arguments, though some of their arguments – particularly those that did not align with their own ideology – were of low quality. This presents a particularly challenging but interesting NLP problem: from a purely linguistic point of view, there is nothing wrong with these texts; they make perfect grammatical sense. However, they are poor arguments in a more meaningful sense – they offer no evidence or justification, and may not have a coherent premise.

This second dataset, therefore, allows us to further interrogate the value of conceptual network structures in understanding political behavior. While the Mechanical Turk study allows for the examination of correlations between network structure and ideological and personality traits, it remains agnostic as to the primary drivers of that structure. That is, those data cannot indicate whether inferred network structure truly is an individual-level trait or merely a reflection of ideological talking points. Here, the Turing test formulation allows for the exploration of a “reasoning fingerprint”
by disambiguating individual style from partisan messaging. While we might expect conceptual network structures to be primarily tied to individual-level covariates, this setup allows us to test whether it is merely a reflection of the ideological position being espoused.

Second, given the high rate of ironic responses, we can examine the extent to which conceptual network structure serves as an indicator of argument quality. Can we tell whether someone is participating in good faith or responding ironically based on the inferred structure of their text? There is no strong a priori reason to believe this would be the case, yet, if there is indeed a behavioral signal captured by conceptual network structure, it would serve as further evidence that conceptual networks convey meaningful information and that this classic line of inquiry should indeed be revived using modern computational techniques.

3.4 Methods

This paper presents a method for inferring the latent conceptual network structure of short text. While the literature suggests that cognitive reasoning and linguistic expression are both best modeled through network structure, two important theoretical questions must be considered when developing such a model. First, what precisely is being connected, and second, what is the nature of those connections? In other words, in the resulting network model, what do the nodes represent, and what do the edges represent?

This section will theoretically motivate the node and edge representations in a conceptual network, and describe a new method for inferring these constructs. The section will then describe the challenges of network measurement and present several tools to measure and compare inferred network structures.
3.4.1 Inferring concepts

A conceptual network is intended to represent the interconnections between concepts, which in turn requires the operationalization of what constitutes a “concept.” In his classic work on semantic memory, Quillian (1967) argues that a “concept” can be understood as a compressed object which contains indefinitely large amounts of information. As a cognitive process, then, concepts serve as a heuristic guide to the boundaries of a topic that would otherwise require an arbitrarily large amount of resources to describe precisely. In this sense, a “concept” is a recursive knowledge structure in which a meta-concept is itself comprised of a network of sub-concepts.

This is the core intuition behind semantic network libraries such as BabelNet (Navigli and Ponzetto, 2012), ConceptNet (Speer and Havasi, 2012), and SNePS (Shapiro and Rapaport, 1987). Notably, these libraries make an additional necessary assumption: “concepts” are encoded as words. A concept can then be described best through its associated concepts, which themselves are encoded as words. In other words, a “concept” can be thought of fundamentally as collections of closely related words. Identifying the concepts in a text then means determining which words refer to the same amorphous topic.

Fortunately, this is exactly the motivation behind word embeddings (Mikolov et al., 2013). Trained on vast corpora of data, embeddings place words in a high-dimensional space representing the contexts in which those words appear. Words which are similar – or, more precisely, which occur in similar contexts (Firth, 1957) – will appear close together in this space and can be clustered into concepts. Because training such models requires enormous quantities of data, this paper uses a publicly available dataset of pre-trained embeddings, which have been found to be appropriate for most tasks (Spirling and Rodríguez, 2019).

Specifically, the model uses embeddings pre-trained on 100 billion words from the
Google News corpus (Mikolov et al., 2013). This dataset embeds words in 300-dimensional space by maximizing the average log probability:

\[
\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)
\]  

(3.1)

For a sequence of training words \( w_1, w_2, \ldots w_T \) and a context window of \( c \). Once embedded, word similarity can be measured as the cosine similarity between two words’ vector representations.

Leveraging word embeddings to take similar words as pointing to the same concept (Firth, 1957) allows subjects to have lexical diversity without considering every unique word as a unique concept. Specifically, clusters of words are taken to refer to the same concept if all words in that cluster have cosine similarity greater than 0.5. Concepts are arbitrarily labeled with one of their constituent words. Stop words and words which do not have trained embeddings are excluded from the analysis. Furthermore, using part of speech tagging, pronouns, and other referent words are replaced by the word to which they refer.

### 3.4.2 Inferring connections

![Figure 3.1: Example of the grammatical parse of a sentence](image)

Figure 3.1: Example of the grammatical parse of a sentence
The next challenge is determining both theoretically and operationally what constitutes the connections between inferred concepts. The simplest approach is to define interconnections based on word co-occurrence: two concepts are connected if constituent words occur within some fixed window of each other. This method, however, is theoretically under-motivated. Defining edges by co-occurrence suggests that linguistic distance is the core driver of conceptual relations: that any concepts which appear near to each other are related and – perhaps more concerning – that concepts must be syntactically near in order to be related.

This belies the nature of linguistic communication: near-ness may be an indicator of conceptual connection, but it is too simplistic a measure for the richness of natural language. Efforts which have sought to infer conceptual network structure through hand-coding (Axelrod, 1976; Shaffer et al., 2009) would have been much more tractable if co-occurrence was a sufficient measure of conceptual connection.

This paper, therefore, proposes an approach which leverages grammatical structure in order to determine conceptual relations. Specifically, the model determines the grammatical parse of a text by identifying each word’s part of speech and their syntactic dependency relations. In future work, we plan to integrate semantic as well as syntactic parsing. This syntactic parse identifies the grammatical relations between words, linking, for example, adjectives to the nouns they modify and subjects to their related objects. An example of grammatical parsing can be seen in Figure 3.1. Importantly, these grammatical connections can be meaningfully interpreted – indeed, the very purpose of these grammatical rules is to serve as a tool to help humans encode and decode linguistic communication.

While the grammatical parse serves as the network’s foundation, this structure is modified through the process of inferring concepts described above. When terms, such as stops words, are removed from the network, any remaining parent and child
nodes are connected in their place. Any concept which occurs multiple times – either through the repetition of a word, use of a referent word, or through conceptually similar words – are taken to be the same node, with all their external links shared. Additionally, negative words (such as “not”) are removed and replaced with a negative tie between grandparent and child terms. These steps result in a weighted, signed network of conceptual interrelations. Example networks using these methods to infer concepts and their relations from text can be seen in figure 3.4.2.

![Networks showing conceptual interrelations](image)

(a) Example liberal position on abortion

(b) Example of conservative position on abortion

3.4.3 Network measures

There are many methods of network comparison, but these frequently rely upon networks having the same content (e.g., nodes and node labels), and measure network distance as differing patterns of interconnection. Here, the networks generated across individuals do not necessarily share any of the same nodes, and we are primarily interested in how structure varies, independent of variations in content.
Portrait divergence (Bagrow and Bollt, 2019) provides one antidote to this, allowing for pairwise comparison of arbitrary networks. This approach defines a graph’s portrait as an asymmetric matrix, $B$, in which the $B_{kl}$ entry captures the number of nodes $k$ that have path length $l$. This portrait contains both high and low dimensional information about the structure of the network. The network’s degree distribution, for example, is captured by the first row ($B_{1,k}$), while the shortest path distribution is encoded as $\frac{1}{2} \sum_{k=0}^{N} k B_{l,k}$. This matrix is normalized to the row-wise cumulative distribution of $B$, and the similarity between two networks is calculated as the Kolmogorov-Smirnov test statistic $K$, e.g., the maximum distance between the two matrices.

While we use portrait divergence to measure pair-wise similarly between networks in Section 3.5.2, this approach does not allow for a fine-grained understanding of the ways in which dissimilar networks are structured. We therefore also describe the topology of inferred networks through seven network measures that can compare resulting structures across several key dimensions. Building off the moral philosophy literature, the model engages measures that capture connectivity, complexity, and hierarchy – dimensions which capture different understandings of good moral “coherence.”

Connectivity serves as the baseline for coherence and is measured here through the percent of nodes which are in the giant component. Complexity is suggestive of the particularist view (Dancy, 1993) of coherence which advocates for richly interconnected and resilient networks. This approach is captured through the measures of average degree, clustering, and density. Finally, hierarchy reflects the view of utilitarianism as advanced by Sidgwick (1907) and covers measures of entropy, disassortativity, and standard deviation of degree. These measures are described in detail in Table 3.1.
<table>
<thead>
<tr>
<th><strong>Connectivity (Baseline)</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Giant component percent</td>
<td>The percent of nodes in the largest component of the network, ( N_G/N ). This measure indicates how cohesive the network is. A value of 1 indicates the network has a single component (e.g., a path exists between any two nodes), while lower values indicate that the network has multiple, disconnected components.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Complexity (Dancy, 1993)</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Degree (k ( \text{avg} ))</td>
<td>The average degree across all nodes in the network. Higher values indicate that nodes have more connections on average.</td>
</tr>
<tr>
<td>Clustering</td>
<td>A measure of how locally-connected a network is. High values indicate triadic closure (Saramäki et al., 2007), while low values indicate a network that is locally tree-like.</td>
</tr>
<tr>
<td>Density</td>
<td>The ratio of existing edges to the total possible edges, ( 2E/(N(N - 1)) ). This is a measures the overall interconnectivity of a network with a value of 1 indicating that every idea is connected to every other idea, and a value of 0 indicated that no concepts (nodes) are connected.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Hierarchy (Sidgwick, 1907)</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>The amount of information contained in the network’s normalized degree distribution (( p_k )) (Shannon, 1948). This measure is dependent on both the length of the distribution (e.g., ( N )) and the heterogeneity of the distribution. Measured as (- \sum (p_k \times \log(p_k)))</td>
</tr>
<tr>
<td>Disassortativity</td>
<td>Measured as the inverse of the Pearson correlation coefficient, ( -r ), disassortativity captures the degree homophily of the network (Newman, 2003). Values range from (-1) to (1), with a value of (1) indicating that high degree nodes tend to connect to low-degree nodes, as in a star network. Note that for this study, we use disassortativity in order to have the same dimension and valence as the standard deviation.</td>
</tr>
</tbody>
</table>
The standard deviation of the network’s degree distribution. Lower numbers indicate that nodes are more homogeneous in their degree, whereas larger values indicate a greater difference between the lowest and highest degree nodes in the network. For the purposes of this study, \( k_{\text{std}} \) is normalized against a hub-and-spoke network with the same number of nodes \( N \).

**Table 3.1**: Measures of network structure.

While each individual measure captures a single feature of network structure, together these measures provide a holistic description of a network’s local and global characteristics. For example, while the average degree – the number of connections nodes have on average – is a valuable piece of information, it alone does not provide detailed topological insight. From that single statistic, we cannot tell whether a network is heterogeneous (has nodes of differing degrees) or homogeneous (has nodes of similar degrees), whether it is connected or has multiple components, nor whether it is densely interconnected or sparse.

To provide a more intuitive sense of what these measures indicate, Figure 3.2 compares these statistics across three stylized example networks. Each network has a fixed number of nodes (\( N = 16 \)) and edges (\( E = 24 \)) – resulting in equal density (\( d = 0.2 \)) – and are constructed to have an equal average degree (\( k = 3 \)). However, these networks display strikingly different topological properties, which are conveyed through additional network statistics. In particular, we see that higher standard deviation and higher disassortativity are both indicative of heterogeneous, hub-and-spoke like structures. Entropy provides a weakly opposite indicator with homogeneous networks having slightly higher entropy than heterogeneous networks. Given that entropy is calculated as \(- \sum (p_k \times \log p_k)\), the minimal effect of variations in degree distribution suggests that higher entropy is more likely to be indicative of higher node count.
The final two network measures, giant component percent and clustering, each provide unique topological insight not captured by the other network measures. Specifically, the giant component percent indicates whether a network is connected or fractured into multiple components, while clustering indicates the presence of triangles – e.g., the tendency of nodes that share a neighbor to themselves be connected.

It should also be noted that these network measures differ in how robust they are to noise. Statistics such as average degree, standard deviation of degree, and density are
among the more robust measures, and will not change significantly with the random addition or removal of edges. Giant component is perhaps the least robust measure, as the random removal of a single edge could result in an isolated node and thus prevent a network from being completely connected.

Given these seven measures, we can then compare structural proprieties across networks, determining which networks are topologically similar and which are divergent. Furthermore, by examining the full set of metric-level comparisons, we can gain insight into the drivers of topological similarity or difference.

3.5 Results

This paper presents a method for inferring the latent network structure of concepts within textual documents. While the literature clearly supports the theoretic motivation for the existence of such latent network structure, it remains to be seen whether there is value in developing such a method. This paper, therefore, demonstrates the value and implications of this approach through three applications. First, using the Mechanical Turk dataset, the paper demonstrates that the network structure inferred from an individual’s text is meaningfully correlated with ideology as well as known latent personality traits. This suggests that expressions of political preferences – not just the preferences themselves – are tied to behavioral traits. Second, the YouGov dataset illustrates that these correlations are not simply ideological and more reflective of ideological correlates than simple partisan talking points. Finally, the YouGov data are further used to show that the method presented here can distinguish between ironic responses and texts which authentically reflect an ideological view. Together, these findings suggest that there is a meaningful signal within the very structure of these expressed reasons.
### 3.5.1 Personality and Reasoning

![Graph showing correlations between network statistics and latent personality and demographic measures]

**Figure 3.3:** Correlations between network statistics and latent personality and demographic measures

As described in Section 3.3, subjects in the Mechanical Turk study completed three network elicitation activities for two issue-area prompts. Therefore, in order to assess the relationship between network statistic $s$ and personal trait $p$, we employ a multilevel model which includes topic and method random effects:

$$ s = \beta p + \alpha_t + \alpha_m + \epsilon \quad (3.2) $$

The resulting correlations between inferred structure and personality measures are shown in Figure 3.3. We find a striking left/right divide in the structural properties of subjects’ inferred networks. Again, it is worth noting that this structure is separate
from the content of that reasoning, suggesting these subjects differ not only in what they say but fundamentally in how they say it. This divide can be seen through the fact that subjects with conservative ideology (Pew Research Center, 2017) tend to have similar structural properties to those who score high on the traditionally conservative Moral Foundations dimensions of Purity and Authority (Haidt and Joseph, 2008). Those who score high on the traditionally liberal dimensions of Fairness, Openness, and with a high Progressivism total seem to also share similar structural properties. Notably, subjects with high political knowledge do not appear to fit neatly into either a progressive or a conservative track, suggesting – as we would expect – that knowledge is a trait orthogonal to ideology. This finding further supports the external validity of our construct. Additionally, we see these patterns repeated across demographic measures, with subjects who are older, Republican, white not of Hispanic origin, and male more likely to demonstrate “conservative” properties.

Specifically, we see that “progressive” subjects tend to create networks that have a higher standard deviation of degree ($k_{std}$), entropy, average degree ($k_{avg}$), clustering, and disassortativity while “conservative” subjects tend to be lower on each of these dimensions. As illustrated by the example networks in Figure 3.4.2, higher values of $k_{std}$ and disassortativity suggest more heterogeneous, hub-and-spoke like networks. On the other hand, higher values of $k_{avg}$ and clustering suggest more interconnected networks, while high entropy suggests either a more homogeneous structure or more content (nodes). Taken together, this combination of network statistics suggests that progressive subjects tend to form networks with a core-periphery structure – that is, networks with an interconnected core of central ideas surrounded by a periphery of loosely connected auxiliary ideas. The weak signal sent by the density metric is further suggestive of this, as a network with a dense core and sparse periphery would have a non-remarkable density on average. While the higher values of $k_{std}$,
entropy, and disassortativity are suggestive of hub-and-spoke type networks and may be reflective of Sidgwick (1907)’s utilitarian view of coherence, the higher values of $k_{avg}$ and giant component percent suggest these networks have a richer, more interconnected structure which may be more in line with the particularist philosophy (Dancy, 1993).

Conservative subjects, on the other hand, produce networks with lower $k_{std}$, $k_{avg}$, entropy, and clustering. Taken together, this suggests that these subjects produce more homogeneous networks in which each idea is roughly similarly connected, but further suggests these subjects tend to produce less content overall. We also see, through the giant component metric, that conservative subjects are more likely to produce networks with multiple, disconnected components while progressives are more like to produce connected networks, suggesting that a major difference in structure may be a tendency to “bridge” between different clusters of distinct thought, with progressives more likely to tie disparate concepts together and conservatives more likely to articulate differing strains of though separately.

While the conceptual network structure inferred for liberal respondents is reflective of certain liberal traditions that would claim these structures to be morally superior (Rorty et al., 1989), it is too soon to make any claims as what constitutes a “good” conceptual network structure. Given the method’s reliance on grammatical structure, text that results in a sparse network could simply be reflective of non-standard English rather than deep moral views. While this analysis suggests grammatical structure is a meaningful correlate of personality traits, we should be wary of extending that to a normative assessment. Furthermore, the analysis of the Mechanical Turk data cannot disambiguate between possible effects. These same personality traits are believed to lead to ideological positions (Haidt and Joseph, 2008), making it unclear whether variations in structure are indeed an indication of personality or, more generally, a reflection of a given position’s talking points.
3.5.2 Reasoning Fingerprint

This work ultimately aims to provide a tool that can provide insight into individual-level reasoning phenomena, but it can only meaningfully do so if there is individual variation in inferred structure. That is, if a method has any potential to bring insight to the dynamics of individual opinion change and conversation quality, it must be able to pick up on meaningful signals at the individual level. Furthermore, if we are to think of this as an individual measure, we need to demonstrate that it is not merely capturing some element of group identity – such as common talking points around a shared ideological view.

In the YouGov dataset, each respondent provides two, ideologically opposed reasons which have been judged by a human coder to be authentic attempts to represent those points of view. We can, therefore, ask whether individuals tend to produce networks similar to themselves or similar to others within the same category. That is, will the structure of $C_i$, the conservative essay produced by respondent $i$, be more similar to that same user’s liberal structure, $L_i$, or to the structure $C_j$: user $j$’s conservative essay on the same topic? If $C_i$ and $C_j$ are more similar, it suggests that any structural features are driven to some degree by the content; e.g., that conservative arguments have similar structure regardless of who is doing the arguing. If $C_i$ and $L_i$ are more similar to each other, it suggests that there is some individual argument style – that $i$ will produce similar structures across dissimilar topics. Finally, we may find no patterns in similarity, suggesting that there is neither an individual nor ideological signal within the inferred structure of text.

Using portrait divergence (Bagrow and Boltt, 2019), we generate a single point estimate of pairwise similarity. For each subject who participated in good faith, we compare the similarity between the two networks produced by that individual to the networks produced by others. Because we would expect several individual-level covariates to result in self-similarity, we restrict the comparison set to those most like
the subject being considered. Specifically, comparison texts are restricted to those that are on the same topic, are written by subjects with the same ideology as the comparison subject, and present the subject’s true ideological view.

This creates two distributions: one capturing self-comparison and the other illustrating each subject’s comparison to the ideologically compatible portion of the subject pool. These distributions are shown in Figure 3.4.

![Figure 3.4](image)

**Figure 3.4**: Distribution of network similarity for authentic respondents. “Self” captures similarities between a single respondent’s liberal and conservative text, while “Ideological” captures between-subject similarity for a fixed topic and ideology. A similarity of 0 indicates that networks have identical portraits.

As we can see, networks inferred from a single individual are slightly more likely to be more similar than networks inferred from different individuals. A t-test shows that this difference is significant (p < 0.05). While there may be other individual correlates, such as education, driving this result, this finding does suggest that respondent essays do not merely reflect ideological talking points. Particularly on subjects like abortion, when ideological views are very established and well known, it is entirely possible that conceptual network structure would have been largely driven by ideology. This finding,
therefore, suggests that the inferred network structure is capturing something about individual style or preference. Put differently, the structure of reasoning appears to be an individual characteristic rather than a topical or ideological one.

### 3.5.3 Authenticity and Irony

As described in Section 3.3, nearly half of respondents in the YouGov dataset provided inauthentic, but linguistically meaningful answers. We can, therefore, aim to separate authentic from inauthentic answers using our inferred structural features. If such a classification can be done, it suggests that the inferred networks are indeed meaningfully encoding the latent structure of the text. I compare two approaches to this task.

First, as a baseline measure, I consider two common measures of text sophistication: word count and Flesch-Kincaid readability. A text’s Flesch-Kincaid score is calculated based on the number of words, sentences, and syllables in a text, with higher scores indicating more complicated texts. Figure 3.5 shows the cumulative distribution of...
these measures within both the authentic and inauthentic responses. Here we see that inauthentic responses do tend to have fewer words, but not dramatically so. Texts in both samples have nearly identical Flesch-Kincaid scores, suggesting this will not be a helpful feature for separating these categories.

The second model considers the inferred network structure of the text, using the network measures described in Section 3.4.3, and a third model includes all features from Models 1 and 2.

Table 3.2 shows the results of a logistic regression for each of these models. Comparing out-of-sample accuracy\(^4\), we find that Model 1 accurately classifies 70% of the texts, while the network features of Model 2 improves upon this to accurately classify 74% of the texts. This suggests that the network features of Model 2 provide some signal as to the authenticity of a text that is not captured by the course features of Model 1.

Looking at the effects of each feature, we see that word count is indeed driving the performance of Model 1, with the Flesch-Kincaid score producing a small and insignificant result. In Model 2, we see that several network features appear to encode a signal supporting the classification. Specifically, the average degree and density both help indicate whether or not a response is authentic. In Model 3, we see that these effects continue when all features are considered, with word count, average degree, and density, all containing information. Specifically, as we would expect, texts with more words are more likely to be authentic. From the network measures, we further see that a higher average degree is more likely to indicate authentic texts, while networks with higher density are less likely to indicate authentic texts. Noting that word count is highly correlated with the number of nodes in a network, this suggests that authentic networks are characterized not only by more content but by content with meaningful interconnections. That is, nodes on average have a higher

\(^4\)With an 80% in-sample, 20% out-sample split.
degree, but, due to the higher number of nodes, authentic networks are overall less dense. Ironic responses, on the other hand, are characterized by denser networks with overall less content. While highly interconnected, these networks are limited in how many other nodes can be connected to, resulting in a lower average degree than the authentic responses.

One way to interpret these results is to think of these network features illustrating the strength of an argument in the Aristotelian sense. An argument structured as a major premise followed by a minor premise (Toulmin, 1958), would have network characteristics similar to what we see in the authentic arguments: multiple points which are strongly, but not overly densely, connected to each other. This could suggest that ironic responses are lacking any such minor premise and rather state a simple and densely connected claim without providing any justification.

We might further think of these structures in terms of their texts’ Gricean Implicature (Grice, 1975). That is, ironic texts explicitly say one thing while meaning another. Note that having such an implicature is itself not necessarily a sign of an inauthentic response – in conversational language, enthymemes are frequently employed to refer implicitly to shared knowledge without making an explicit argument. In this corpus, however, the implicature of a text is what determines its category – either it says what it means and is authentic, or it says something other than what it means and is ironic. These network structures, then, may indicate whether or not a respondent is genuinely aiming to be understood (Grice, 1975), with ironic responses less thought through and developed and authentic responses more intentionally constructed.
### Table 3.2: Results of logistic regressions.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>word count</td>
<td>3.122***</td>
<td></td>
<td>1.189**</td>
</tr>
<tr>
<td></td>
<td>(0.295)</td>
<td></td>
<td>(0.475)</td>
</tr>
<tr>
<td>Flesch Kincaid</td>
<td>0.480</td>
<td>-0.357</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.450)</td>
<td>(.509)</td>
<td></td>
</tr>
<tr>
<td>giant component</td>
<td>-0.543</td>
<td>-0.260</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.394)</td>
<td>(0.409)</td>
<td></td>
</tr>
<tr>
<td>k avg</td>
<td>3.011***</td>
<td>2.672**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.100)</td>
<td>(1.129)</td>
<td></td>
</tr>
<tr>
<td>clustering</td>
<td>0.610</td>
<td>0.705</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.655)</td>
<td>(0.659)</td>
<td></td>
</tr>
<tr>
<td>density</td>
<td>-2.612***</td>
<td>-2.273***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.494)</td>
<td>(0.518)</td>
<td></td>
</tr>
<tr>
<td>entropy</td>
<td>-0.111</td>
<td>0.170</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.614)</td>
<td>(0.624)</td>
<td></td>
</tr>
<tr>
<td>disassortativity</td>
<td>0.058</td>
<td>0.198</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.390)</td>
<td>(0.396)</td>
<td></td>
</tr>
<tr>
<td>k std</td>
<td>-0.876</td>
<td>-1.711</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.918)</td>
<td>(0.986)</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- *p<0.1; **p<0.05; ***p<0.01
- Model (1): 70% predictive accuracy
- Model (2): 74% predictive accuracy
- Model (3): 74% predictive accuracy
3.6 Discussion

Arguments for conceptual networks have been made in a variety of fields for decades. Particularly within the public opinion literature, such models have been seen as a crucial tool for understanding political behavior and making sense of the deliberative system (Lane, 1962; Axelrod, 1976; Campbell, 1960). However, this line of work was largely abandoned due to the prior lack of data and computational resources. This paper calls for a revival of these methods and demonstrates that conceptual network structure holds the potential to generate meaningful behavioral insight.

Citizens express themselves in individually distinctive ways, and being able to measure those expressions has the potential for behavioral insight along a number of dimensions. In terms of deliberative democracy, the ways in which individuals express themselves and understand each other is likely to influence the ability of people to successfully deliberate and collaborate around matters of common concern. Furthermore, even if we decline to believe that citizens are rational creatures capable of formulating their own views, studying individual variation in expressions of reasoning still holds the potential to illuminate popular reflection of elite messaging. Given modern computational tools, now is the ideal time to revive this classic line of work and develop models for individual expressions of reasoning.

This paper calls for a computational revival of these classic techniques for understanding political opinions. By demonstrating that conceptual network structures can be computationally inferred from text and meaningfully reflect individuals' personality traits, this paper underscores the value of revisiting questions around individual expressions of political reasoning. Due to the relative lack of modern work in this area, this paper is largely exploratory and demonstrates the value of considering the structure of expressed opinion in the face of conflicting behavioral priors.

Specifically, this paper has presented a text-based method for inferring conceptual
network structure. This method utilizes grammatical structure to capture the implicit connections between concepts and uses word embeddings (Mikolov et al., 2013) in order to identify which words point to the same concept. This approach captures the richness of natural language and the subtle ways in which individual expressions differ. Building off a long line of work in moral philosophy, inferred network structures are considered in terms of differing understandings of moral “coherence.” A range of network measures capture the richness of these structures and are used to classify the argumentative quality of texts.

This method is applied to two datasets: an original study conducted on Amazon’s Mechanical Turk and a nationally representative YouGov study conducted by Dr. Dan Hopkins and Dr. Hans Noel. In the first dataset, 100 subjects completed short response essays on two different political prompts. These subjects further completed a battery of demographic and personality questions. These measures covered Big 5 personality traits (John and Srivastava, 1999), Moral Foundations (Haidt and Joseph, 2008), and other politically relevant topics (Gastil et al., 2012; Pew Research Center, 2017; Carpini and Keeter, 1993). As expected, we found a distinctive divide between ideologically conservative and ideologically liberal respondents. Subjects with a conservative ideology or who scored high on the traditionally conservative measures of authority and purity all produced consistent network structures. These structures were overall sparse and less likely to be connected. Liberals, on the other hand, all produced relatively more content and generated network structures consistent with then particularist view of moral coherence (Dancy, 1993).

In the YouGov dataset, subjects responded to a single issue prompt but were asked to provide two texts – one professing the conservative position and the other describing the liberal position. Respondents were asked to write both responses as though the genuinely held that view, though only about 56% of respondents did so in good faith. While the first dataset illustrates the ideological similarity of inferred
network structures, this dataset underscores the continued need to consider individual covariates. In general, subjects tend to produce more self-similar network structures than ideologically-similar network structures. That is, a single subject’s liberal and conservative essays are likely to be more similar to each other than that subject’s essay is to an ideologically similar essay. This relationship holds even when the comparison is limited to essays on that subject’s true ideological views written by a subject with a shared ideology. While we perhaps may not be too surprised that individual covariates play an important role here, this serves as an noteworthy reminder that there is more than mere ideology at play.

Furthermore, since nearly half of subjects did not participate in good faith, the network method is used to differentiate between genuine, “authentic” responses and “ironic” responses that do not appropriately capture an ideological view. This analysis found that coarse features, especially word count, send a strong signal as to the authenticity or irony of a response, but further indicate that network features provide additional information as well. Specifically, a model with both word count and network features classifies these responses with 74% accuracy, a significant improvement over the baseline, and a notable improvement over word count alone (70%).

Taken together, these results underscore the potential of conceptual network methods to shed meaningful insight into questions of political behavior. This classic line of public opinion research bolsters deliberative democracy, valuing individuals’ political expressions rather than being content to examine preferences only in aggregate. Such a fine-grained approach to public opinion was once beyond our reach, making aggregate models the only practical solution. However, modern computational techniques and data accessibility have made this problem tractable. This paper demonstrates that we can indeed derive meaningful behavioral insight from individual variation in expression. It is time to revive this classic line of research.
Chapter 4

Why Keep Arguing? Predicting Engagement in Political Conversations Online

4.1 Introduction

Digital communication plays a critical role in our political infrastructure. Online platforms have expanded the reach of “kitchen table conversations,” as people increasingly turn to social media as a primary news source (O’Connor et al., 2010; Lee and Ma, 2012; Bakshy et al., 2015) and elected officials use digital channels to communicate with their constituents (Kavanaugh et al., 2012; Farina et al., 2013). Such interactions are often modeled as one-shot games or as evidence of long-term links in a social network (Feng and Wang, 2013; Myers and Leskovec, 2014). However,

\(^{1}\)Replication materials for this chapter can be found at https://github.com/sshugars/conversations.

\(^{2}\)This chapter is coauthored with Dr. Nick Beauchamp.
much online activity consists not of single-shot, unidirectional interactions with elites, but repeated interactions among peers. These iterated interactions—conversations—have important implications for political theory: while conventional wisdom claims that brief social media interactions have little effect on subsequent online behavior, a number of recent experiments have shown modest but real effects of single-shot interactions (Friggeri et al., 2014; Munger, 2017). Deliberative theory suggests that repeated interpersonal interactions where individuals engage in extended conversation may have even more substantial effects (Bednar and Page, 2007; Habermas, 1984; Axelrod, 1987).

In this paper, we focus not on persuasive outcomes, but on the more fundamental question of what leads people to engage in extended online conversation and argument in the first place. Existing work in this area has generally had a more practical bent, focusing on tweet- or conversation-level recommendation, and aiming to predict user interest in conversation threads in order to better curate and recommend targeted content. Such work has looked at user engagement in various forms of online conversation (Chen et al., 2012; Yan et al., 2012; Vosecky et al., 2014; He and Tan, 2015), as well as via retweeting (Feng and Wang, 2013; Hong et al., 2013) and re-entry back into existing conversations (Backstrom et al., 2013). Our work here is closest to the latter: we are more interested in the extended dynamics of conversation, particularly the decision to reengage or exit, than in the initial decision to interact. We focus on users who have already made that initial participation and seek to understand and predict whether and when they reengage based on user, tweet, and thread-level features.

While ongoing deliberative conversation is the substantive focus here, this framing also turns an impossible problem—predicting initial responses to a tweet out of the entire pool of twitter users—into a practicable prediction task—predicting re-participation of users whom we know are already part of a conversation. This
approach also conditions out the even harder problem of explaining the origins of an initial tweet or conversation, particularly given the immense variety of motivations behind those first moves. Instead, we focus on existing conversations – at least a first move followed by a response – and model the processes that lead to extended and branching conversations among existing participants. Twitter might seem less suited to such models than traditional online forums, but Twitter, in fact, produces immense quantities of impromptu extended, branching conversations, and by focusing only on re-entry by existing participants, we can study what causes individuals to continue an argument or drop out, bracketing the question of initial engagement.

By conditioning on existing user interaction, we aim to get more deeply at the question of why people bother arguing online. What brings them back to a repeated argument? What factors contribute to an individual returning to or abandoning an argument? While in face-to-face settings, social etiquette suggests that a comment will most likely be greeted with a response, there is no \textit{a priori} reason to expect a response to the vast majority of online posts. While we expect to find many types of conversations occurring online, we might expect that more extreme content (positive or negative) will increase engagement, as trolls successfully incite arguments and partisan allies reinforce each other’s positions (Cheng, 2017). Between the extremes lies a more productive and interesting mode of engagement: true deliberative argument, in which participants exchange content in a genuine attempt to persuade or inform. Such behavior is not as uncommon as skeptics might assume, and is prevalent in knowledge-sharing platforms such as StackOverflow, Yahoo! Answers, and other such forums, where users may be motivated to some degree out of a general sense of community (Adamic et al., 2008; Oktay et al., 2010; Anderson et al., 2012) even as they argue over better or worse solutions to shared problems.

We find evidence for all of these behaviors in our data, and in particular, show that while many of these engagements are negative, conversations often cover a range of
emotions and go on far longer than a single-shot attack or mutual trolling might suggest. While we leave for later the ultimate question of persuasive effect, we establish here that even a medium as apparently unpromising as Twitter is full of complex, extended political conversations, and that individuals’ decision to repeatedly reengage in those conversations is surprisingly systematic.

4.2 Related Work

Much of the theoretical work around conversational dynamics has been done within the literatures of deliberation and persuasion. While both these approaches focus their attention on back-and-forth conversations, they vary in their characterization of those conversations. The persuasion literature looks broadly at how people convince others and “win” arguments, while the deliberative ideal imagines thoughtful participants reasoning together to generate public opinion centered on the common good (Cohen, 1989; Habermas, 1984). ‘Reasons’ may constitute factual arguments or emotional appeals (Mansbridge, 2015), but ideal deliberation is often taken to be free of persuasion, coercion, or other forms of instrumental action (Habermas, 1984). Contra persuasion models, ideal deliberators should engage in rational speech acts – aiming to honestly express themselves and truly trying to understand the other. Huckfeldt et al. (2004) argue that ideal citizens “are those individuals who are able to occupy the roles of tolerant gladiators – combatants with the capacity to recognize and respect the rights and responsibilities of their political adversaries.” If political debate serves to sharpen our own understanding and build our collective knowledge, then we owe it to our interlocutors to press them on their positions; to find the holes in their armor and encourage refinement of beliefs. The process of debate makes us all better - thus allowing tolerant gladiators to walk away as friends. Citizens who silence their discussants, seek to coerce others, are easily persuaded by false beliefs,
or who otherwise refuse to engage in rational argument, therefore, do a disservice to themselves and their communities.

Experience tells us, however, that such a lofty deliberative ideal is rarely met in political conversation. Sunstein (2002) advances the “law of group polarization,” finding through numerous empirical studies that “deliberation tends to move groups, and the individuals who compose them, toward a more extreme point in the direction indicated by their own pre-deliberation judgments.” Sunstein argues this polarization is the natural result of the social context, which serves as a significant driver of individual actions and opinions. Hearing friends express a view makes a person socially inclined to express the same view. In other words, deliberating groups tend towards extremism in the direction of the pre-deliberation median because nobody wants to take the social risk of expressing an unpopular view. Sanders (1997) similarly argues that the broader context of power dynamics frequently has a debilitating but under-recognized effect on deliberation, as marginalized individuals feel silenced and unable to share their true opinions. Importantly, the majority of participants may mistakenly assume that such power effects are negligible if “deliberation appears to be proceeding.”

Another line of work has tackled conversational dynamics from the perspective of the data-processing problem of platform curation, e.g., trying to predict which posts will be popular for the purpose of highlighting those posts for users. Much of this work focuses on post-level engagement, predicting engagement as a function of topics (Hong et al., 2013) or social network structures (Pan et al., 2013; He and Tan, 2015). Much of this work has considered ‘popularity’ as a raw aggregate of engagement with an initial post, finding, perhaps unsurprisingly, that the popularity of a user’s past content is a strong predictor for the popularity of their future content (Artzi et al., 2012). Backstrom et al. (2013) break the task into related subtasks: length prediction and re-entry prediction. Intuitively, these subtasks indicate distinctive types of threads:
threads which are long because a high number of users chime in a small number of times - to offer congratulations or condolence, for example - while other threads are long because a small number of users contribute a large number of times in a back-and-forth conversation. Supporting this theory, Backstrom et al. (2013) find that the number of distinct users in long threads follows a bimodal distribution. Using data from Facebook and Wikipedia, Backstrom et al. (2013) find the identities of recent commenters is most predictive of conversation re-entry.

This lattermost line of work is largely a-theoretical and not particularly concerned with normative issues. While the current study borrows many of their methods, we are also fundamentally interested in the dynamics of online conversation from a deliberative perspective. Thus we are interested less in conversation recommendation or modeling engagement in conversations per se and more focused on how individual speech acts (tweets) lead existing discussants to re-engage with each other or abandon a conversation. Regardless of the outcome of a conversation, it is important to understand what sustains conversations – particularly acrimonious ones – and keeps mutual opponents or supporters engaged with each other. As we will see, this engagement can take more or less productive forms, but simply understanding the deliberative dynamics is an important first step.

4.3 Data

For this study, we collected a corpus of 7053 Twitter conversations. These conversations were identified by first using Twitter’s Streaming API to collect all tweets which contained the keyword “Trump” during a week in October 2017. We returned to these tweets approximately three months following the collection window to extract the entire conversation of preceding and following replies. Using the Twitter REST API, we retrieve full metadata for every tweet in the conversation. We discard conversations
shorter than a minimum of three exchanges or in which tweets have been deleted, as metadata for those tweets cannot be retrieved.

Our collection method, then, results in a sample of conversations in which at least one tweet took place during the October 2017 collection window and included the keyword “Trump.” While the majority of sampled conversations both begin and end in fall 2017 (See Figure 4.2), the latest tweets in our sample come from the time of conversation extraction (approximately January 2018) while the earliest comes from 2016.

The seed keyword “Trump” was intentionally selected in the hopes of identifying particularly charged political conversations. This corpus should by no means be considered a random sample or representative of Twitter discourse more broadly. Rather, these data should be seen as a small collection of spirited political conversations that we use here to build and test a model of political engagement online. In future work, we hope to see this model applied to a more diverse array of conversations.

![Figure 4.1: Distributions of length and users by thread](image)

Our full corpus contains 7,053 conversations comprised of 63,671 unique tweets. As described in more detail in Section 4.4, “a conversation” in the context of Twitter can be understood as a tree containing multiple branching threads, each connecting
to the same root tweet. As shown in Figure 4.1, the distribution of thread length in our sample is heavy-tailed: by construction, the minimum thread length is 2, while the longest thread contains 108 tweets. The average length of a thread is 5.6 tweets with a standard deviation of 4.1, and the mean number of unique users in a conversation is 3.7, with a standard deviation of 1.2. As we might expect from social media engagement, responses tend to occur within a relatively compressed time period. Just under half (40%) of the tweets in our sample are posted within 5 minutes of the tweet that proceeds it in the conversation. About three-quarters (70%) are posted within 1 hour, and nearly all (95%) take place within a day. The cumulative distribution of inter-event time – i.e., the number of hours between tweets – can be seen in Figure 4.2.

![Figure 4.2: Cumulative distribution of the time taken to reply to a tweet for those tweets with replies.](image-url)
4.4 Methods

4.4.1 Model

Our fundamental question is why someone continues to engage in (or stops engaging in) an online conversation. What makes a person participate in or abandon a discussion? We examine this question by modeling conversation as an interlaced exchange between two or more participants. For a tweet observed at time step $t$, we wish to predict whether existing members of the conversation, as defined below, will respond or not respond to that tweet at time step $t + 1$.

While anyone in the Twitter universe may conceivably reply to a tweet, our interest is in modeling the actions of those who are already part of a conversation in a loose sense: first, because this makes the prediction problem practical, and second, because it allows us to model engagement in dialogue, not just taking pot-shots on a microblog. Furthermore, since many threads on Twitter are initiated by entities unlikely to participate in conversation - such as corporations, celebrities, politicians, and bots - we take our pool of candidates to be users who have already responded at least once in a given thread, ruling out the user who initiated the thread unless they also replied to another tweet in the conversation. While this does not fully remove user heterogeneity, it does limit the sample to accounts which actively engage in an exchange. We consider self-replies to be a continuation of a thought, and thus not a ‘reply’ in the traditional sense. We, therefore, do not include a tweet’s author in the list of candidates who may respond to that tweet.

Our predictive model is structured as follows: for conversation $j$, at every time step $t > 2$, we construct a candidate list of active participants who might respond to the current tweet. Those candidates who do reply at time $t + 1$ are assigned an observed outcome of 1 while all potential respondents who do not choose to reply are assigned an outcome of 0. Note that multiple users may respond directly to a
single tweet. In this analysis, we focus on the temporal sequence of replies and use $t$ as an index of that temporal order. Users return to Twitter on their own schedules, generally due to exogenous constraints on their free time, and have the opportunity to respond to any available tweet when they do. Only two-thirds (66%) of users in our dataset make all their comments within an hour of their initial activity, suggesting that it is common for users to engage in conversation over multiple Twitter sessions. While we may generally expect returning users to respond to the most recent tweet, conversational engagement need not follow this temporal order, and users in our dataset seem to flit between the threads of a conversation tree, going back to respond to earlier tweets nearly half (43%) of the time. For all these reasons, measuring $t$ via clock time intervals, e.g., within a Poisson framework (George and Kibria, 2011; Shen et al., 2014), would overlook this exogenously-constrained, bursty, and often non-sequential reply behavior. Using temporal sequence order allows us to relax this assumption and look at the conversational points at which a user re-engages with the understanding that non-engagement could be due to any number of factors. We do, however, continue to use tweets’ timestamp for the calculation of certain features, specifically the number of likes, retweets, and comments visible at the time of a candidate’s reply, as well as to control for time-of-day patterns that affect when users generally engage with Twitter.

Table 4.1: Example conversation flow between participants $A, B$ and $C$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>Conversation Order at time $t$</th>
<th>Candidates</th>
<th>Observed at $t + 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>$A \rightarrow B \rightarrow A$</td>
<td>$B$</td>
<td>$B = 0$</td>
</tr>
<tr>
<td>4</td>
<td>$A \rightarrow B \rightarrow A \rightarrow C$</td>
<td>$A, B$</td>
<td>$A = 0; B = 1$</td>
</tr>
<tr>
<td>5</td>
<td>$A \rightarrow B \rightarrow A \rightarrow C \rightarrow B$</td>
<td>$A, C$</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1 illustrates a conversation thread from time step $t = 3$ to time step $t = 5$. At each step, we show the potential candidates for re-entry, and the outcomes associated with that time step; the number of observations at each time step is equal to the
number of potential respondents, only those candidates who respond are scored as
a 1, with the rest assigned a 0 for that time step in that conversation. Note that
candidate responses are not mutually exclusive and that multiple candidates can
respond (observed result of 1) to a single tweet.

For our dataset of 7053 conversations, this results in 1,016,492 total observations,
with 110,035 observed instances of 1 (candidate users who responded) and 906,457
observed instances of 0 (candidate users who did not respond). This gives us a
baseline prediction accuracy of 89% if we guess that all candidates never respond.

4.4.2 Features

We expect a user’s tendency to reply to a conversation to be influenced by a number
of factors and their interactions. Previous work (Backstrom et al., 2013; Artzi
et al., 2012; Hong et al., 2013; Feng and Wang, 2013) has generally focused on
predicting conversation-level engagement (i.e., whether a user participates anywhere
in a conversation), and has therefore primarily focused on the candidate user who
might reply as well as features of the overall conversational thread. Since we are
interested here in the more specific problem of predicting the points at which an
existing participant replies or does not reply, we further include features related to
the tweet that might receive a reply, as well as features of that tweet’s author.

This gives us three sets of features and related hypotheses which we discuss in the
following subsections:

1. **Candidate and recent tweet features**: This includes features related to the
candidate user’s activity \((H1.1 - H1.3)\) as well as features of that candidate’s
previous tweet in the thread \((H1.4-H1.5)\).

2. **Conversation thread features**: Features related to thread length and en-
gagement ($H2.1 - H2.2$).

3. **Author and current tweet features**: Includes features of the author who may receive a reply ($H3.1-H3.2$) as well as the tweet at time $t$ that may be replied to ($H3.3-H3.5$).

**Candidate & Recent Tweet Features**

At the most basic level, we would expect that active users will be more likely to reply at any given point in the conversation ($H1.1$). The most readily available measures of engagement for a user are the number of others they follow (following count), the number of followers they have (follower count), the total number of tweets they have posted (statuses count), the number of favorites or likes they have given (favourites count), and whether they have a verified account.\(^3\) While measures of activity such as the number of users followed, number of tweets, and number of favorites given would all presumably have a positive effect on reply probability, measures of popularity such as the number of followers or being verified could possibly have the opposite effect, since more popular users may be less likely to enter a scrum with the hoi polloi ($H1.2$).

At the level of thread-user interactions, we would also expect that a user is more likely to reply as a function of how engaged they have already been in the conversation (number of replies) up until that time, and presumably less likely to respond the longer it has been since their last comment ($H1.3$). We include two features to capture this dynamic: a binary variable prev response, which indicates whether the current tweet $t$ was in response to this candidate user, and time since prev, which provides a raw count of how many time steps it has been since the candidate’s last

\(^3\)While there are a broad range of users/entities with ‘verified’ status, the badge is intended to indicate authentic accounts “of public interest.” [https://help.twitter.com/en/managing-your-account/twitter-verified-accounts](https://help.twitter.com/en/managing-your-account/twitter-verified-accounts).
comment. Similarly, we would expect that a candidate respondent will be more likely to re-engage if their most recent tweet in the conversation received positive feedback, as measured by the number of favorites, retweets, and replies the candidate’s previous tweet received (favorite count, retweet count, and reply count respectively; H1.4).

We also examine content-level characteristics for candidates by evaluating the topical distribution and emotional valence of their most recent tweet in the conversation prior to time \( t \). We describe these features in detail in Section 4.4.3. We expect that candidates whose previous tweet was negative or more extreme in its valence are more likely to keep a conversation going (H1.5). Additionally, comparing the topical content of a candidate’s previous tweet to the content of the current tweet, for which we are predicting response, allows for a measure of interest similarity between the two users. After constructing topic vectors, as described in Section 4.4.3, for a candidate’s previous tweet and an author’s current tweet, we calculate the absolute and squared euclidean distance between the two vectors. We treat this as an inferred measure of ideological difference, indicating whether conversations are likely to stay within topic or alternate between users.

**Conversation Thread Features**

Following recent work in this area (Backstrom et al., 2013; Artzi et al., 2012; Hong et al., 2013; Feng and Wang, 2013), we would expect various features of the conversation up until time \( t \) to affect user engagement. The length of the conversation thread (thread length), for instance, is a good indicator of the amount of interest the conversation has generated and therefore is expected to increase the probability of

---

\(^4\)Following the Twitter API’s naming conventions, we use favorite count as the tweet-level feature measuring the number of times a tweet has been liked and favourites count as the user-level feature indicating the number of tweets a user has liked in their account lifetime.
reply ($H2.1$). This effect, however, may decrease with increasing thread length ($H2.2$) as the conversation becomes more disjointed, unwieldy, or difficult to display via the Twitter interface. We therefore also include a quadratic term for thread length.

### 4.4.3 Current Tweet & Author Features

Our approach differs most significantly from past models in that we are interested not only in user engagement overall in a conversation but in predicting the conversational points at which a user chooses to engage. Modeling this requires accounting specifically for the features of each tweet that may be replied to, as well as features of that tweet’s author.

For author characteristics, many of the user features discussed in Section 4.4.2 may also influence the probability of an author receiving a reply. This may be mediated indirectly via the tweet, or directly in those cases where the author is known to the potential respondent. The author’s activity levels, for instance, might be positively correlated with the likelihood of reply, indicating a tendency to produce more engaging tweets ($H3.1$). Conversely, while popularity may decrease a candidate respondent’s likelihood of replying, a tweet from a popular author may be more likely to receive a response ($H3.2$).

In regards to the current tweet itself, there are many coarse structural (as opposed to content) aspects that may reflect latent characteristics of the tweet, such as its general popularity or interest. Using tweets’ timestamps, we calculate the count of the current tweet’s favorites, retweets, and replies that were visible at the time of a candidate’s response. We expect these features to generally have a positive effect on reply probability ($H3.3$), although the first – favorites – may have the opposite effect if this action reflects silent agreement rather than a tendency to respond. A related measure of tweet “quality” is the ratio of retweets to replies, which we also include.
We also include the length of the tweet and the device used to post it.

Because we would expect cyclical variation in activity as users are naturally more likely to be active during certain hours of the day and on certain days of the week, we control for this tendency using cyclic transformations (Cox et al., 2006) of the day and hour at which a tweet was posted. These features are represented with the xday, yday, and xhour, yhour features.

Finally, at the level of the current tweet’s content, there are a wide variety of semantic and other linguistic characteristics that might increase the likelihood of reply (H3.4). A tweet which mentions a large number of users may be more likely to elicit a response from those users or others; a larger number of hashtags may similarly increase the probability of response; and there is some evidence to suggest that tweets which include information such as a URL will be more popular as well (Bakshy et al., 2011).

At the level of sentiment and emotion, we hypothesize that users in our sample will be more likely to engage in emotionally extreme conversations - whether participating in shouting matches of negative emotion or vigorously reinforcing each other with positive emotion (H3.5). Note that this hypothesis is partly constrained by the construction of our sample, which requires candidate users to have previously engaged in the conversation. Our data cannot tell us how many people choose not to engage in heated political conversations at all. We therefore expect that, regardless of tendencies within the broader population, the active users who are candidates in our data would be more likely to be attracted to intense emotion rather than to shy away from it.

We measure the emotional content of a tweet using several methods. These approaches use existing dictionaries to assign a valence score to each word and calculate the overall emotional value of a tweet as the average valence of its component words. We capture a tweet’s sentiment using AFINN (Nielsen, 2011), and its valence using the extended ANEW lexicon (Warriner et al., 2013). We also use VADER (Hutto
and Gilbert, 2014) to examine more closely the negative or positive charge of a tweet. In all of these methods, lower scores indicate negative words, while higher scores indicate positive words. VADER provides separate measures for valence along positive, negative, and neutral dimensions, as well as a compound score, which provides a single valence measure compiled from the three dimensions. Both AFINN and VADER were developed primarily for measuring sentiment in social media corpora. We also use ANEW (Warriner et al., 2013) to calculate arousal and dominance scores for each tweet. Arousal indicates the intensity of emotion, from calm to intense, while dominance scores indicate the degree of control, from vulnerable to powerful.

Finally, to capture higher-level semantic content, we use Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) to identify topics in the corpus. In this model, each topic has an associated word distribution, and each document (tweet) has an associated topic distribution; by inspecting the former, one can discern the “meaning” of each topic, and by inspecting the latter, one can discern the topical focus of a tweet or, aggregated over all an author’s tweets, the topical interests of that author. We pre-process tweets by removing punctuation, user handles, and standard English and Spanish stopwords - the latter because tweets in our corpus contain code-switching between English and Spanish. Additionally, we treat “Trump” as a stopword for this corpus since it is the search term from which conversations were collected. Running LDA for 10 topics\(^5\), we take as features the topic distribution of the current tweet (i.e., 10 features), as well as the topic distribution of the most recent tweet in the conversation by the candidate user.

Table 4.2 shows the top 10 words associated with each of the 10 topics derived from our corpus of tweets. Note that topics are arbitrarily numbered, and the labels presented here do not reflect a ranking. We can see that our collection of political

---

\(^5\)This number of topics was selected because it yields meaningful topics which are coherent to a human reader, but is relatively arbitrary; topic counts of 5, 15, or 20 all produce similar results.
tweets from October 2017 focused on stories such as emergency response in Puerto Rico (topics 3, 4, and 7), NFL players kneeling during the national anthem (topic 9), racism (topic 5), and comparisons between President Trump and Democratic leaders (topics 1 and 8).

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
<th>Topic 9</th>
<th>Topic 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>hope</td>
<td>love</td>
<td>people</td>
<td>good</td>
<td>people</td>
<td>thought</td>
<td>puerto</td>
<td>true</td>
<td>news</td>
<td>live</td>
</tr>
<tr>
<td>hillary</td>
<td>sad</td>
<td>pr</td>
<td>mayor</td>
<td>lol</td>
<td>evidence</td>
<td>rico</td>
<td>wrong</td>
<td>fake</td>
<td>usa</td>
</tr>
<tr>
<td>bot</td>
<td>big</td>
<td>money</td>
<td>day</td>
<td>black</td>
<td>funny</td>
<td>years</td>
<td>obama</td>
<td>real</td>
<td>war</td>
</tr>
<tr>
<td>agree</td>
<td>yeah</td>
<td>power</td>
<td>god</td>
<td>white</td>
<td>russian</td>
<td>lies</td>
<td>people</td>
<td>time</td>
<td>matter</td>
</tr>
<tr>
<td>cnn</td>
<td>people</td>
<td>dying</td>
<td>work</td>
<td>racist</td>
<td>means</td>
<td>people</td>
<td>president</td>
<td>flag</td>
<td>country</td>
</tr>
<tr>
<td>happen</td>
<td>dont</td>
<td>water</td>
<td>supplies</td>
<td>point</td>
<td>food</td>
<td>understand</td>
<td>vote</td>
<td>protest</td>
<td>marathi</td>
</tr>
<tr>
<td>states</td>
<td>person</td>
<td>tax</td>
<td>great</td>
<td>hate</td>
<td>read</td>
<td>party</td>
<td>thing</td>
<td>stand</td>
<td>tweeting</td>
</tr>
<tr>
<td>liar</td>
<td>blame</td>
<td>hurricane</td>
<td>job</td>
<td>guy</td>
<td>act</td>
<td>white</td>
<td>shit</td>
<td>talking</td>
<td>class</td>
</tr>
<tr>
<td>argument</td>
<td>wow</td>
<td>taking</td>
<td>san</td>
<td>bad</td>
<td>facts</td>
<td>rich</td>
<td>care</td>
<td>watch</td>
<td>leader</td>
</tr>
<tr>
<td>clinton</td>
<td>pr</td>
<td>days</td>
<td>juan</td>
<td>problem</td>
<td>helping</td>
<td>world</td>
<td>donald</td>
<td>anthem</td>
<td>place</td>
</tr>
</tbody>
</table>

Table 4.2: Top 10 words for each topic.

4.5 Results

4.5.1 Prediction Accuracy

Since responses are so rare, for a little under 90% of the observations one would predict a response correctly by simply guessing a non-response for every possible respondent. This sets a relatively high baseline for prediction, but we find that a straightforward logistic regression with the features described above significantly improves upon this baseline, achieving 94% out-of-sample accuracy\(^6\) in predicting exactly who among the previous participants in a conversation will and will not respond to a given tweet.

Using a support vector machine (SVM) (Boser et al., 1992), we are able to increase that accuracy to 98%, suggesting that there may be significant interactions and

\(^6\)Using an 80% in-sample, 20% out-sample split.
nonlinear effects among our features. Our SVM model is especially good at predicting non-responses, erroneously predicting a response when the truth was a non-response only about 1% of the time. Conversely, since there are so few responses, we more often erroneously predict a non-response, getting about 13% of the true responses wrong, although that only amounts to another 1% of the total sample. In all, about half of our errors are false 0s and half are false 1s, showing that the model does a very good job overall of predicting both when people will choose to respond, and when they will choose not to.

The gain in predictive accuracy suggests that the proposed models are good descriptors of the underlying mechanism which drives response behavior (Shmueli, 2010). That is, while there may be unknown correlates missing from our model, these features alone are enough to accurately explain the response outcomes we are seeing. Given that our features are selected and tested based on theory-based hypotheses, this predictive power supports the existence of our proposed causal mechanisms (Cranmer and Desmarais, 2017). To be clear, it is not enough to fully prove a causal claim, but the predictive power of our models is strongly suggestive of a causal mechanism in which our selected features are driving observed response behavior (Cranmer and Desmarais, 2017).

Tables 4.3-4.6 show the coefficients from the logistic regression model, since interpreting per-feature effects for SVMs is notoriously problematic. These coefficients indicate the strength of each feature’s predictive power: positive coefficients mean a response is more likely while negative coefficients mean a response is less likely. All features were scaled prior to fitting the logistic regression, meaning that the size of these effects are on the same scale.

However, identifying which features are “significant” for the logistic regression is a non-trivial problem with so many features. With 1,016,49 observations, by most
traditional measures of statistical significance, almost all of our features are statistically significant, regardless of the substantive magnitude of their effect. Even after multiple testing correction⁷ most coefficients are still significant.

However, in another sense, we do not have nearly as many observations as it may appear since, for any given tweet, very few choose to respond, and most responses are 0s. Furthermore, all tweet- and author-level conditions are shared across all the individuals who may or may not respond to that tweet. Thus it makes sense to cluster errors at the current-tweet level, reflecting the fact that the number of observations with variation in tweet- and author-level features is far fewer than the simple count of observations would imply. After doing this, approximately half of our features lose their significance, and even more do so if we run FDR correction after error clustering. However, from a prediction point of view, this may be going too far since our testing suggests that almost every feature – even if not significant by traditional statistical measures – does increase out-of-sample accuracy. This gap between the prediction and statistics literatures (Lo et al., 2015; Shmueli, 2010) goes beyond the scope of this article, so we present significance levels for all three corrections and focus on the cluster-corrected version in most cases as being the most prevalent approach in the social sciences.

As in the previous section, we discuss the feature effects by category, since each category speaks to a different family of hypotheses. But it should be reiterated that Tables 4.3-4.6 all derive from the same single logistic regression, and are only broken up for convenience.

⁷We use the Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 1995), setting a false positive rate of 10%.
4.5.2 Response Predictors: Candidate Respondent

Table 4.3 shows results for features pertaining to the candidate respondent who has previously participated in the conversation and now may decide whether to respond to the current tweet or not. At the general level, we find that, as expected in hypothesis $H1.2$, more popular users are less likely to re-engage even though they have done so already. Similarly, users are less likely to respond if they are verified or have more followers, although this effect is more fragile to cluster-correction. Converting these coefficients to odds ratios, we see that for every follower a candidate has, they are slightly less likely ($2.6e - 24$) less likely to engage. While this is a small individual effect, it captures the large disparity in candidate follower counts.

Interestingly, although it is also non-significant after clustered-error correction, while users who are generally more active on Twitter are more likely to respond, as predicted in hypothesis $H1.1$, a user who has been more active in a given conversation may actually be less likely to respond. We measure conversation activity (comments count) as the number of comments made prior to time $t$. In terms of the odds ratio, for every tweet a candidate has sent over their account’s lifetime, they are 0.5 times more likely to respond, while for every tweet a candidate has had within a given conversation, they are 0.5 times less likely to respond. While initially surprising at the user-level, this finding lends support to $H2.2$. That is, the difference in these effects may indicate that conversations have a natural ending point where users do not re-engage because they have nothing more to add, or that users may suffer from conversation fatigue - eventually getting bored or tired of engaging in the same back-and-forth.

Table 4.4 shows the effects for the candidate respondent’s previous tweet in the conversation. Note that in addition to the features discussed below, these tables also show our controls for time-varying effects using cyclic transformations of hours and days of the week, which control for periodicities such as the tendency to reply more
Table 4.3: Response predictors: Candidate Respondent

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>FDR</th>
<th>Clust</th>
<th>FDR+Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>verified</td>
<td>-0.625</td>
<td>***</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>followers count</td>
<td>-54.294</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>following count</td>
<td>-0.187</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>statuses count</td>
<td>0.026</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>favourites count</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comments count</td>
<td>-0.170</td>
<td>***</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

in the evenings or on weekends (Cox et al., 2006). Many of these features are strongly predictive of response even after various error corrections. In support of H1.3, users are significantly more (0.7 times) likely to respond if they were the author of the previous response (prev response), that is: if we are predicting the response to a tweet that was itself a response to the candidate user. Similarly, candidates who have not engaged in the recent conversation (time since prev), become less (0.3 times) likely to rejoin as time goes on. Taken together, these results may indicate that, while ostensibly a platform for multi-person conversation, dialogue on Twitter may be largely rapid-fire and dyadic in nature.

We find some evidence to suggest that candidates whose most recent tweet was longer (chars) are more (0.6 times) likely to return to the conversation. This may indicate that tweet-length reflects a user’s enthusiasm for the conversation or for tweeting in general. Conversely, having used a hashtag in their previous tweet is negatively related to a further response, perhaps either because the purpose of the first response was to promulgate the hashtag and that job is done, or because the interlocutor did not respond in kind, demotivating the candidate respondent.

We also see interesting effects around the emotions of a candidate’s most recent
tweet. Based on a tweet’s VADER score, it appears that, in support of \( H1.5 \), if a candidates’ most recent tweet was negative, they are more (0.7 times) likely to maintain engagement in a conversation while if their tweet was positive, they are less (0.4 times) likely to continue interacting. We will return to interactive emotional dynamics in Section 4.5.5.

Finally, in this corpus, users whose previous tweets focused on topics 8 and 10 are more likely to return to the conversation. Topic 8 seems to indicate negative views of the democratic party, while topic 10 may indicate a level of nationalistic pride. This suggests that people who engage with these topics tend to be more argumentative and less likely to let a debate go than the average Twitter user in our sample.

### 4.5.3 Response Predictors: Conversation

Table 4.5 summarizes the effects of the features of the conversation thread itself, specifically the number of participants up until \( t \) and the thread length, as measured by the number of tweets in the conversation at time \( t \). While, at first, it appears that thread length serves as a positive indicator of the interest in a thread (\( H2.1 \)), we do find this effect decreases for longer threads (\( H2.2 \)), as indicated by the quadratic thread length\(^2\) term. However, by setting the first derivative of the resulting curve to 0, we find that a reply is maximally likely for a thread length of 2 – the minimum possible in our dataset. This emphasizes the tendency for non-reply and suggests that the likelihood of a thread continuing decreases monotonically as a function of thread length as users lose interest, feel the conversation is exhausted, have difficulty viewing or following the conversation, or simply move on to other things.
Table 4.4: Response predictors: Candidate respondent’s previous tweet

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>FDR</th>
<th>Clust</th>
<th>FDR+Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>prev response</td>
<td>0.883</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>favorite count</td>
<td>-0.311</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retweet count</td>
<td>-262.234</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reply count</td>
<td>0.141</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quality</td>
<td>262.523</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>source</td>
<td>0.037</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>xday</td>
<td>0.169</td>
<td>***</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>yday</td>
<td>0.239</td>
<td>***</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>xhour</td>
<td>0.048</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yhour</td>
<td>0.193</td>
<td>***</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>chars</td>
<td>0.367</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>has url</td>
<td>0.037</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mentions</td>
<td>0.155</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hashtags</td>
<td>-0.078</td>
<td>***</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>sentiment</td>
<td>0.362</td>
<td>***</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>vader neg</td>
<td>0.641</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>vader pos</td>
<td>-0.313</td>
<td>***</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>valence</td>
<td>-0.084</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>arousal</td>
<td>0.151</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dominance</td>
<td>-0.174</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time since prev</td>
<td>-0.658</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>topic 2</td>
<td>1.853</td>
<td>***</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>topic 3</td>
<td>-0.037</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 4</td>
<td>-0.364</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 5</td>
<td>0.246</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 6</td>
<td>-0.536</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 7</td>
<td>-1.153</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 8</td>
<td>-2.787</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>topic 9</td>
<td>-0.573</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 10</td>
<td>2.404</td>
<td>***</td>
<td>**</td>
<td>*</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01
Table 4.5: Response predictors: Conversation features

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>FDR</th>
<th>Clust</th>
<th>FDR+Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>participants</td>
<td>-0.179</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread length</td>
<td>0.105</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thread length$^2$</td>
<td>-0.026</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
</tbody>
</table>

*Note:* *p*<0.1; **p**<0.05; ***p***<0.01

4.5.4 Response Predictors: Current Tweet & Author

Perhaps the most interesting predictors are those involving the current tweet that may or may not receive a response. This is the area in which our model extends beyond previous efforts (Artzi et al., 2012; Hong et al., 2013; Feng and Wang, 2013), in that we are examining not just engagement in a conversation, but responses and re-engagement at specific moments and in response to specific tweets.

The coefficient estimates for the tweet for which we are predicting replies are listed in Table 4.7, while the estimated coefficients for the features of that tweet’s author are displayed in Table 4.6. For the latter, we see that, as predicted in *H3.1*, users who are more active in the conversation, e.g., have contributed more comments, are 0.7 times more likely to receive replies to their tweets. While users with verified accounts and more followers may be slightly more likely to receive replies (0.5 times and 0.7 times, respectively), there is little evidence to support *H3.2*: that the popularity of a user mediates whether or not their tweets receive a response. This result may be in part because we focus only on users who have engaged at least once in the conversation, excluding the initial tweet. Thus while an extremely popular user or entity may kick off a discussion - posting a tweet that receives numerous responses - the people who actually engage in back and forth conversations seem to be less affected by the popularity of their interlocutors.
However, at best, author-level characteristics are indirect effects unless the respondent actually knows the author; the most interesting and direct effects – as well as those potentially most subject to manipulation by the tweeter – are via the tweet itself, as shown in Table 4.7. As before, we find that the day of the week has a strong effect, although the time of day has less of an effect after clustering. Interestingly, the source from which a tweet is posted - iPhone, Android, web interface, or third party software - also seems to be strongly predictive, with mobile users more likely to respond. This may suggest latent features of users or simply linguistic variation of tweets mediated by the platform interface. Users posting from digital devices, for example, may be more succinct in their tweets.

In contrast to the behavior predicted by H3.3, the previous reply count of a tweet may have a small, negative effect on whether a tweet will receive additional responses. For each potential reply, we consider how many replies, if any, had been made up to that point. The negative effect of this coefficient suggests diminishing returns to the number of replies a single tweet is likely to receive, and that, reasonably enough, one may feel a tweet has been adequately rebutted if many others have already replied. For each response a tweet has already received, however, it is only $4.2e^{-5}$ times less likely to receive an additional response, meaning that it can take many replies before a conversation moves on.

Interestingly, while tweets with more characters (chars) are 0.6 times more likely to receive replies, contra H3.4, mentions and hashtags were negative indicators for our dataset, making reply 0.4 times less likely and 0.5 times less likely, respectively. This may suggest that while longer tweets have more content to reply to, those who are overwhelmed by too many hashtags or ‘@’-mentions are less engaged. Additionally, in support of 3.5, we see that tweets with higher valence, e.g., a higher number of ‘pleasant’ words, are 0.4 times less likely to receive replies. This is similar to our previous result that if a potential respondent’s previous tweet in the conversation was
positive, they are less likely to reply again. Conversations about President Trump are, perhaps unsurprisingly, not just negative, but require negativity to persist.

Tweets focused on topics 3, 5, and 7, were more likely to receive replies. Two of these topics, 3 and 7, are both focused on the humanitarian crises in Puerto Rico, indicating that this was a particularly active topic of back-and-forth discussion in our sample. Topic 5 is more diffuse but points to issues of racism with words like “black,” “white,” and “racist” weighted within the top 5. This topic may or may not have been tied to Puerto Rico, but indicates another area of fervent debate around the president.

Table 4.6: Response predictors: Current tweet author

<table>
<thead>
<tr>
<th></th>
<th>Significance after p correction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef</td>
</tr>
<tr>
<td>verified</td>
<td>0.069</td>
</tr>
<tr>
<td>followers count</td>
<td>0.889</td>
</tr>
<tr>
<td>following count</td>
<td>0.334</td>
</tr>
<tr>
<td>statuses count</td>
<td>-0.031</td>
</tr>
<tr>
<td>favourites count</td>
<td>-0.202</td>
</tr>
<tr>
<td>comments count</td>
<td>0.637</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

Finally, the last two rows in Table 4.7 show the difference between the current tweet and the candidate respondent’s previous tweet. This difference is measured as the Euclidean distance between the topic vectors associated with a pair of tweets. In other words, it captures how topically different a tweet at time $t-1$ is from a tweet at time $t$. This is perhaps the most psychologically interesting variable since it speaks to a deep question about who we choose to converse with: those most like ourselves (presumably to agree), those most unlike ourselves (presumably to disagree), or something in between. We find that respondents are 0.5 times more likely to reply to comments very unlike their previous comment. Interestingly, the
### Table 4.7: Response predictors: Current tweet

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>FDR</th>
<th>Clust</th>
<th>FDR+Cl</th>
</tr>
</thead>
<tbody>
<tr>
<td>favorite count</td>
<td>1.657</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retweet count</td>
<td>9.171</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reply count</td>
<td>-10.055</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>quality</td>
<td>-43.260</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>source</td>
<td>-0.348</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>xday</td>
<td>-0.354</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>yday</td>
<td>-0.345</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>xhour</td>
<td>0.146</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yhour</td>
<td>-0.044</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chars</td>
<td>0.649</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>has url</td>
<td>0.075</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mentions</td>
<td>-0.412</td>
<td>***</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>hashtags</td>
<td>-0.083</td>
<td>***</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>sentiment</td>
<td>-0.135</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vader neg</td>
<td>-0.111</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vader pos</td>
<td>0.152</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>valence</td>
<td>-0.524</td>
<td>***</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>arousal</td>
<td>-0.116</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dominance</td>
<td>0.343</td>
<td>***</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>topic 2</td>
<td>0.815</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 3</td>
<td>2.140</td>
<td>***</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>topic 4</td>
<td>1.541</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 5</td>
<td>2.913</td>
<td>***</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>topic 6</td>
<td>1.024</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 7</td>
<td>2.669</td>
<td>***</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>topic 8</td>
<td>1.537</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 9</td>
<td>-0.148</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic 10</td>
<td>2.043</td>
<td>***</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>difference</td>
<td>-0.036</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>difference$^2$</td>
<td>0.188</td>
<td>***</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>

**Note:**  
* p < 0.1; ** p < 0.05; *** p < 0.01
estimated minimum reply likelihood is at a distance of 0.1, which is about 1.5 standard deviations below the mean reply distance. This suggests that while there is a slight tendency to respond to comments very similar to one's own (distance = 0), for most users, the likelihood of reply monotonically increases with distance.

Optimistically, this engagement across difference could be interpreted as people engaging in cross-cutting dialogue. Alternatively, it could be merely reflective of users who intentionally troll those whose views are different from their own. To investigate this further, we first examine the correlations between observed sentiment and Euclidean distance. If interactions over a larger distance are indeed merely trolling, we would expect sentiment to be more negative for distant tweets. However, we find no evidence for a correlation between sentiment and distance. Using the VADER compound score, the sentiment of the earlier tweet in a pair has only a 0.04 correlation to distance, while the later tweet in a pair has 0.02 correlation to distance. These null effects are consistent across included sentiment measures.

The lack of correlation suggests that there may be more than mutual trolling going on in these conversations, and we therefore further hand-code a random collection of tweet pairs. For our 1 million observations, we randomly sample 100 pairs from the bottom 1% of closest replies (difference < 0.4) and 100 pairs from the top 1% of furthest replies (difference > 0.85). Note that the mean distance is 0.7, with a standard deviation of 0.2. These dyadic tweet conversations are coded along two dimensions: whether the authors appear to be in agreement with each other and the sentiment of the later tweet in the conversation. These are independently coded along a 3-point scale: agree (1), neutral (0), disagree (-1), and positive sentiment (1), neutral (0), and negative sentiment (-1). This creates nine possible sectors into which a given tweet pair can fall.

Our findings paint a complex picture of political engagement on Twitter. As we see
in Figure 4.3, our measure of topical difference captures notably divergent types of behavior. Though neither end of the distance extreme describes a single type of conversational engagement, difference does seem to capture user agreement. In our sample, subjects agreed with each other in nearly two-thirds (65%) of close tweets and less than one quarter (23%) of topically different tweets. In other words, independent of sentiment, users who agreed tended to “stay on topic,” while users who disagreed tended to shift topics. However, user agreement is not consistently tied to a single sentiment, though we do see notable patterns when these features are considered together. Specifically, close users who agree are roughly evenly split between tweet pairs that are positive in tone (24%) and negative in tone (28%). This suggests at least two distinctive types of “agreeing” behavior. People may use positive sentiment to affirm each other’s views, or they may use negative sentiment as they mutually abuse “the other side.” Furthermore, a small portion of hand-coded tweet pairs (15%) appear to have the
potential to be productive, political talk: these tweets are on the same topic, reflect users who disagree with each other, and have a positive sentiment.

Looking at tweet pairs that are further away from each other further reveals two additional types of behavior. This topic-shifting may indeed indicate the use of negative sentiment and disagreement (35%), but it may also indicate purely inane responses (such as tagging additional accounts) to neutral topics (32%). The full results in Figure 4.3 suggest even that more types of sentiment and agreeing behavior occur on Twitter, though possibly to a lesser degree. Together, these findings suggest the need to consider conversations as capturing multi-dimensional behavior that cannot easily be classified into simple categories of “trolling” or “agreement.”

### 4.5.5 Contents of Responses

Although this project focuses mainly on the decision to reply, rather than on the content of replies, we can briefly examine content interactions between a current tweet and its replies. These results are mainly suggestive at this point since we do not embed this within a nested model that also controls for the decision to reply as a first stage.

Figure 4.4 provides heat map illustrations of correlations between tweets and their replies, both on topics (left) and emotions (right). Red indicates positive correlations, while blue shows negative correlations. Topics are sorted by the first eigenvector of the correlation matrix in order to cluster similar topical or emotional response patterns together.

On topics, we see that for this corpus, the debate around NFL players kneeling during the national anthem seems to be highly contained, with tweets on this topic (topic 9) often receiving responses on the same topic. On the other hand, the conversation
about disaster relief following a devastating hurricane in Puerto Rico seems to be more diffuse. For example, topic 4, which appears to be largely about the mayor of San Juan, is somewhat correlated with itself but is more strongly correlated with topic 1 - a topic focusing on another democratic, female politician: former Secretary of State Hillary Clinton. Tweets primarily engaged in topic 5, which seems to focus on racism, most commonly receive responses focused on topic 6. While somewhat less coherent than other topics, it is telling that topic 6 is most commonly followed by topic 4. Since topic 4 praises the work of the mayor of San Juan, this may suggest our corpus contains conversations in which users are arguing about whether disaster response in Puerto Rico is tied to racism.

Additionally, there are negative correlations between topics. Topic 1 and topic 8, for example, seem to represent different views on the Democratic party. Topic 8, which potentially expresses negative views of the party, has a strong negative correlation with topic 1, which seems to express positive views on the subject. Unsurprisingly, negative judgments of the Democratic party are rarely met with positive ones as a response, but instead some other form of negative attack.

We can also see some of these dynamics in the purely emotional content of tweets and replies. Tweets with a positive VADER score are likely to receive replies which also have a positive VADER score, even if they are less likely to receive a reply at all (as we saw earlier). Tweets scored as neutral (rather than negative) are least likely to follow positive tweets, suggesting that many of these conversations consist of like-minded people reinforcing each other’s beliefs or using charged language for anyone who disagrees. Tweets that score high on the dominance measure are most highly correlated with arousal, indicating that words of strength and power are met with words of excitement - either to eagerly agree or to voraciously disagree. Neutral tweets are met with neutral tweets and are unlikely to elicit a positive response, but are also presumably less likely to receive any reply at all.
4.6 Discussion

To summarize our results, we find that a number of user-, thread-, and tweet-level features are critical in predicting the dynamics of online conversations. While previous studies (Backstrom et al., 2013; Artzi et al., 2012; Hong et al., 2013; Feng and Wang, 2013) have primarily focused on predicting post-level engagement for the purposes of algorithmic content curation, we predict comment-level engagement as users exit and re-enter a conversation. Particularly novel is our inclusion of features pertaining to the individual tweet that may be responded to, particularly the emotional and topical content of those tweets. Our logistic regression model predicts user response remarkably well, achieving 94% out-of-sample accuracy. The 98% accuracy achieved by our SVM model suggests that there may be further non-linear and interactive effects among our features to explore in later work, perhaps via additional machine learning methods (such as random forests or deep neural networks), although at 98%, we are already near the ceiling of predictive accuracy.
Our constructed corpus is intentionally centered around highly charged political content, and we would not necessarily expect these findings to hold in a more general context. Furthermore, while we have restricted our user analysis to those who have engaged in conversation, these users are still likely to represent a heterogeneous mixture of entities, including everyday citizens, celebrities, and brands. These findings are, therefore, not necessarily reflective of political talk simply between everyday citizens but can be better understood as a small slice of the public sphere in which a range of entities engage.

We find support for many of the hypotheses outlined in Section 4.4.2. In support of $H1.1$, $H1.2$, and $H3.1$, we find that the features of both candidate respondents and current tweet authors have small but important effects on predicting response. Interestingly, in contrast to $H3.2$, we find that the popularity of a tweet’s author has little effect on predicting whether or not a tweet will receive a reply. Because we only include users who have been active in a back-and-forth exchange, this suggests that conversations on Twitter are relatively free from the sort of social influence we may have found if we were examining replies to the initial tweet of a thread.

In support of $H2.2$, we find that longer conversations are decreasingly likely to receive replies. Similarly, while in $H1.3$, we expected a candidate’s previous activity in a thread to predict additional activity, we found that users who have already offered numerous comments are less likely to re-engage. Taken together, this suggests that the time, attention, and cognitive energy that goes into participating in a back-and-forth exchange leads to a natural cut-off where conversations, though popular at first, become too much effort to continue or have their subject matter exhausted.

Additionally, and in support of $H1.4$ and $H3.5$, we find that the emotional and topical contents of tweets seem to play a significant role in driving the continuation of conversation. Users with negative-sentiment tweets are more likely to re-enter
conversations and tweets with fewer pleasant words are more likely to receive a response. While a small corner of our corpus may be primarily engaged in positive-to-positive conversations, it seems that the vast majority of Twitter dialogue around President Trump consists of acrimonious argumentation. This is reinforced by the finding that candidates are more likely to respond to tweets unlike their previous tweet. This could optimistically be interpreted as people engaging in dialogue across difference, but could just as well be mutual trolling – though if the latter, at least we do observe extensive repeated interactions rather than simple one-off attacks.

Through sentiment analysis and the hand-coding of tweets, we find evidence for both of these behaviors as well as more complex user interactions. Users may troll those they disagree with or may agree with negative commentary about “the other side.” Users may further engage in positive discussions with those they already agree with, and, in some cases, may even engage in positive discussions with people with whom they disagree.

These findings paint a picture consistent with what some avid Twitter users might expect: at least when it comes to political dialogue around a controversial figure, most conversations are emotionally charged and negative in tenor. While such conversations fall far below the ideals of democratic deliberation (Habermas, 1984; Cohen, 1989; Dewey, 1927), our findings suggest this may not be the end of the story. First, within the broader deliberative system, it is commonly acknowledged that many “everyday” conversations will frequently fail to meet deliberative ideals (Mansbridge, 1999). Nevertheless, these conversations may still play an important and positive role in expanding people’s viewpoints and encouraging the refinement of beliefs (Mansbridge, 1999; Huckfeldt et al., 2004). Despite the predominately negative tone of our corpus, we do find that users are responding to topics outside their own talking points and that users who are active in a conversation are more likely to receive a response. In other words, conversations are happening, and those conversations do not appear to
be strictly confined within partisan bubbles.

Furthermore, it is difficult to fully characterize the democratic value of a conversation based on sentiment analysis alone. For example, within our corpus, we find that positive tweets are most likely to receive positive-sentiment responses. While on the surface, this may suggest a collection of more civil exchanges, it is also possible that positive-to-positive conversations represent little more than in-party affirmation, with little deliberative value. The deliberative ideal imagines citizens as ‘tolerant gladiators’ (Huckfeldt et al., 2004), who fight with strong words but who emerge from confrontation as friends. Our corpus finds evidence that there is no shortage of strong words - rather, it is the long-term effects of these conversations that remains to be seen.

Taken together, these findings suggest that within this corpus, citizens are participating in rich and engaging political conversations. While the extent to which these conversations support democratic ideals remains to be seen, if we wish to extend and enrich these interactions, we should seek to broadly increase conversational activity online, developing tools to make it easier to engage and follow long threads. Additionally, given users’ willingness to respond to those unlike themselves, these findings suggest that there is value in adding noise to recommender systems – showing users new and different content, rather than overfitting recommendations based on the content with which they have already interacted.

In future work, we intend to analyze the dynamics of political conversations over time, looking at sentiment and opinion flows through whole threads of conversation, and we are developing a model to infer the role of latent ideology in these exchanges. In this paper, we have identified a number of key factors predictive of conversation engagement and shown that these conversations are systematic and highly predictable, reflecting a complex interplay of circumstance, topic, and emotion.
Chapter 5

Conclusion

The full deliberative system is a richly interconnected framework of formal and informal political talk. Such a system is the backbone of the democratic ideal – providing venues through which citizens can clarify their thinking, share information, advocate for their views, and reason together. In the ideal, such a system facilitates the generation, aggregation, and filtering of ideas – allowing all voices to be heard and the best solutions to rise to the surface. This ideal does not rely upon some fantastical notion of perfect, ideal citizens, but rather imagines a complex system of continual change and gradual improvement. Understanding this full system, then, requires a more detailed understanding of its component parts – each of which have their own networked dynamics.

This dissertation has studied three elements of the deliberative system: the exchange of arguments, the structure of individual-level reasoning, and the dynamics of ongoing conversation. Each of these elements have their own networked dynamics, and each chapter has aimed to build methods and models to strengthen our understanding of these systems. Like any complex system, the full deliberative system cannot be
wholly understood through the elaboration of its pieces, and yet this work provides tools for measuring and improving our collective deliberative life. Each paper presents a conceptual framework for understanding these the interconnections of systems and provides insights on how we can work towards meeting the democratic ideal of a functional deliberative system.

5.1 Giving and asking for reasons

One key challenge of democratic life is that differing normative interpretations mean that people with the same information will not necessarily agree on what actions are best. Chapter 2, therefore, examines the process of giving and asking for reasons within a normative setting of policy implementation. The model imagines a small group of deliberators considering a set of policies which are all aimed at addressing some overarching social issue, such as education or healthcare. Before voting on which policies to implement, agents engage in discussion close to the deliberative ideal – sharing and receiving information in good faith, aiming to convince others and to be convinced when appropriate (Mercier and Landemore, 2012). Empirical evidence suggests that such an exchange could influence real-world policy votes (Fishkin, 2014; Knobloch et al., 2013; Neblo et al., 2010).

This chapter specifically examines three elements of this piece of the deliberative system: the effects of cognitive capacity, ideological factioning, and open-mindedness. Each dimension examines the influence of a canonical deliberative failure. People may be less likely to come to good decisions if humans have limited cognitive capacity, if they are factioned and unable to engage with people unlike themselves, or if they are too set – or too flexible – in their opinions. We examine these effects through groups comprised of uninformed, intellectual, or ideologue agents who operate with differing rules for adopting others’ beliefs.
The key finding is that factioned groups do surprisingly well at identifying optimal policy solutions. While uninformed agents perform as expected – achieving worse outcomes as cognitive capacity decreases – groups composed of oppositely-skewed ideologues appear to be resilient to the effects of declining cognition. This finding is in line with deliberative theory (Mansbridge, 1999) and suggests that heterogeneous agents can achieve good outcomes if they are able to engage in good-faith discussions.

This, of course, is a very restrictive assumption – very often, people do not engage in good-faith exchange and are more concerned with winning than in collaboratively discovering the truth. In such settings, factions would likely lead to entrenchment rather than optimal outcomes. What this finding suggests, though, is that the presence of opposing factions may not itself be the biggest concern. Indeed, having access to a diversity of opinions has the potential to lead to better outcomes than might be achieved otherwise. This benefit can only be realized, however, if we are able to build systems, institutions, and interventions which help people truly listen to and learn from each other.

The chapter further finds that this ideological effect does not appear to be moderated by our other parameters of interest. That is, while ideologues do far better than uninformed agents, their performance is not significantly improved by the presence of highly cognitive intellectuals or by being increasingly open to, or skeptical of, others’ views. This suggests that the main factor in driving ideologues’ performance is their inherent counter-balance to each other; by constantly pushing the “other side” to be better, both groups, and ultimately the whole, can improve.

It is particularly interesting to note that the conditions under which agents accept each other’s views have little effect on observed outcomes. While we would see entrenchment if agents were wholly unwilling to consider opposing views, even a modest openness to others’ opinions can lead to better outcomes.
Taken together, these results suggest a need to further invest in studying and creating deliberative spaces. The deliberative ideal is indeed lofty – and we should not expect that every person in every conversation will fulfill the ideals of good-faith reasoned exchange (Mansbridge, 1999). However, these results suggest that we do not have to. A world in which discussion and debate lead to better policy outcomes does not have to be perfect – it does not have to rely on highly cognitive people wholly free of partisan bias. It is okay for people to be imperfect and to be flawed in their thinking – but only if they are modestly willing to listen to each other.

5.2 The structure of reasoning

Chapter 3 examines the structure of individual-level reasoning. Such structure can be observed through linguistic expressions – either through speech or written words (Toulmin, 1958; Walton, 1996; Quillian, 1967; Axelrod, 1997) – but may reflect underlying cognitive processes (Collins and Loftus, 1975; Quillian, 1967; Shavelson, 1974; Shaffer et al., 2009). A notable line of classic work in public opinion argued that this structure has the potential to provide insights into variations of political behavior (Lane, 1962; Axelrod, 1976; Campbell, 1960). Within the context of the deliberative system, such structure reflects the process of self-reflective “deliberation within” (Goodin, 2000) and holds the potential to influence how messages are expressed and received during interpersonal dialogue. That is, if we take the observed structure to be a reflection of cognitive processes, it brings insight to the individual reasoning process. If we take it to be merely a semantic representation, it allows us to separate the influence of content – what someone says – from structure – how someone says it.

While past work has relied upon hand-coding of text or semi-structured interviews, this chapter presents a computational, text-based method for inferring conceptual network structure. This method relies upon the grammatical parse of a text to
identify which expressed concepts are connected and leverages word embeddings in order to identify which words point to the same concept.

Using a sample of 100 Mechanical Turk respondents, we find that this method produces structures which meaningfully correlate to known personality traits. Specifically, conservative respondents and those who score high on traditionally conservative personality measures – such as purity and authority – tend to produce network structures that are sparse and more likely to be disconnected. Liberal respondents, on the other hand, are more likely to produce interconnected networks with a more hierarchical structure. While the conceptual network structure inferred for liberal respondents is reflective of certain liberal traditions that would claim these structures to be morally superior (Rorty et al., 1989), it is too soon to make any claims as what constitutes a “good” conceptual network structure. Given the method’s reliance on grammatical structure, text that results in a sparse network could simply be reflective of non-standard English rather than deep moral views. The approach of using grammatical structure is still a valuable tool in assessing underlying cognitive connections, but we should be wary of making normative assessments based on grammatical style.

Furthermore, using a secondary dataset in which nearly 1,000 respondents were asked to argue both the “liberal” and “conservative” position on a topic, we demonstrate that this method captures a “reasoning fingerprint” of individual expression and reasoning quality. The inferred network structure can seemingly pick up on latent patterns of argument quality that are missed by coarser textual measures, and these patterns appear to be driven more by individual style than by ideological content.

Together, these findings suggest that people do express themselves in individually distinctive ways. There appears to be a meaningful behavioral signal in the structure of these expressions, independent of the content of ideas being expressed. Examining
this structure, and developing models capable of incorporating both structure and content, could, therefore, be a fruitful line for further study, with the potential to bring better understanding to both individual reasoning processes as well as attempts to reason together.

5.3 Dynamics of ongoing conversation

The final empirical chapter, chapter 4, examines the dynamics of real-world conversations. We look particularly at online conversations, which play an increasingly important role in the public sphere. This setting is perhaps one of the most common venues for modern “everyday talk,” a critical element of the deliberative system. This chapter presents a framework for interpreting threads of asynchronous conversation and demonstrates that a number of user-, thread-, and tweet-level features are critical in predicting the dynamics of online conversations.

Specifically, we find that features of both candidate respondents and current tweet authors have small but important effects on predicting response. Interestingly, the popularity of a tweet’s author has little effect on predicting whether or not a tweet will receive a reply, suggesting that once users choose to engage, ongoing engagement does not appear to be driven by popularity. We further find that users who have already offered numerous comments are less likely to re-engage and that longer conversations are decreasingly likely to receive replies. This suggests that the time, attention, and cognitive energy that goes into participating in an ongoing exchange leads to a natural conversation drop off.

Additionally, we find that the emotional and topical contents of tweets seem to play a significant role in driving the continuation of a conversation. Users with negative-sentiment tweets are more likely to re-enter conversations and tweets with
fewer pleasant words are more likely to receive a response. While a small corner of the corpus is primarily engaged in positive-to-positive conversations, it seems that the vast majority of Twitter dialogue around President Trump consists of acrimonious argumentation. This is reinforced by the finding that candidates are more likely to respond to tweets unlike their previous tweet. Hand-coding a sample of tweets on dimensions of agreement and sentiment reveals varied and nuanced dynamics. Conversations with negative sentiment may reflect users who are trolling those who disagree with them – or, it may capture users discussing their mutual distaste for “the other side.” Conversations with positive sentiment are most likely to reflect mutual affirmation, but may, in some cases, reflect people genuinely engaging in dialogue across their differences.

While these findings paint a picture of engagement that falls far below the ideals of democratic deliberation (Habermas, 1984; Cohen, 1989; Dewey, 1927), they are notable in that they do show engagement. Not every conversation within the deliberative system needs to meet deliberative ideals (Mansbridge, 1999), and indeed we would generally not expect this ideal from a corpus of tweets about President Trump. Nevertheless, these conversations may still play an important and positive role in expanding people’s viewpoints and encouraging a refinement of beliefs (Mansbridge, 1999; Huckfeldt et al., 2004). Despite the predominately negative tone of the corpus, users appear to be engaging in cross-cutting dialogue and actively engaging in political conversation.

Taken together, these findings suggest that there is a rich and engaging world of everyday political talk taking place in online spaces. While the extent to which these conversations support democratic ideals remains to be seen, if we wish to extend and enrich these interactions, we should seek to broadly increase conversational activity online, developing tools to make it easier to engage and follow long threads. Additionally, given user’s willingness to respond to those unlike themselves, these
findings suggest that there is value in adding noise to recommender systems – showing users new and different content rather than overfitting recommendations based on the content with which they have already interacted.

5.4 Future work

While this dissertation has examined several key elements of the deliberative system, there is much which remains to be done. This includes developing models that incorporate reasoning structure along with content, as well as examining the interactions between reasoning structure and conversational dynamics. Each of these chapters could expand into a dissertation of its own, each capturing just one element of the full, complex, deliberative system.

Specifically, future work will expand the analysis of real-world, multi-party conversations. While simulations can bring insight to the macroscopic effects of localized decisions, reasoning or “deliberation within” is an important piece of the deliberative system, and conversational structure can reveal a lot about deliberative engagement, conversations between people – real-world attempts to argue, persuade, learn, and decide – are the ultimate elements of interest. The deliberative system cannot be understood without appropriate methods and measures for considering its constituent parts, and yet there is so much more to do if we are to work towards the democratic ideal of deliberation.

Further missing from this current work is a proper accounting of the deliberative system itself. This dissertation has described “deliberation” as a complex system that engages many levels of citizens in many types of conversations. However, both here and within the deliberative literature, accounts go little beyond that. Just what is the deliberative system? How do we know when its working and how can we best aim
to improve it? The common agreement is that the current state of political discourse is broken, yet we cannot truly begin to mend that without knowing what we are trying to fix.

This dissertation began by arguing that collaborative reasoning lies at the core of democracy, and it will end by arguing the same. Whether we like it or not, we are forced to share this world with each other. And whether we believe it or not, we each have the power to shape the communities we inhabit. Democracy is the art of associated living, the art of doing your best to do what is right. Democracy is the belief that we must work together, that no person is an island. We all have something to learn, and we all have something to share. Indeed, democracy is not a system of government – it is the firm commitment to lifting each other’s voices, to listening when you need to listen, and speaking when you need to speak. It is the conviction that – flawed and imperfect though we may be – reasoning together is the only path forward.
References


Levine, P. (2015). *We are the ones we have been waiting for: the promise of civic renewal in America*. Oxford University Press.


Madison, J. (1787). Federalist no. 10: ”the same subject continued: The union as a safeguard against domestic faction and insurrection.”. *New York Daily Advertiser.*


158


Appendices

6.1 Chapter 2 Appendix

Chapter 2: Collaborative Reasoning on Value-Laden Topics: A Game of Giving and Asking for Reasons presents an agent-based model of collaborative reasoning on value-laden topics. This appendix reviews model details and presents additional analyses that demonstrate that the core findings presented in the paper hold across a range of parameter inputs. Full replication materials can be found online at https://github.com/sshugars/collaborative_reasoning

6.1.1 Model Parameters

Like all agent-based models, this model has several parameters that jointly described the system and agent behavior. All parameters are described below, along with a summary of their role and treatment within this paper. Section 6.1.3 goes into additional detail about the parameters investigated in this study. Section 6.1.4 concludes by discussing the parameters which are considered fixed and the limitations of this study.
1. **Solution space complexity:**

   - $N$: Number of policies under consideration. All simulations presented in this paper use $N = 4$.
   - $K$: Number of policies a given policy is influenced by. This parameter governs the complexity of the terrain. All simulations presented in this paper use $K = 2$.
   - $A$: Number of states in which a policy can exist. All simulations in this paper use $A = 2$. While, in principle, the model can support an arbitrary number of states, most $NK$ models fix the value of $A$ at 2 and assume binary on/off states.
   - **true distribution**: The distribution from which ground truth solutions are drawn. Across all simulations, this is assumed to be a uniform distribution.
   - **true bounds**: The range of possible ground truth influence values. This is fixed at $[-10, 10]$ across all simulations. While, in principle, these values could be modified, it is most interesting to consider true noise in relation to agent noise, described below. This paper, therefore, fixes the true bounds while sweeping the agent’s noise parameter.

   Together, these parameters describe a system of $A^N$ possible solutions, where each of $N$ possible items is influenced by the state of $K$ other items. The values of this influence are drawn from **true distribution** with bounds **true bounds**.

2. **Simulation parameters:**

   - **min_noise**: The minimum amount of noise to add when initiating agent networks. This number is bounded at 0. For all simulations in this paper, **min_noise** is fixed at 1. This discards the trivial case, where agents are initialized directly to the ground truth, and begins a sweep of agent’s noise parameter.
   - **max_noise**: The maximum amount of noise to add when initiating agent networks. For all simulations in this paper, **max_noise** is fixed at 210. This is 20 times larger than the bounds on the ground truth influence values $([-10, 10])$, resulting in agent networks that are essentially random noise.
• **runs**: number of simulations to run at a single noise level. This value is fixed at 100 for every simulation in this paper.

• **max_iter**: The maximum number of timesteps (speech acts) to allow within a single run. Simulations will stop earlier if a group reaches the ground truth solution. For every simulation in this paper, this value is fixed at 5000. Anecdotal investigation found that after that many timesteps, simulations had typically identified the true solution or reached a steady-state with no further changes.

A full simulation sweeps from **min_noise** to **max_noise**. At each noise level, a ground truth solution is initialized once. Agents are initialized **runs** times. A single run ends when agents have discovered the ground truth solution or after **max_iter** timesteps.

### 3. Agent parameters

• **agent_distribution**: The distribution from which agent’s perceived influence values are drawn. Across all simulations, this is taken to be a uniform distribution for all agents.

• **over**: The upper bound on the amount of noise that may be added to a specific agent’s initialized network. This varies by agent type, which will be discussed in detail in Section 6.1.3, but never exceeds **max_noise**.

• **under**: The lower bound on the amount of noise that may be added to a specific agent’s initialized network. This varies by agent type but is never less than **min_noise**.

For every ground truth influence value \( v \), an agent is initialized with belief \( v + \text{agent_distribution}([\text{under}, \text{over}]) \).

### 4. Group parameters

• **n_users**: Number of agents in a simulation. The paper presents results for groups of 25 agents. Results for groups of 5, 15, and 25 agents are presented in Section 6.1.3 below and yield similar results as those presented within the paper.

• **openness**: The cosine similarity required to accept a proposed view. This value is bounded between -1 (always accept) and +1 (always reject).
The paper presents results for groups with openness of -0.75, 0, and +0.75. Additional results illustrating the consistency of those findings are presented in Section 6.1.3.

• agent_types : The types of agents participating and the number of agents of each type. Agent types provide differing bounds on agents’ initialization conditions. The specific types of agents (uninformed, intellectuals, ideologues) will be discussed in Section 6.1.3 along with variations on these types.

For every run in runs, a group of n_users agents are initialized. Each agent is initialized with noise bounds determined by agent_types. Agents then accept a vector of shared beliefs if the cosine similarity between the speakers’ vector and the listeners’ vector is greater than openness. This is described further in Section 6.1.2.

6.1.2 Simulation Steps

Given the above parameters, a full simulation proceeds in the following steps:

1. Select noise level n in range [min_noise, max_noise].

2. Initiate ground truth solution, an NK model with A states. Influence values v are drawn from a true_distribution with bounds true_bounds. These values determine the ground truth value of each of A^N possible solutions.

3. For runs times:
   (a) Initiate n_users of agent_types.
   (b) For max_iter timesteps or until the group identifies the optimal solution:
      i. Randomly select one of n_user agents to share a belief. This agent will be the speaker.
      ii. Randomly select one of N polices to share a position on.
      iii. Randomly select one of A states to discuss.
iv. Determine belief vector for speaking agent \((\text{speaker\_vector})\). This \(A^K\) length vector describes every other policy’s influence on the selected policy being in the selected state.

v. Compare the \(\text{speaker\_vector}\) to each agent’s own \(A^K\) length vector of beliefs \((\text{listener\_vector})\).

vi. All agents for whom \((\text{speaker\_vector} \cdot \text{listener\_vector}) > \text{openness}\) moves towards the proposed beliefs as follows:

\[
\text{new\_beliefs} = \text{listener\_vector} + \left( (\text{speaker\_vector} - \text{listener\_vector}) / 2 \right)
\]

vii. All agents hold an up or down vote on each of \(N\) policies. Each agent selects their preferred policy platform (e.g., the platform with the highest value) and votes for each policy to be in their preferred state.

viii. The implemented state of each policy is decided independently by majority vote, and the “quality” of the implemented policy platform is measured against the ground truth value of that platform.

4. Repeat from Step 1 until \(n > \text{max\_noise}\).

6.1.3 Explored Parameters and Findings

This paper focuses largely on manipulating the group parameters of \(\text{n\_users, openness, and agent\_type}\). This section first demonstrates the robustness of findings for differing values of \(\text{n\_users}\) and \(\text{openness}\) and concludes with a discussion of different \(\text{agent\_types}\).

While there are many system-level parameters which could further be explored, this paper takes those items as fixed in order to focus more closely on group dynamics and possible failures of deliberation. These static parameters and the limitations of this approach are discussed in Section 6.1.4.
Number of Agents

This model is intended to capture the dynamics of small group deliberation. While $n_{\text{users}}$ could be any arbitrarily large, positive integer, we focus here on relatively small numbers of agents. All results presented in the paper are for groups of 25 agents, but these findings are robust across smaller group sizes. Figures 6.1-6.3 below compare findings for groups of 5, 15, and 25 agents and illustrate that these results are not significantly changed by group size. Figure 6.1 shows this consistency for uninformed agents, while Figure 6.2 shows intellectual agents, and Figure 6.3 shows ideological agents.

![Figure 6.1: Uninformed: Comparing the Number of Agents](image)

(a) Percent of good policies enacted.

(b) Distance from optimal policy platform.

(c) Percent of agents in the largest coalition.

Figure 6.1: Uninformed: Comparing the Number of Agents
Figure 6.2: Intellectuals: Comparing the Number of Agents
Figure 6.3: Ideologues: Comparing the Number of Agents
Openness

The openness parameter governs how close (in cosine similarity) a shared belief needs to be in order for an agent to accept it. This parameter is taken to be shared across all agents in a simulation. The paper contains results for skeptical (openness = 0.75), moderate (openness = 0.0), and open (openness = -0.75) levels of this parameter. Figures 6.4-6.6, show that these results are representative. As we would expect, the openness level has a direct, monotonic effect on the coalition behavior of uninformed agents (Figure 6.4). When agents always accept shared information (openness = -1), they ultimately form a single coalition. When they never accept shared information (openness = +1), they remain factioned into multiple coalitions. This coalition behavior does not significantly influence the ability of uninformed groups to come to good policy decisions. Again, this is as we would expect – since uninformed agents hold essentially random views, the sharing and acceptance of information neither helps nor harms their group knowledge.

In Figure 6.5, we see that intellectual agents perform approximately equally well across all openness levels. The exception to this finding is only at the extreme when agents never accept information (openness = +1). At this extreme, intellectual agents still easily outperform their uninformed counterparts, but, as we would expect, they do experience some factioning behavior which has the potential to lead to worse policy implementations.

Finally, in Figure 6.6, we see consistent results for ideological agents. The outlying case here is on the acceptance side, where ideologues who always accept information (openness = -1) perform more like uninformed agents than like their skeptical ideologue peers. This, again, is to be expected and emphasizes the paper’s main finding that the counterbalancing of factioned groups can lead to better decisions. If these factioned agents always agree, then they no longer provide balance and ultimately make worse decisions then they would otherwise.
Figure 6.4: Uninformed: Comparing Different Openness Levels (10 Agents)
Figure 6.5: Intellectuals: Comparing Different Openness Levels (10 Agents)
Figure 6.6: Ideologues: Comparing Different Openness Levels (10 Agents)
Agent Types

This is the most complex parameter in the model and the most difficult to explore. Broadly, agent types captures both the bounding conditions on agents and the group composition of those bounding conditions. Note that across all simulations, the given bounds are taken to describe a uniform distribution.

Chapter 2 present the analysis of the following agent types:

- **Uninformed**: Noise added to ground truth values is bounded by \([-n, n]\) for noise level \(n\).

- **Intellectual**: Noise added to ground truth values is bounded by \([-1, 1]\) for all noise levels \(n\).

- **Ideologue**: Ideologue agents come in two types:
  - Positive-Skew Ideologues: Noise added to ground truth values is bounded by \([((n - 2) \times \text{sign}(v), (n) \times \text{sign}(v))]\) for noise level \(n\) and ground truth value \(v\).
  - Negative-Skew Ideologues: Noise added to ground truth values is bounded by \([-n \times \text{sign}(v), (-n + 2) \times \text{sign}(v)]\) for noise level \(n\) and ground truth value \(v\).

Note that the sign of ground truth value \(v\) is included in order to allow for non-parametric interactions, which, for high levels of noise, prevent ideologues from simply being optimists who think all interactions are good or pessimists who think all interactions are bad.

Using these agent types, Chapter 2 explores the following group compositions:

- All uninformed: All \(n\) users agents are initialized using uninformed bounds.

- All intellectuals: All \(n\) users agents are initialized using intellectual bounds.

- All ideologues: \(n\) users are split evening among positive_skew and negative_skew ideologues. If \(n\) is odd, the positive_skew coalition is arbitrarily always one agent larger.
- Majority ideologues: If \( n_{\text{users}} \) is odd, these groups have one intellectual agent. If \( n_{\text{users}} \) is even, these groups have two intellectual agents. The remaining \( n_{\text{users}} \) are split evenly between \text{positive skew} and \text{negative skew} ideologues.

- Majority intellectuals: Half of agents are intellectuals, rounding up in systems where \( n_{\text{users}} \) is odd. The remaining \( n_{\text{users}} \) are split evenly between \text{positive skew} and \text{negative skew} ideologues.

A full sweep of the parameter space would require a detailed investigation of different agent types as well as different group compositions. In this paper, the specific agent types (uninformed, intellectual, and ideologue) were selected to test specific deliberative challenges; specifically concerns about everyday citizens being too cognitively limited to hold well-informed opinions (Lippmann, 1922) and about polarized groups being unable to properly coordinate on solutions (Dworkin, 2006; Madison, 1787). While these agent types are limited compared to real-world deliberation, they do offer meaningful insight on the specific research questions being addressed. Therefore, the additional analysis here takes these agent types as fixed and examines different group compositions, showing that the groups presented within Chapter 2 are representative of a range of group compositions.

For example, we can look for a tipping point of non-intellectual agents. While intellectuals alone perform perfectly by design, we can evaluate the performance of groups that have more than half intellectuals, the composition presented in the paper. In Figures 6.7-6.8, we see that, as we would expect, reducing the number of intellectuals leads to a monotonic decrease in group performance.

Figure 6.7 shows the transition from a group of intellectual agents to a group composed entirely of uninformed agents. Here we see a notable drop in group performance once a single uninformed agent is added to the mix. This performance continues to decline as more uninformed agents are added, though it quickly reaches a steady state. Groups in which more than half the agents are uniformed are virtually indistinguishable from
Figure 6.8 demonstrates a similar pattern for groups which transition from intellectuals to ideologues. While extremes – groups entirely of intellectuals or entirely of ideologues – display markedly different behavior, any mix of these agents types perform similarly. We do see some monotonic splitting the coalition behavior – with groups with more intellectuals likely to develop larger coalitions – this behavior does not appear to influence actual policy outcomes.

(a) Percent of good policies enacted.  
(b) Distance from optimal policy platform.

(c) Percent of agents in the largest coalition.

**Figure 6.7**: Comparing Mix of Uninformed and Intellectual Agents
Figure 6.8: Comparing Mix of Ideologue and Intellectual Agents
6.1.4 Static Parameters and Limitations

While many of the model’s parameters have been examined within Chapter 2 and this appendix, there are still many parameters left unexplored. These fixed parameters generally fall into two categories: those related to the complexity of the task and those related to the complexity of individual and group behavior.

This first set of parameters describe the \( NK \) model itself. All simulations run in the paper use a simple terrain of \( N = 4, K = 2, \) and \( A = 2, \) with influence values drawn from a uniform distribution with bounds \([-10, 10]\). Modifying any of these parameters could make the solution space more complicated and therefore degrade agent performance. These parameters are intentionally kept fixed within this paper in order to better focus on individual reasoning and group dynamics. However, by using this relatively simple terrain, it limits what can be inferred about normative reasoning on even more complex tasks.

The second set of parameters capture a wide number of choices about individual and group behavior. While agent-based models must intentionally start with the simplest of assumptions, further exploration of this space could be greatly enhanced by examining more complex individual and group processes. As it is, the findings presented here only capture caricatures of common concerns around deliberation. Further exploring these parameters would allow for a richer model which more accurately captures deliberative reality.

Specifically, the model makes a number of simplifying assumptions about certain parameters being shared at the group level. \textit{Openness} and \textit{noise}, specifically, are treated as group-level phenomena. While some simulations explore different levels of \textit{openness}, this value is always assumed to be the same across all agents. Future models could benefit from making this a heterogeneous parameter, with some agents more eager to accept than others. Similarly, \textit{noise} is taken to affect most agents
equally. While intellectuals are protected from this effect, both uninformed agents and ideologues have their initial beliefs set directly in relation to the noise level. This, too, should be appropriately complicated in future models, with more variation in agents’ assumed cognitive capacity.

Another simplifying assumption of the model is in narrowing the agent types which are explored. All considered agents hold beliefs that are either randomly distributed from the truth (uninformed agents), closely bound to the truth (intellectuals), or intentionally skewed from the truth (ideologues). These are extremely limiting conditions, and while these classes of agents shed light on the specific deliberative failings explored within this paper, future models should examine a more exhaustive range of possible agent bounds.

Finally, there is a small group of parameters that could greatly affect the dynamics of group behavior but which go entirely unexplored here. First, both the true values and the noise added to agents’ beliefs are drawn from uniform distributions. This simplifying assumption allows the model here to focus on other elements of group dynamics, yet this could be a particularly interesting line of further inquiry – examining the dynamics of group decision making when either the truth or individual’s beliefs are inherently skewed.

Additionally, all the models presented here assume that all agents share the same set of influence patterns. That is, while agents disagree on the value of policy A’s influence on policy B, all agents agree that those policies are related. Such a strict assumption is necessary in order to focus on the parameters explored here, yet, going down a path of imagining fundamentally different world views – where agents not only disagree on the value of influence but also disagree on what influences what – could be a fruitful line of research that captures some of the real challenges of deliberation today.
On the whole, this paper presents an agent-based model aimed explicitly at examining three canonical deliberative failures. While the model makes several necessarily simplifying assumptions in order to focus on the parameters of interest, it does raise some interesting questions about the true challenges of deliberation. In finding that polarized, factioned, groups do surprisingly well at finding good solutions and that even cognitively limited agents can deliberate successfully, this work suggests that individuals and groups do not have to be perfect – they just have to be willing to listen.
6.2 Chapter 3 Appendix

Chapter 3: *The Structure of Reasoning: Measuring Justification and Preferences in Text* presents and validates a model for inferring conceptual network structure from text. Full replication materials can be found online at https://github.com/sshugars/conceptual_networks. This appendix includes details on one of the datasets used within that paper; an original survey conducted on Amazon’s Mechanical Turk.

6.2.1 Prompts

We select prompts which would be accessible to participants with a wide range of political knowledge, might activate different core concepts, and are well-studied issue areas in political psychology (Ditto et al., 2018):

**Abortion**

“Do you think abortions should be legal under any circumstances, legal only under certain circumstances, or never legal under any circumstances?” *From Smith et al. (2012)*

**Healthcare**

“Do you think it is the responsibility of the federal government to make sure all Americans have health care coverage?” *From Smith et al. (2012)*

**Childrearing/Authoritarianism**

“If you had to choose, which traits are more important for a child to have: obedience and good manners, or independence and curiosity?” *Feldman and Stenner (1997)*
6.2.2 Survey Questions

Demographics

1. Do you currently reside in the United States? (Yes/No)

2. What is your age? ____

3. What is your gender?
   □ Male
   □ Female
   □ Self-identify as ____

4. Do you identify as Hispanic or Latino? (Yes/No)

5. What is your race? (Check all that apply)
   □ Asian/Pacific Islander
   □ Black/African American
   □ Native American
   □ White
   □ Self-identify as ____

6. Where do you currently live? Please select your state and enter your city or town. ____

7. Your country/countries of origin is/are: ____

8. In which of the past elections did you vote? (Check all that apply)
   □ 2016 presidential election
   □ 2016 presidential primary
   □ 2014 election for governor, senator, or other offices

9. What is the highest level of education you have completed?
No high school degree
High school graduate
Some college, but no degree (yet)
2-year college degree
4-year college degree
Postgraduate degree

10. What was your total household income before taxes during the past 12 months?

Less than $25,000.
$25,000 to $34,999.
$35,000 to $49,999.
$50,000 to $74,999.
$75,000 to $99,999.
$100,000 to $149,999.
$150,000 to $199,999.
$200,000 or more.

11. Thinking about politics these days, how would you describe your own political viewpoint?

Very liberal
Liberal
Moderate
Conservative
Very conservative
Not sure
12. Aside from weddings and funerals, how often do you attend religious services?

☐ More than once a week
☐ Once a week
☐ Once or twice a month
☐ A few times a year
☐ Seldom
☐ Never
☐ Don’t know/prefer not to say

13. What is your present religion, if any?

☐ Protestant
☐ Roman Catholic
☐ Mormon
☐ Eastern or Greek Orthodox
☐ Jewish
☐ Muslim
☐ Buddhist
☐ Hindu
☐ Atheist
☐ Agnostic
☐ Nothing in particular
☐ Something else
Moral Foundations Questionnaire

*From Graham et al. (2011)* When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using a scale from 0 (not at all relevant; this consideration has nothing to do with my judgments of right and wrong) to 5 (extremely relevant; this is one of the most important factors when I judge right and wrong).

<table>
<thead>
<tr>
<th>Consideration</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whether or not someone suffered emotionally</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not some people were treated differently than others</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone’s action showed love for his or her country</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone showed a lack of respect for authority</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone violated standards of purity and decency</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone was good at math</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone cared for someone weak or vulnerable</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone acted unfairly</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone did something to betray his or her group</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone conformed to the traditions of society</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone did something disgusting</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone was cruel</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone was denied his or her rights</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone showed a lack of loyalty</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not an action caused chaos or disorder</td>
<td>0</td>
</tr>
<tr>
<td>Whether or not someone acted in a way that God would approve of</td>
<td>0</td>
</tr>
</tbody>
</table>

Please read the following sentences and indicate your agreement or disagreement.

*Compassion for those who are suffering is the most crucial virtue.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*I am proud of my country’s history.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*Respect for authority is something all children need to learn.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree
People should not do things that are disgusting, even if no one is harmed.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

It is better to do good than to do bad.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

One of the worst things a person could do is hurt a defenseless animal.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

Justice is the most important requirement for a society.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

People should be loyal to their family members, even when they have done something wrong.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

Men and women each have different roles to play in society.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

I would call some acts wrong on the grounds that they are unnatural.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

It can never be right to kill a human being.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

I think it’s morally wrong that rich children inherit a lot of money while poor children inherit nothing.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

It is more important to be a team player than to express oneself.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

If I were a soldier and disagreed with my commanding officer’s orders, I would obey anyway because that is my duty.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

Chastity is an important and valuable virtue.
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

6.2.3 The Big Five Inventory

From John and Srivastava (1999)
I see myself as someone who: *Is talkative*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Tends to find fault with others*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Does a thorough job*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is depressed, blue*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is original, comes up with new ideas*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is reserved*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is helpful and unselfish with others*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Can be somewhat careless*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is relaxed, handles stress well*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is curious about many different things*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is full of energy*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Starts quarrels with others*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is a reliable worker*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Can be tense*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

*Is ingenious, a deep thinker*

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trait</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>-------------------</td>
<td>----------</td>
<td>-----------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Generates a lot of enthusiasm</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Has a forgiving nature</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Tends to be disorganized</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Worries a lot</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Has an active imagination</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Tends to be quiet</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Is generally trusting</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Tends to be lazy</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Is emotionally stable, not easily upset</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Is inventive</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Has an assertive personality</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Can be cold and aloof</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Perseveres until the task is finished</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Can be moody</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Values artistic, aesthetic experiences</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>Statement</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>Neither Agree nor Disagree</td>
<td>Agree</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>----------</td>
<td>-----------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Is sometimes shy, inhibited</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is considerate and kind to almost everyone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does things efficiently</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remains calm in tense situations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prefers work that is routine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is outgoing, sociable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is sometimes rude to others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Makes plans and follows through with them</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gets nervous easily</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likes to reflect, play with ideas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has few artistic interests</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likes to cooperate with others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is easily distracted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is sophisticated in art, music, or literature</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Deliberative Values

*From Gastil et al. (2012)*

Please say how much you agree or disagree with these statements. There are no correct or incorrect responses to these opinion questions.

*Even people who strongly disagree can make sound decisions if they sit down and talk.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*Everyday people from different parties can have civil, respectful conversations about politics.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*The first step in solving our common problems is to discuss them together.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*Individual leaders make better decisions than groups like committees, more often than not.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*When we hear claims that are false or misleading, we have a responsibility to speak up and correct them.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*People who make political claims must always back up their arguments with solid evidence.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

*People should always present logical arguments in support of their views.*
Strongly Disagree  Disagree  Neither Agree nor Disagree  Agree  Strongly Agree

Political Knowledge

*From Carpini and Keeter (1993)*

Here are a few questions about the government in Washington. Many people don’t know the answers to these questions.

- Do you happen to know what job or political office is now held by Mike Pence?
  - □

- Whose responsibility is it to determine if a law is constitutional or not?
  - □ President
• How much of a majority is required for the U.S. Senate and House to override a presidential veto? ____

• Do you happen to know which party had the most members in the House of Representatives in Washington before the election this/last month? ____

• ’Would you say that one of the parties is more conservative than the other at the national level? Which party is more conservative? ____

Political Ideology

From Pew Research Center (2017)

Next are some pairs of statements that will help us understand how you feel about a number of things. Please choose the statement that comes closer to your own views – even if neither is exactly right.

a.

• Government is almost always wasteful and inefficient

• Government often does a better job than people give it credit for

• No Answer

b.

• Government regulation of business is necessary to protect the public interest

• Government regulation of business usually does more harm than good

• No Answer

c.

• Poor people today have it easy because they can get government benefits without doing anything in return
• Poor people have hard lives because government benefits don’t go far enough to help them live decently

• No Answer

d.

• The government should do more to help needy Americans, even if it means going deeper into debt

• The government today can’t afford to do much more to help the poor

• No Answer

e.

• Racial discrimination is the main reason why many black people can’t get ahead these days

• Blacks who can’t get ahead in this country are mostly responsible for their own condition

• No Answer

f.

• Immigrants today strengthen our country because of their hard work and talents

• Immigrants today are a burden on our country because they take our jobs, housing and health

• No Answer

g.

• The best way to ensure peace is through military strength

• Good diplomacy is the best way to ensure peace

• No Answer

h.

• Business corporations make too much profit
• Most corporations make a fair and reasonable amount of profit
• No Answer

i.
• Homosexuality should be accepted by society
• Homosexuality should be discouraged by society
• No Answer

j.
• Stricter environmental laws and regulations cost too many jobs and hurt the economy
• Stricter environmental laws and regulations are worth the cost
• No Answer