Modeling and Measuring Deliberation Online

Nick Beauchamp[1]

Northeastern University

Abstract

Online communication is often characterized as predominated by antagonism or groupthink, with little in the way of meaningful interaction or persuasion. This essay examines how we might detect and measure instances of more productive conversation online, considered through the lens of deliberative theory. It begins with an examination of traditional deliberative democracy, and then explores how these concepts have been applied to online deliberation, and by those studying interpersonal conversation in social media more generally. These efforts to characterize and measure deliberative quality have resulted in a myriad of criteria, with elaborate checklists that are often as superficial as they are complex. This essay instead proposes that we target what is arguably the core deliberative process – a mutual consideration of conceptually interrelated ideas in order to distinguish the better from the worse and to construct better conceptual structures. The essay finishes by discussing two computational models of argument quality and interdependence as templates for richer, scalable, nonpartisan measures of deliberative discussion online.

Keywords: social media, deliberation, interpersonal communication, argument, debate, persuasion, natural language processing, big data, political science, argument mining, machine learning, network analysis, framing, collective decisions.

[1] Email: n.beauchamp@northeastern.edu; web: nickbeauchamp.com. The author would like to thank Sarah Shugars, Lu Wang, and Kechen Qin for their help in writing this essay.

1 Introduction

As the online world continues its exponential growth, the role of interpersonal communication has become increasingly central in understanding how opinions form, change, and affect behavior. While face-to-face communication has always played a fundamental role in shaping our views, verbal interactions have remained difficult to study for obvious reasons, but the rise of social media reveals and documents the sheer scale of these interpersonal interactions. But social media has also bred pessimism about the effects of these interactions, often being depicted as a den of bullying, trolling, flaming, groupthink, and worse. To understand and foster its more positive potential, however, requires a robust model of productive conversation. This essay examines that challenge through the lens of deliberative theory, exploring how one might model and measure deliberation in the online world, potentially as a first step towards improving it. Deliberative theory began as an outgrowth of democratic theory, recognizing that pure procedural voting was often insufficient to achieve collective decisions that best reflect the fundamental interests of participants. Instead, it was proposed that through careful, informed discussion, a deliberating group can arrive at collective decisions that most participants would agree are more informed and better reflective of their true preferences and beliefs than would have been achieved with a quick, majority-rule vote. As the role of interpersonal communication on social media has become especially vexed recently, it becomes essential to discover when, where and how online conversations might lead to better decisions, more informed participants,

and increased mutual understanding. This essay ultimately argues that there are some grounds for optimism, at least about the potential for online deliberation. To reach that conclusion, though, requires a careful examination of the origins and many competing definitions of deliberation; the challenges of its measurement, especially online; and deeper models of the fundamental deliberative qualities that productive conversation might achieve.

Although each of these steps is challenging, for scholars of deliberation and argument, the world of social media would seem to be the perfect dataset and testbed. In many online domains conversations are threaded, so you can tell exactly who is addressing whom; identities are clear and persistent (albeit often pseudonymous); and since communities form and interact consistently for weeks, months, or years, long-term changes in speech, opinion, and online behavior can be measured and tracked over long periods of time. Furthermore, the same conditions that make these domains appealing to the social scientist may often make them exceedingly useful and productive for their participants, who can benefit from consistent, long-term and transparent interactions with other members to build deliberative conversations and communities that can expand and enrich their understanding of the world and provide essential long-term emotional support. Tempering this optimism is a prevailing sense that social media is often an echo chamber of fact-free self-reinforcing bubbles of like-minded users (Sunstein, 2009), or worse, a breeding ground for bullying, misogyny, racism, and other forms of harassment that reinforce existing prejudices and power-structures, and drive out all those who aren't aggressive straight white cisgender men (Kayany, 1998; Hobman et al., 2002; Jones et al., 2013). Evidence of such unproductive (or anti-productive) speech online abounds, of course, but like many aspects of social

media, there are also large degrees of selectivity bias at work here: even if such things are prevalent and even dominant on some platform, it may also be the case that many pockets exist where speech is more deliberative, conversational, and useful for millions of users across billions of posts. One of the goals of this essay is to explore what the conditions and criteria might be for more productive, deliberative conversation online, even if such things are relatively rare (or appear to be) today. Once these criteria are better delineated and their real-world conditions better understood, we can work towards enhancing the conditions conducive to productive speech.

Although social media may be relatively new, these debates about deliberation and its pitfalls, as well as the general skepticism towards the whole conversational endeavor, are not at all new. This essay begins with a discussion of deliberative theory in the traditional "offline," "real world" setting, in order to better understand the normative goals and practical challenges of deliberative conversation. At their most extensive, theories of deliberation can entail dozens of potential criteria, at various levels ranging from the institutional, to the individual, to the content of each sentence spoken; one of the challenges is abstracting from this menagerie a core set of criteria, if that is possible. Another challenge is turning theory into empirics: how are each of these criteria measurable, particularly at the massive scale allowed by social media? A particular challenge is measuring substantive criteria rather than just a series of more easily observable epiphenominal markers which may be overly specific to various platforms or transcription methods. This leads directly to the various computational methods that may be necessary to measure these deliberative processes in the freeform text one finds in transcripts from deliberative exercises and in social media.

Once the concepts and challenges posed by deliberation more generally have been delineated, we can turn to online deliberation more specifically. This encompasses both online environments that have been carefully constructed in order to further deliberative discussion (and often emerge out of offline deliberative work), as well as more "natural" online environments such as social media, that may be more or less deliberative as a function of their membership and online structures. While constructed online deliberative platforms often are accompanied by surveys and other outcome measures, they are often as rare, expensive, and unselfsustaining as offline exercises. By contrast, social media abounds in data, but it is often difficult or impossible to survey participants to see what they have learned, how they have changed, and how they feel about their conversational activities. Because of this, with social media – and often even with purpose-build online platforms – we need to measure deliberative quality and outcomes directly from the textual content and patterns of interactions. As we will see, there are many approaches to this, particularly as we move beyond subjective human judgment into more automated computational methods introducing tools from natural language processing (NLP), argument mining, and network theory. But while many of these methods seem well-suited to the empirical analysis of deliberation in the wild, they are hampered both theoretically and empirically. Theoretically, there appears to be an increasing proliferation of criteria and subcriteria, numbering now in the dozens across the literature. And empirically, there is a tendency to focus of superficial, easily-measured markers of argument and deliberation, rather than deeper structures such as argument strength or conceptual interconnections – understandably enough, given that the latter are much harder to measure.

This essay finishes, then, with a look at a couple of recent efforts to measure deeper deliberative structures. In the ideal world, deliberation is valuable because it allows people to better learn facts, ideas, and the underlying conceptual structures relating between them. Participating in or observing a discussion or argument should be beneficial because, ideally, not all arguments are equal, and not all communication is dominated by superficial rhetorical features: in a productive debate, for instance, the better side should win not because the winner is a better rhetorician, but because they are in possession of the better arguments. But to measure these things – if indeed they exist – requires us to have a model of better and worse arguments, in a way that is objective and unbiased by the prejudices of the measurer. And to measure genuine deliberative thought requires not just superficial notions of facts or lists of words associated with reasonable talk, but models of how ideas are interrelated and how deliberative conversation, like deliberative thought, sifts through and reorganizes these ideas to arrive a better global decisions. The purpose here is not to solve entirely these deeper issues, but more to raise and highlight their importance in moving beyond the myriad ad hoc approaches currently employed, toward a objective, scalable, and theoretically coherent model for measuring deliberation online.

2 Deliberation

Whether explicitly or implicitly, much recent research into online communication is deeply normative: Are people learning from each other, or are they in self-reinforcing bubbles, or are they engaged in combat that is at best polarizing and at worst deeply destructive? Even many descriptive or positive theories of communicative behavior online tend to have normative

questions underlying at least an aspect of their hypotheses, particularly as the researchers encounter one or another type of dysfunctional collective online behavior (such as "flaming" or "trolling"). But if we wish to understand more systematically how deliberation online can go wrong, it is useful to develop better theories about how we might want it to go right. Deliberative theory has been developed over the last thirty years in part as a response to a similar set of problems in the offline world, and suggests that the problem of developing, measuring, and understanding deliberative communication – and its absence – is considerably more complex than merely the absence of bullying or acrimony (for instance). This section examines existing, traditional "offline" deliberative theory, and in particular the fundamental challenges of measurement that carry over to a more extensive model of online deliberation.

Although the very concept and definition of deliberation is deeply vexed, as we will see below, a rough and minimal definition could be: an extended conversation among two or more people in order to come to a better understanding of some issue. There are many other aspects that various scholars consider essential – institutional, formal, and outcome-oriented, eg – but most share this core idea of people engaged in conversation with the purpose of increasing their understanding. But because there remain many disagreements about these criteria and their normative importances, it is worth spend a little time understanding the development of deliberative theory and the current status of the complex and far-ranging discipline it has become.

Theories of deliberative democracy were developed by Habermas and others (Habermas et al., 1985; Cohen, 1989; Dryzek, 1994) in part as a response to limitations in traditional democratic theory that are analogous to some of the limitations we've already discussed in the

online world. In particular, even when we are satisfied with a particular electoral system as a fair aggregate of existing public opinion, it is clear that these democratic procedures do not always succeed in producing the best outcomes according to various theories of justice (Bohman, 2000; Ackerman and Fishkin, 2002; Bächtiger et al., 2005; Thompson, 2008; Gutmann and Thompson, 2009; Mansbridge et al., 2010). Voters may be deeply uninformed, for instance, producing not just sub-optimal outcomes according to some objective measure, but even according to what those voters themselves would wish were they to learn all the facts (Fishkin, 2013). Similarly, voters may not have worked through their own ideas sufficiently (by either objective measures or their own lights); or may be ignorant not just of facts but of important ideas, or connections between those ideas; or may misunderstand in important ways the ideas and beliefs of other voters in ways that mislead (Habermas, 1990; Manin, 2005). Mere voting may reflect the current attitudes of the voters, but perhaps not their "better" selves – what they would opt for were they better informed about the world, themselves, and others. Thus the need for deliberation before voting, to allow participants to better understand themselves, each other, and the issues, fact and ideas at work.

Early deliberative theory was deeply rooted in the rational enlightenment tradition, where the main goal of deliberation was to learn information and share reasons, and one of the primary goals was ideally to reach some form of consensus (Habermas et al., 1985; Cohen, 1989; Habermas, 1990; Manin, 2005; Thompson, 2008). Out of this framework, the original criteria for practical deliberation were developed, with the goal of producing rational, factual discussion that ideally leads to consensual decision-making. To better understand these deliberative criteria, it is worth distinguishing here three stages of the deliberative process – stages which will persist

even after the definition of deliberation has been subsequently expanded beyond its narrow rationalist origins. Roughly speaking, these stages are (1) the input, such as the environment or institutional framework under which deliberation takes place; (2) the deliberative process itself, including the action of the participants and the content of their communication; and (3) the output, such as opinions changes, decisions, or votes. The environment or institutional input may extend from large-scale social structures down to the rules governing the structure of a single deliberative group; the process may include individual behavior as well as well as the content of individual speech acts; and the output may range from individual knowledge gain to more consistent or factually correct group decisions (Landwehr and Holzinger, 2010). Deliberative criteria in the rationalist tradition are often framed most explicitly in terms of both inputs and processes: the environment must be fair, equal and unbiased; individuals must emphasize reason-giving, respect and honest expression rather than emotion and strategic manipulation; and the content of what they say should actually contain arguments, ideas, facts, etc (Habermas et al., 1985; Cohen, 1989; Steenbergen et al., 2003; Mansbridge et al., 2010). But none of these criteria can really be considered successful in themselves unless the individuals have gained in their rational understanding of the issues and moved towards a consensual truth (output). For instance, a (purportedly) reasonable, respectful setting that produced acrimony and group-think would not ultimately be judged a deliberative environment. But conversely, just as we would judge a democracy somehow deficient if a benign dictator merely appointed a representative body, or merely dictated policies reflective of the democratic will without actually holding polls or elections, so too in deliberation, if we imagine an outcome where everyone simple ends up better informed

(eg, via reading some material on their own) without the deliberative process, it may be beneficial, but isn't truly deliberative. Thus neither input nor output are themselves sufficient, and this sense in which the process itself is the most fundamental part of deliberation becomes even more pronounced in the second wave of deliberative theory.

That a rational consensus exists and that the participants should strive to achieve it was central to the first wave of deliberative theory (Cohen, 1989; Habermas, 1990; Bächtiger et al., 2005; Niemann, 2006; Gastil and Black, 2007; Thompson, 2008), but this has since been loosened and expanded to encompass a wider variety of normative goals. In part, this is an empirical concession: as we can see from the online world, not only is consensus difficult to practically achieve even in the best of worlds, it may not be even theoretically possible when interests are fundamentally opposed, or when the participants occupy a variety of identities or other positions that, while not inherently in opposition, are not something that we would actually want to merge (Holzinger, 2004; Dryzek, 2009; Gutmann and Thompson, 2009; Mansbridge et al., 2010). Moreover, even where practically achievable, we have seen that there are many flawed forms of consensus, where either participants naturally converge via groupthink on a suboptimal position, or merely bully a subset of participants into agreement (Sunstein, 2009; Garrett, 2009).

In reaction to these various flaws in this rationalist, consensual version of deliberation, a second wave of deliberative theory arose that allows for a broader conception of the goals and procedures of deliberation. Here, deliberation can potentially progress even with self-interested actors (Mansbridge et al., 2010), implacable disagreements about the truth (Gutmann and Thompson, 2009), a diversity of backgrounds, and some degree of continuing ignorance or

incomprehension about the experiences or backgrounds of the other participants (Mutz, 2006). However, without the core idea of rational progression towards a consensual truth, the second wave of deliberative criteria tend to be more fundamentally procedural rather than outcome oriented. If the process is fair and unbiased, encompasses the full diversity of viewpoints and is respectful to all differences, and in some loose sense encourages justification, exploration, and explanation over rhetoric, attack, and strategy, then it fulfills the environmental criteria, with much less emphasis on the outcome (Mutz, 2006; Dryzek, 2009; Gutmann and Thompson, 2009; Mansbridge et al., 2010; Fishkin, 2013; Mansbridge, 2015). Clearly, an outcome full of acrimony and misunderstanding would be a failure, but beyond that relatively loose criterion, procedural rather than outcome-based criteria tend to dominate.

Measuring process, though – especially as the definitions broaden beyond the purely rational – leads to an array of tricky questions: If a process claims to be fair and open, does it actually achieve equal speech from all participants (Steenbergen et al., 2003; Steiner, 2004; Gutmann and Thompson, 2009)? If not, is the content of what is said at least diverse enough that it encompasses all the major positions at the table, and in rough proportion to the individual representation of those positions (Manin, 2005; Mutz, 2006; Thompson, 2008)? Do participants seem to present justifications, explanations, and exploratory questioning, rather than aggressive or strategic speech designed to "win" (Manin, 2005; Mansbridge et al., 2010)? Even more fundamentally, do they seem to be exchanging arguments and ideas – "ideas" conceived broadly as things that can be personal, anecdotal, or emotional and not just abstract reasons – rather than rhetorical jousting? And, even more fundamental to the deliberative processes itself, are their exchanges of

arguments and ideas responsive to each other and reflective of underlying concepts and structures, rather than the sorts of purely performative or rhetorical speech one finds so often online (Holzinger, 2004; Gutmann and Thompson, 2009; Mansbridge et al., 2010)?

These questions are arguably at the core deliberation, as will be discussed below. But each are quite difficult to answer empirically, and each lead to multiple nests of thorny theoretical problems. Again, one might be tempted to try to shortcut the discourse analysis by turning back to outcomes: eg, to simply survey participants at the end about their knowledge, understanding of the issues, satisfaction, etc. But again, those outcomes can be achieved without a deliberative process at all. If we are interested (as with democratic theory more generally) in achieving the correct outcome via the correct process, then the most fundamental and direct measure is not at the institutional (input) or outcome (output) levels, but at the discourse level – with all of its attendant empirical and theoretical challenges.

Perhaps as a result of the challenges of measuring process, much of the early and concrete work on deliberative measurement and institutional design was done at the input and output levels, but that has gradually expanded to capture the more elusive but arguably more fundamental procedural criteria (Steenbergen et al., 2003; Holzinger, 2004; Steiner, 2004; Gutmann and Thompson, 2009). In addition to more straightforward measures of institutional/environmental suitability (open, fair proceedings that at least attempt to encourage constructive, reasonable conversation) and outcomes (surveys of knowledge gain, satisfaction, and absence of groupthink or polarization), the tricky middle – the discourse itself – has seen important strides in measurement. Perhaps the most substantial is the Discourse Quality Index (DQI), which

attempts to measure a number of types of quality, including breadth of participation; depth and content of justifications; respect towards other groups and their arguments; and constructive behavior (Steenbergen et al., 2003). All of these are fairly abstract and subjective criteria, however, and while the DQI and and other measures of course subdivide these qualities into more detailed and potentially objective sub-criteria and measurements, the evaluation of them has generally required intensive work from fairly expert human coders, each with their various potential biases regarding what counts as a justification, constructive behavior, rational, manipulative, etc.

Partly as a response to these types of bias, and partly as a response to the sheer cost and effort of such hand-coded measurement, various more automated methods have recently been developed. Such methods are absolutely necessary to expand these measurements to raw online discourse at scale. As an illustration of one of the more comprehensive efforts a multifaceted computational measurement of deliberation, Gold et al. (2015) seeks to measure four different core aspects of deliberative discourse: equal participation; mutual respect; justification; and persuasive effects. Like most fundamental deliberative qualities, each of these is a theoretically complex and multifaceted concept, without any obvious computational measure than can be easily automated (hence the DQI's reliance on human coders). Equal participation is perhaps the most directly operationalized, since it operates on the individual level and one can use speaking time as a proxy for this. But for the content-based criteria, rather than construct deep substantive measures that somehow reflect the complex concepts at work in the minds of human coders – a challenging task – the authors find superficial markers that are hopefully associated with the

deeper measures. This approach is not specific to them, but rather illustrative of how many automated attempts have grappled with these difficult measurement issues. Respect, or its absence, is measured via interruptions as notated in a transcript; justification is measured via a couple grammatical constructs (in German) that are sometimes used when justifying remarks; and persuasiveness is tracked by verb associated with changes in opinion ("accept," "believe," etc). So while it may be that these markers are systematically associated with a number of fundamental measures, such an approach requires extensive validation against human coding, does not export readily to other languages or contexts, and does little to capture the core concepts of deliberation. However, with some work, it can presumably be converted over into a variety of languages and function as a quick and large-scale, though approximate, measure of deliberation along a number of different procedural dimensions. As we see in the next section, as the deliberative field turns to online environments to test and develop their theories, they increasingly employ similar measurement strategies – strategies which are scalable and automateable, but which are also necessarily somewhat superficial.

3 Online deliberation

Within the body of research into specifically online deliberation, there have been two somewhat distinct communities that have only recently begun to merge. On the one hand are those who emerge out of the deliberation tradition, interested in cataloging the various criteria enumerated above, distinguishing the pros and cons of online vs offline discussion, and constructing deliberative environments online. On the other hand is work that emerges more from computer

science and communications, which pays greater attention to behavior in existing rather than purpose-build online communities, and therefore grapples with more varied network topologies connecting interlocutors rather than simple chatrooms modeled on literal deliberative rooms, and which tends to employ more automated content analysis rather than using survey measures of outcomes (Himelboim, 2008; Himelboim et al., 2009; Gonzalez-Bailon et al., 2010; Himelboim, 2011; Choi, 2014). But although originally more descriptive than normative, much of the latter work has been drawn to normative deliberative questions very similar to those developed in the deliberation community.

Within the outgrowth of traditional deliberative theory into online communication, one encounters many of the same sets of checklists that we saw above. For instance, in Schneider we have four criteria: equality, diversity, reciprocity, quality (Schneider, 1997). In Dahlberg, seven: reasoning, reflexivity, ideal role taking, sincerity, equality, and autonomy (Dahlberg, 2001). In Janssen we have six: form, dialogue, openness, tone, argumentation, reciprocity (Janssen and Kies, 2005). And in the International Association of Public Participation (IAP2, 2007), five: inform, consult, involve, collaborate, and empower.(Nabatchi, 2012) Once again, we are presented with a congeries of theories and criteria, although of course plenty of overlap can be found among these. Each list, though, does tend to focus more on a single level discussed above: Schneider more on the environmental or input criteria; Dahlberg and Janssen more on individuals, process and content; and the IAP more on outcomes.

Perhaps the most wide-ranging recent effort to systematize many of these qualities is in Friess and Elders (Friess and Eilders, 2015), where they distinguish three phases of deliberation similar

to those outlined above: "Institutional/Input/Design" "Communicative/Throughput/Process" and "Productive/Outcome/Results." Within each of these three basic stages, they in turn enumerate the usual menagerie of criteria, although helpfully targeted for specifically online deliberation. For the institution: asynchrony, self-identification, moderation, empowerment, division of labor, and information; for the process: rationality, interactivity, equality and inclusion, civility, common good, and constructiveness; and for the output, knowledge gain, reason learning, opinion change, social trust, and political engagement (on the individual level) and consensus, error avoidance, epistemic quality, and legitimacy (on the collective level). They present a tidy table summarizing these aspects, but such tidiness somewhat masks the inherent complexity and even messiness of even this "systematic" set of criteria, with its 21 variables. The historical trajectory seems to be an ever-increasing list of criteria, with measurement and assessment falling farther and farther behind.

Furthermore, as challenging as it may be to enumerate and categorize all these criteria, much more challenging is implementing their measurement, particularly at the most central stage, the deliberative process. This challenge is illustrated by Nelimarkka et al. (2014), who closely examine three online systems from the perspective of Dahlberg and the DQI: The Living Voters Guide, including its earlier iterations Consider.it and Reflect (Kriplean et al., 2012); the Open Town Hall (Vogel et al., 2014); and the authors own California Report Card (CRC). The California Report Card is notable for the sensitivity of its attention to equality and autonomy, randomizing individual encounters and even their arguments in ways to prevent dominance by certain participants or their ideas. Much more challenging than these input criteria, though, are

the procedural measures at the heart of deliberation. The authors devote close attention to the difficulties of operationalizing reason, for instance, which like many others they take to be one of the core criteria for the procedural stage. They divide reason into reciprocity and justification, and while justification can often be measured via fact-giving and other superficial syntactic measures (as Gold et al. (2015) do), a deep measure of reciprocity is quite tricky, since simply counting responses in discussions fails to capture the degree that individuals really are taking in and thoughtfully responding to the ideas of their interlocutors (which Trénel (2004) distinguishes as formal interactivity vs substantial interactivity). This idea of reciprocity, which in many ways is close to the core ideal of deliberation, will be returned to shortly, but for now it can be taken simply as another instance of how difficult it is to measure process, particularly at the core level of interactive content. Yet as difficult as this may be in carefully crafted online deliberative environments, it is much harder to capture in the wild. Many of the approaches discussed above also turn their sights to natural online settings, but even recent work that careful applies of one of these sets of four, seven, or twenty-two criteria, can feel immediately obsolete or irrelevant when applied to ever-changing social media.

4 Deliberation in social media

On the one hand, this definitional proliferation and its associated measurement and design problems demonstrates the breadth and ambition of what modern deliberative theory has become. On the other hand, even relatively minimal real-world deliberative polls are sufficiently expensive that it is difficult to do the sort of extensive iterative experimentation necessary to design

effective institutions applicable to a wide range of domains. And when moving from experiment to implementation, even a well-designed system can be prohibitively costly to deploy at scale. Even if we take the idea of deliberative "polling" to heart and hope that, like a well-chosen focus group, a well-selected deliberative poll might somehow be representative of a larger polity, the situation is even worse than a focus group, inasmuch as the interactive aspect between people's opinions leads to a combinatorial explosion of possible outcomes that might be very sensitive to the exact backgrounds and behaviors of the participants. The appeal of online communication is therefore both at the implementation and experimentation stages: it is easier and cheaper to build and deploy deliberative systems online at the scale necessary for deliberative democracy, and more fundamentally, insofar as we still do not have ideal models or measurements of deliberation, it is far easier not just to design online deliberative experiments, but to use the copious quantities of online observational data to hone and test our theories.

One of the most helpful developments in this regard has emerged out of computer science and other fields outside of the explicitly deliberative. This research has traditionally been more interested in a positivistic understanding of online communication, but has been drawn into the same normative issues underlying deliberative work. Most usefully, given the complex and varying nature of environments and behaviors found online, a wide array of tools emerging out of network theory, natural language processing, and other domains has been developed that are useful for operationalizing some of these theoretically elusive deliberative criteria. Somewhat less attention, though, has been given to the "outcome" side of things, a point Friess and Eilders (2015) also

make about online deliberative research more generally – and a point that will be returned to later.

4.1 Early work: flames and bubbles

Some of the earliest work in this area begins with a striking characteristic of early online communication, an observation that is both descriptive and normative: the emergence and rapid prevalence of "flaming" (Reinig et al., 1997; Kayany, 1998). In addition to trying to explain this phenomena, some of the earlier work explicitly examines the tradeoffs between this tendency online and counterbalancing advantages that may emerge from the same set of underlying online features. On the one hand, Usenet discussion boards and blogs had quickly become famous for "flame wars," where conversations dissolve into vitriolic attacks that basically epitomize the opposite of deliberative discussion. On the other hand, while the absence of social cues was often blamed for the flaming tendencies (Kiesler et al., 1984; Kayany, 1998), the ano- or pseudonymity of the online fora can produce more equality between participants (Dubrovsky et al., 1991; Bordia, 1997; Albrecht, 2006), potentially leading to more engagement (Price and Cappella, 2002), diversity (Hargittai et al., 2008; Garrett, 2009; Wojcieszak and Mutz, 2009; Brundidge, 2010), and participation (Boulianne, 2009) – desiderata we've already seen in the roughly contemporaneous deliberation literature. But while the prevalence of "flaming" was taken as a given, less well examined was its cause: in particular, was it due more to an absence of social cues and conventions, or more to the very diversity of opinion that had been lauded, bringing people

together from further points along the ideological spectrum (eg) than would normally encounter each other in everyday life (Hobman et al., 2002; Papacharissi, 2004)?

A second line of investigation that subsequently emerged, while not directly responding to this research question, does seem to flow directly out of it. In the early 2000s the dominant form of social media was the weblog, and in the political domain it was soon noted that the most distinctive characteristic of the linkages between political blogs was dense intraparty connections with weak or absent interparty connections (Adamic and Glance, 2005). This also marked one of the first notable applications of network theory to examine not just homogenous environments, but the complex, self-selected interpersonal connections online that deeply affect the deliberative outcome (Himelboim, 2008; Gonzalez-Bailon et al., 2010; Himelboim, 2011; Eveland Jr and Kleinman, 2013). These "bubbles" of self-selected interlocutors were taken at the time, and have often been taken since, as self-evidently deleterious, blocking information flow and leading to less-informed participants. While perhaps the earlier forms of online communication led to somewhat greater diversity of participants, as the medium matured and people had greater ability to self-sort, bubbles arguably become more dominant, eroding many of the benefits of diversity.

Alongside this, more substantial theoretical work examining the bubble and groupthink phenomena was developing in the deliberative and social science literatures, with perhaps the most well-known early crossover being Sunstein (2009). This can also be seen as a counter-balance to the lauded "wisdom of the crowd" (Surowiecki, 2005; Page, 2008), where the most stylized result is that a diversity of opinion can produce group judgments more accurate than those of any of the individual participants, but that this increased accuracy is destroyed if the participants are

allowed to discuss their individual judgments first: discussion is precisely what collapses the wisdom of the crowd back to groupthink. There are of course many more important and theoretically interesting details about exactly how much diversity of opinion is optimal and how much discussion is sufficient to ruin the group wisdom, but this tension between diversity and self-destruction mirrors the earlier research into flaming and bubbles. How are we to know when we have a productively diverse discussion, vs an unproductive flame war? Except in the most artificial of circumstances, real-world deliberative outcomes are rarely as clear-cut as guessing the number of jelly beans in a jar. Instead of relying on outcome measures, mirroring the progression in deliberative work, online studies have turned more of their attention to the discourse itself, seeking more automated measures of discourse quality. Perhaps the most direct and prevalent content measure is emotion and sentiment, particularly as automated methods were developed to measure such things (Papacharissi, 2009; Berger and Milkman, 2012; González-Bailón et al., 2012). More sophisticated content measures have also emerged out of the natural language processing community, but less as a response to these problems than out of an entire different sub-discipline, which will be returned to shortly.

4.2 More complex measures of environment: network analysis

Perhaps a more substantial and novel contribution emerging from the study of social media behavior has been on the environmental or input side, as network measures were developed to model and explain deliberative behaviors. Because modern social media since Usenet is arranged

not in closed room-like silos but in open-ended networks of interconnected users, the entire first level of deliberative analysis – the environment – becomes problematic. Can Facebook or Twitter even be subdivided into smaller communities whose deliberative qualities can be independently assessed? And if not, we need to model both the environment and individual behaviors in a more open-ended network structure to discover which topologies are associated with deliberative quality, either as input or potentially as output. But like the more sophisticated and granular analysis of wisdom-of-the-crowd effects in Page (2008), this is ultimately a good thing: it forces us to understand how the environment and the individual interact, how the former shapes the behavior of the latter, and how the latter is constituted and shaped in turn by the self-selecting and linking behavior of the individuals. After all, barring costly deliberative democratic institutions, the most prevalent form of deliberation in actual life is interpersonal and thus networky, whether in person or online.

A review of network-based approaches to communication, even from a normative deliberation point of view, is beyond the scope of what can be covered here (Himelboim, 2008; Brundidge, 2010; Gonzalez-Bailon et al., 2010; Himelboim, 2011; Eveland Jr and Kleinman, 2013; Choi, 2014). Even something as far afield as retweet behavior is apropos, since those behaviors map onto our previous questions of bubbles, knowledge (and error) transmission, and even cascades of vitriol and "flaming" – all of which presumably have analogs in the more artificial and controlled environment of a deliberative meeting. How can the topology be tweaked to boost deliberative outcomes, and how do less deliberative behaviors shape their own self-reinforcing environments? These sorts of questions remain largely unanswered, but recent work has begun

to bring together the network and other computational tools with the more established deliberative concerns.

One recent illustrative multimodal example of this is Choi (2014), who examines four now-familiar criteria – discussion flow; diversity of opinion; rationality of discussion; and persuasion – but from a perspective that mixes network methods and automated content analysis. Since the domain is Twitter, the approach is more descriptive than normative or design-oriented, but the fundamental research questions are once again driven by the core deliberative concerns. Perhaps most interesting is the analysis of discussion flow or dynamics, which focuses on retweets but which is analogous to many forms of information transmission. Using an Exponential Random Graph Model, they examine how different local network topologies linking Twitter users affect tendencies to retweet. This is a common approach in network analysis, but the normative deliberative application is more relevant, where the question here is whether topologies tend to reinforce existing dominant speakers or diffuse communication out towards peripheral players, and on a second level whether cliques of speakers tend to form or not. In both cases, from a deliberative perspective we would prefer the latter options, with diffusion and equality rather than concentration and cliqueishness. The predominant result in network analysis is that concentration rather than diffusion dominates, although in this case Choi finds that this is less the case in this specific Twitter dataset than expected; whether that is due to increased deliberation or merely an underpowered sample is less clear.

Their examination of the other three qualities – diversity, rationality, and persuasion – likewise illustrate the sorts of automated content analysis that are now prevalent, although this

entire field remains limited by the same sorts of superficial content measures we saw in Gold et al. (2015). Diversity is operationalized via the domains in quoted URLs, which is used to measure the relative prevalence of inter- vs intra-ideological discussions based on the known ideology of different online news sources. While a fine measure, this illustrates how so many of these measures can be very specific to the form of the social media in which they transpire, since obviously URL quotation is a useless measure in live communication. It also illustrates some of the weaknesses of the purely descriptive approach, since while they of course find less inter- than intra-ideological discussion, it is unclear (a) what we would most desire normatively, and (b) what variations in conditions might affect these relative quantities. Content – rationality and its alternatives – are operationalized using now-standard NLP measures, in particularly the relatively basic tools included in toolkits such as the Linguistic Inquiry and Word Count toolset (Pennebaker et al., 2001), which purport to be able to measure content qualities as diverse as sentiment, anxiety, anger, or sadness, as well as high-level cognitive features such as causal reasoning, reflection, speculation, assertion, and so on. However, although reasonably well-validated against human judgments, these automated measures are mainly assessed via simple wordlists (which may be compiled either by human experts or more unsupervised computational methods) – so it is unlikely that they are able to capture deep conceptual structures, just as the syntactic indicators of "justification" in Gold et al. (2015) tended to be more superficial than deep. Consistent with other work (Berger and Milkman, 2012; González-Bailón et al., 2012), Choi finds that negative emotion tends to increase retransmission of content, and of the cognitive measures, strong assertion rather than causal or speculative thinking seems to dominate. But the generality of these

results, particularly given the superficial measures, remains a problem shared by almost all researchers of online deliberation.

4.3 More complex measures of content: argument mining

While the introduction of network analysis has been useful for pushing deliberative theory and practice away from simplistic environments to the sort of complex social structures common in actual human dynamics, the content side has often lagged, relying mainly on either human coding of broad categories, or automated methods that amount to little more than word counting. However, another branch of computer science and computational linguistics has recently been making great strides in deeper measures of deliberative content, under the broad umbrella term of "argument mining." Just as network analysis pushes deliberative theory to examine local interpersonal structures and not just homogenous institutions, argument mining in NLP pushes previously superficial content-based deliberative criteria into greater levels of argumentative detail. Like deliberative theory, argument analysis began a few decades ago as a more theoretical endeavor (Toulmin, 1974; Douglas, 1996), and really only blossomed as an empirical and computational program in the last 20 years. Some of the earliest work was with legal texts (Moens et al., 2007; Wyner et al., 2010; Mochales and Moens, 2011), attempting to identify arguments and then to identify and classify the substructures of different arguments, such as premises and claims. And as with deliberative theory, the types of arguments quickly proliferated, with for instance 96 different "schemes" (eg, from precedent, from effects, from authority, from fear, ad hominem, slippery-slope, etc) in Walton et al. (2008). These techniques were initially most often

applied to well-structured texts such as legal documents, where it is clear that (a) writers are indeed engaged in formal argument, and (b) the forms of those arguments are often sufficiently stereotyped to make for easier automatic retrieval. They have been also heavily applied in essay scoring (Shermis and Burstein, 2013; Stab and Gurevych, 2014a; Beigman and Deane, 2014; Stab and Gurevych, 2014b), where one desires to detect not just the presence and kinds of arguments, but to answer more normative claims about quality, but have expanded to many other domains. And of course much contemporary work has now shifted to the domain of online communication.

Early empirical work began with human classification, and then moved on to simple word-count approaches like those discussed above, eg looking for reasoning terms such as "because" or "therefore." But computational methods have subsequently advanced considerably, using machine learning methods to classify argument types (usually trained with human-coded examples) rather than (or in addition to) human-derived terms. Some of the most interesting recent work has been in argument sub- and super-structures: the relationships between premises, evidence, and claims, eg, on the substructure scale, and the relationships between larger arguments that support or contradict each other, eg, on the larger scale (Douglas, 1996; Walton et al., 2008; Palau and Moens, 2009; Feng and Hirst, 2011; Peldszus, 2014; Peldszus and Stede, 2015; Yanase et al., 2015). What is particularly interesting here from the deliberative perspective is how many of these structures go well beyond what deliberative theory has normatively classified. "Justification," for instance, is a very broad category (Biran and Rambow, 2011; Park and Cardie, 2014; Oraby et al., 2015; Park et al., 2015; Eckle-Kohler et al., 2015; Rinott et al.,

2015), and dozens of Walton's schemes could plausibly fit under than heading. This approach has also begun to incorporate less logical structures, such as emotional or personal appeals (Wang and Cardie, 2014; Wachsmuth et al., 2014; Oraby et al., 2015), which remain a vexed issue in deliberative theory as well. As with network structure vs institutions, argument classification in the mining domain has expanded far beyond even the 21 criteria we saw in Friess and Eilders (2015), but if anything the overall organizational structure unifying these components is even less clear.

4.4 Proliferating criteria and the core

We have thus seen an extreme proliferation in all of the major branches of deliberative analysis of online discussion. Deliberative theory itself has grown into dozens of potential criteria, and these have only increased as the specific peculiarities of online communication have entered the mix. Even if we aggregate some of these criteria into input, process, and output stages, developments in network theory have greatly complicated the environmental and individual levels, and developments in argument mining and NLP have greatly complicated the sorts of content we can evaluate from a deliberative perspective, even using raw text and large scales. Meanwhile, as Friess and Eilders (2015) also discuss, the outcome side of the process remains relatively underdeveloped, languishing in older measures such as knowledge and satisfaction surveys, or crudely unrealistic measures of collective accuracy that have little bearing on the complex subjective decisions characteristic of deliberative democracy (Page, 2008). All of these issues are, furthermore, fundamentally connected to the shift to online content. Deliberative

experiments and polls were problematized by the introduction of online deliberative polls, but these issues become much more pressing once we acknowledge that, despite its reputation, social media and other emergent modes of online communication are deeply deliberative, albeit erratically, with topologically complex structures and difficult to measure content (eg, Twitter jargon).

So what is the long-term solution here? The hierarchical classification approach (21 criteria, 96 schemes, etc) seems to only grow and become more baroque over time, whereas the search for one or a few core deliberative criteria seems to have been long abandoned, particularly with the turn away from the ideal of rational consensus. One advantage of traditional democratic and social choice theory is that it provides (theoretically) distinct notions of individual preference and collective outcome, so while opinion varies about what collective outcomes are ideal given individual inputs, or even whether those outcomes are theoretically achievable (Arrow, 1987), at least both ends of the process are relatively stable. With deliberation, particularly online, the extremely complex communicative process becomes the ends itself, and diversity of process, as well as opinion input, becomes almost an end in itself.

That said, perhaps we can make some headway in trying to narrow down the deliberative process to something closer to a core, hugging as close as possible to our earlier, rough-and-ready formulation: an extended conversation among two or more people in order to come to a better understanding of some issue. On the environmental input side, the core criterion most often appears to be something like equality: while in a deliberative poll one may need to seek out diversity of opinion, if we take the participants as given (eg, in a self-selected online community),

most of the sub-criteria commonly enumerated are means towards achieving the ends of equal participation. From a design point of view, there are questions of how to achieve this – equal speaking time, equalized and moderated content, limitations on self-sorting and/or enforced cross-ideological communication – but it is unlikely that any of these questions will have single definitive answers without the sorts of collective outcome criteria (such as factual accuracy) that vary from situation to situation. From an online point of view, though – momentarily ignoring the often egregious communicative content – the institutional setting would seem relatively well suited to deliberative communication, with relatively open and equal forums yielding an apparent equality of participation. The fact that the results are often so poor – from flaming to trolling – would suggest that this equality threshold may not be sufficient, and that indeed there may be no solution without looking more explicitly at communicative content and behavior.

On the other end of the process, outcomes seem to vary most widely depending on the interests of the theorist or institutional designer, and in theory seem to potentially encompass every normative good under the sun, from information to engagement to trust on the individual level to consensuses, increased agreement, or at least meta-agreement (about the justice of the decision procedure itself) on the collective level. The sine qua non, it would seem, would be some form of opinion change or behavioral change, but apart from this narrow core, there may be very little necessary overlap among these criteria. And the deliberative qualities suited to one out come may be entirely different that those suited to another. And from an online perspective, the outcome is perhaps understandably less well examined – especially in the wild – given how little explicit data we often have about opinion and behavior apart from the communicative activity

itself. Even measuring the bare minimum – opinion change of any form – can be quite challenging.

Despite having its own array of measurement challenges, though, the communicative process itself might be the most theoretically cohesive and substantive locus for capturing the core deliberative process. Assuming that participation is relatively equal and outcomes involve some sort of opinion shift other than polarization or groupthink, a deliberative process as distinguished from generic communication seems to involve something like collective deliberation in the same sense than an individual deliberates. For the collective, as for the individual deliberator, the idea is less about acquiring new information and ideas, and more that the facts, ideas, beliefs, memories, etc, that one already has must be considered, brought into bearing with each other, and formed into a more internally coherent structure that yields an overall opinion or behavior that is more internally consistent or accurate than what had come before. By these lights rationality, reason-given, or justification are simply means towards this broader sense of deliberation, as are the more upstream qualities often included in the deliberative process, such as civility, non-negative emotions, or constructive intentions. In the most ideal (if abstract) form of deliberation, one would expect the communicative process to somehow sort through the existing array of ideas and, based on the individual or collective evaluation of them, the best arguments would "win" and allow the individual or group to select a better course of action than before. That is, we would hope that the better arguments gain credence based on their merits, as opposed to the identity of the speaker or style of the presentation.

In the online world, there are presumably moments, individuals, and domains that are more or less deliberative in this way, and the goal would be to distinguish more or less deliberative processes to better establish the types of environments, individuals, and communicative content that are more or less conducive to deliberation. To do so, however, requires measurement of deliberative communication at this relatively abstract and general level: not words associated with justification or non-negative emotion, eg, but a way to distinguish in a content- and politically-agnostic way better arguments or ideas from worse, and to measure the conceptual connections between these ideas so that we can detect when (if ever) the collective or individuals have better sorted through their thoughts after deliberation.

5 Measuring argument quality

Rather than attempt yet another effort to sort through and tabulate the numerous, multi-level, multi-stage criteria for deliberation that have proliferated over the last few decades, the remainder of this essay focuses on better understanding and measuring what is arguably the core deliberative process: the consideration and exchange of ideas in order to discern the better from the worse, and to assemble the whole into a more cohesive and consistent interconnected system. A full model of this process remains beyond the scope of this essay, but the next two sections suggest a couple of the component parts: first, a model that measures the latent persuasive effects of content as separate from style (and thus may discern the side with the more persuasive content); and second, a model that infers the latent connections between ideas, in order to distinguish more interactive, responsive deliberation from less. The hope is that exploring these approaches here

may help us move towards more sophisticated models of online deliberation that actually capture something closer to the core deliberative process, rather than crude word-level correlates or rough individual-level behavioral characteristics.

One of the fundamental assumptions of deliberation is that, in the process of talking through ideas the group may eventually distinguish the better from the worse and thereby arrive at a better conclusion. We have seen that there are many ways to distinguish more or less persuasive styles, but how are we to measure the inherent value of arguments owing to the merit of their content rather than mere style? That is, how might we distinguish better from worse arguments? We could certainly imagine assigning an army of human coders to sift through vast quantities of text and score arguments, but of course that would both lack generality and objectivity. Indeed, it's hard to even conceptualize what an argument's inherent strength or value might be independent of the context and background of the evaluator. And indeed this may be part of the weakness of a theory of deliberation that imagines that arguments win on their own merits. But without something like that assumption, we are left with little more than style and expression, and the entire edifice of rationality, truth and reason seems lost. So for the purposes here, let's consider a slightly more modest goal: inferring not the objective quality of arguments, but the inherent, if latent, persuasive effect of topics or ideas that come by merit of their content rather than the style of their presentation. This at least moves us a little closer to full deliberative model of content quality.

To infer this empirically, consider a specific form of argument strength: the inherent persuasiveness of an idea, subtopic, or sub-issue, within the context of a larger debate on a broader issue (Wang et al., 2016). Our data are currently from live rather than online speech – the

"Intelligence Squared" Oxford-style public debates – so the discussion here will be brief, but the model may be applied to online discussion equally well if we have an overall measure of opinion and persuasion. Our motivating example is a segment of debate about the death penalty quoted in Table 1. Here both sides are discussing the death penalty from the perspective of the mistaken execution of the innocent; both use various rhetorical techniques to strengthen the persuasive effects of their arguments, but both are also constrained by the overall topic. The hypothesis is that some topics of argument (within the overall topic of the death penalty) are inherently more suited to one side or the other, and while the disadvantaged side may do their best to use rhetoric or other smaller pieces of information to bear, they are fundamentally disadvantaged as long as the discussion is on this issue, and would do better to switch the topic to something better suited to their side. So in this case, the discussion of innocent executions is better suited to the anti-death-penalty side, as perhaps would be a discussion of racial disparities in execution rates; whereas a discussion of the specific heinous acts in various murders would perhaps inherently support the pro-death-penalty side. While this framework, with its emphasis on strategy and rhetoric, may seem somewhat counter to the spirit of deliberation with its emphases on justification and sincerity, it acknowledges that no matter how sophisticated the environment or moderation, or how sincere the participants, the reality is that people do their best to persuade others using many different substantive and rhetorical tools. The fundamental assumption for deliberation to be able to take place, however, is that there be at least some substance beneath the rhetoric, and that we have some way (as listeners or researchers) of distinguishing inherently better and worse

arguments and may come closer to the truth (even a multivalent, contingent truth) upon deliberating over these arguments as they are put forth.

TABLE 1

Motion: *Abolish the Death Penalty*

Argument 1 (PRO): What is the error rate of convicting people that are innocent? ...when you look at capital convictions, you can demonstrate on innocence grounds a 4.1 percent error rate, 4.1 percent error rate. I mean, would you accept that in flying airplanes? I mean, really. ...

Argument 2 (CON): ... The risk of an innocent person dying in prison and never getting out is greater if he's sentenced to life in prison than it is if he's sentenced to death. So the death penalty is an important part of our system.

Argument 3 (PRO): ...I think if there were no death penalty, there would be many more resources and much more opportunity to look for and address the question of innocence of people who are serving other sentences.

Model inferences

Rhetorical features: *Questions* (1), *Numerical evidence* (1), *Logical reasoning* (2, 3)

Strength: (1) & (3) inferred *Strong*; (2) inferred *Weak*

An excerpt from a debate on the death penalty, discussing the execution of innocents. Our model scores each argument according to its rhetorical features as well as its latent persuasive strength.

Each of the public debates in this dataset is on a set topic (eg, the death penalty) with experts arguing either side. Crucially for these purposes, the audience for these public debates is surveyed about the issue both before and after the debate, with the "winner" being the side that has gained more supporters, and we take this outcome as our measure of persuasive effect. Thus we are able to train the computational model using the observed debate outcomes, to determine which topics and other features are predictive of winning a debate (ie, which are persuasive). A Hidden Topic Markov Model (Gruber et al., 2007) is used to automatically segment a debate into "arguments" – chunks of text a few sentences long all on the same topic. We also measure per "argument" as

many stylistic features as possible that might have persuasive effects – everything from pronouns and basic sentiment to logical and justification terms, hedges, emotions, concrete language, readability, personality, and raw word counts. The idea is that both substance (the inherent persuasive effects of various topics) and style affect outcome, and if we wish to avoid merely finding the superficial stylistic correlates of substance (e.g., counting "reasoning" words like "because" or "therefore"), we must incorporate those stylistic markers in our model, and hope that in controlling for these superficial effects, the residual will be the inherent topic-specific persuasive effects. We then build a latent variable structural SVM (Yu and Joachims, 2009), which uses both the observed stylistic features and the unobserved, latent argument strengths to predict debate outcomes, where the observed debate winners serve to train the model to infer which features are persuasive and which topics are strong for one side or the other, and then is tested out-of-sample by using those inferred values to predict the outcomes of some debates and compare those predictions with the actual outcomes. There are of course many more technical details to this which are detailed in Wang et al. (2016).

Ultimately, we are able to predict 73% of the debate outcomes correctly (out of sample), a significant improvement over using just the observed stylistic features (66%) or guessing at random (53%). How is it possible to predict the inherent persuasive effects of topics in a debate that the model has never encountered before, on a subject entirely new to the dataset? Because the model estimates the effects of interactions between our myriad stylistic measures and topics, it learns the many subtle stylistic features that tend to be associated with intrinsically stronger or weaker topics, and then uses those observed stylistic features to predict the inherent persuasiveness of new topics.

This is different, however, from merely using style to predict persuasion: only when combined with the model of latent strength does the predictive accuracy jump, suggesting that while we may (with lots of computation) be able to identify persuasive content by the style in which the arguers present it, it is nevertheless the merits of the content that cause those persuasive effects, rather than some elaborate interaction of styles alone.

If one finds this plausible, what it shows is that argument strength – the idea that some topics are inherently more persuasive for one side than the other – does seem to play a significant role in debate. The winning side is usually (but not always) the one with more strong arguments, and most arguments are strong for one side or the other but not both. It may of course seem over-idealistic to imagine that the inherent value of content plays any role in persuasion in public debates, or one may be skeptical of the model's method for inferring these latent values. But the larger point is that, if one believes that one of the core purposes of deliberation is separating the conceptual wheat from the chaff, one has to believe that individual deliberators have a way of doing so that is not merely reflective of superficial stylistic presentation. And to empirically measure this, one needs a human or computational approach that can discern such latent values, either using the methods described here, or some other latent model that goes far beyond the usual superficial word-counting common in other measures of deliberative quality. Of course, much of the value of ideas lies not in some atomic, intrinsic value, but via their interrelationships and mutual consistency, inter-dependencies, and contextual meanings. Inferring those things empirically leads us to the second deep-modeling approach.

6 Measuring conceptual connections

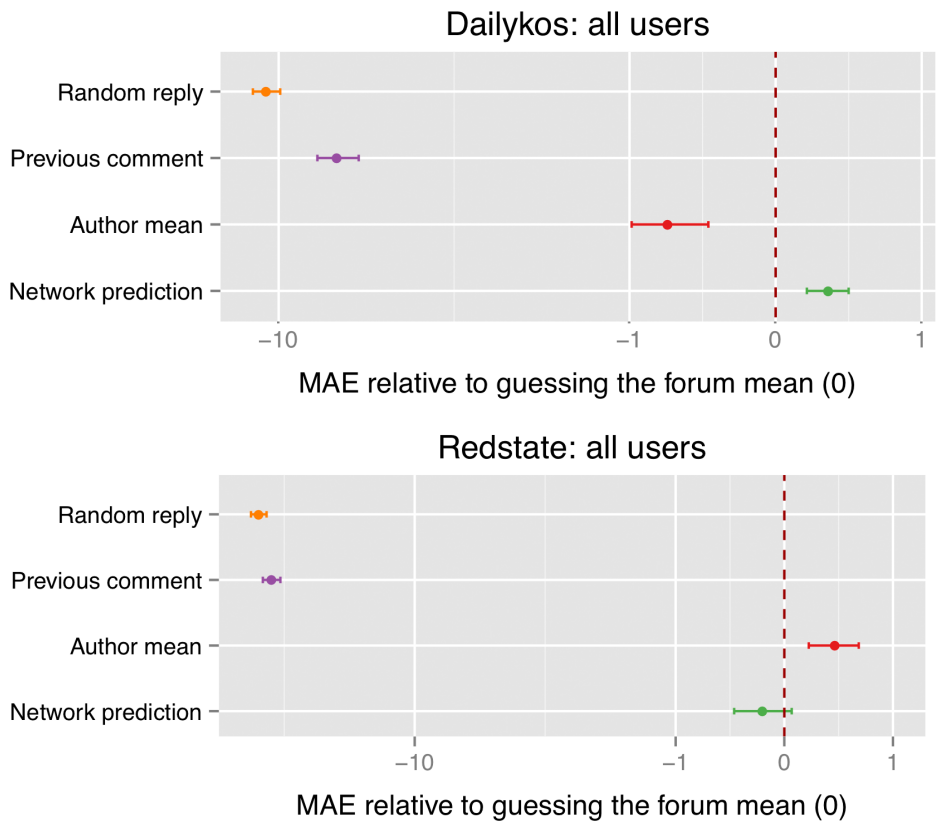
To measure the connections between ideas requires an additional layer: not just the topics of discussion, but a network of links between them. Others have approached this idea of a conceptual network in similar ways (Shaffer et al., 2009; Lodge and Taber, 2013), but the goal here is to infer these networks using purely computational methods. These are similar to the ones used above: first one automatically infers the topics under discussion, and then one infers the connections between those topics, operationalized as the correlations between their usage, where the assumption is that topics that co-occur tend to be more tightly linked in people's minds, and those that are anti-correlated are considered to be inconsistent with each other in some practical sense. The hypothesis is that people have these networks of linked ideas in their minds and, when they are truly deliberating, they don't merely hammer on their own preferences, nor do they mindlessly echo whatever framing content their interlocutor has put forth. Rather, in a truly responsive conversation, a person will consider the topics their interlocutor has raised, and then raise ones that are logically or conceptually related – new ideas that support or contradict the position they just heard. The empirical test of the model, therefore, is whether we can predict the topics of person B responding to person A better using the network-based approach, compared to merely guessing B will either mindlessly echo A, or merely beat on the same drum B always beats on. If this can be established, then we might validate both this specific model of network-based conversation, and also the idea that deliberative responsiveness as a latent quantity can be objectively measured. It may also allow us to discern when and where interlocutors are more or less deliberative, and how conceptual networks may differ and influence deliberation.

To build this model, we need much more interpersonal communication than the formal debates employed above provide. Instead, the two largest online political forums were scraped, each with millions of users over the last decade, one of the left (Dailykos) and one on the right (Redstate), because there are virtually no biparitsan political forums online (Beauchamp, 2016). While the dream of a robust, full-spectrum deliberative environment remains unfulfilled, these uni-partisan forums are still robustly deliberative, probably much more so than domains where the two sides go to war with each other (eg, the comment forums for many newspapers). Indeed, because it is less clear what "ideology" means within party in the US at the moment, this is precisely what makes these discussions more deliberative and potentially fruitful, since the participants themselves are not aware of what labels and identities should apply to themselves (except, arguably, during primary season). Both forums have had extensive discussions with many millions of posts over many years, and discussions tend to be threaded such that one can discern who is speaking to whom.

As before, we begin with a topic model, but then construct a network of their correlations linking topics with positive or negative edges depending on their empirical correlations within the speech acts of all the users. Deliberative responsiveness is modeled as the raising of topics or ideas related to, but different from, what one's interlocutor has just said, which is modeled as a markov process where person A makes a point with some distribution over the topics (eg, 70% about the execution of innocents, 10% about recidivism, etc), and person B responds with topics that are correlated with but in some respect different from those topics (eg, bringing up life imprisonment in addition to the existing topics under discussion), which is operationalized by multiplying the

topic vector through the correlation matrix: $b = \mathbf{C}a$. Figure 1 shows how well various models of responsiveness describe the interactions on these two forums, where each model is tested by seeing how well it predicts the topics of a comment by person B responding to a comment by person A. As we can see, there are interesting differences between the left and right: while on the right, the best predictor of what B will say in response to A is whatever B usually says (B's mean response), for the left the best predictor is the network model of responsiveness: ie, at least for these two forums, the left – whether due to the environment, individuals, or content under discussion – shows more core deliberative responsiveness.

Figure 1



How well various models of user interaction explain the content of post B written in response to post A, across two large political forums, Redstate (conservative) and Dailykos (liberal). Models with higher values predict responses better (negative Mean Absolute Error relative to forum mean, with 95% confidence intervals). For Redstate, the topics of response B are best predicted as the topics the speaker usually uses under all circumstances; for Dailykos, the network model of responsiveness best predicts the content of replies.

This study also establishes that these deliberative arguments do seem to have long-term effects on users' ideology, where ideology is measured by examining which posts users tend to like in a fully automated way (much as the vote record can be used to infer ideology without any human supervision), and also shows that the density and structure of these conceptual networks seems to vary across the left and right in interesting ways (see Beauchamp (2016) for more details). But most fundamentally for our purposes here, it suggests that not only do individual topics and ideas have intrinsic strengths that appear to be correlated with real-world persuasive outcomes (at least, in formal debates), but also that these topics and ideas are connected in conceptual networks that allow us to discern more from less deliberative conversation, where the notion of deliberation here is not some superficial, stylistic measure, but instead a deep model of responsiveness that attempt to model objectively the way in which truly responsive discussants engage in creative and complex ways with what each other are saying. Not only are some topics more inherently persuasive than others, but their effects are intertwined, and to understand deliberation – the careful consideration of ideas to work through their connections and implications – requires something like these high-level latent models in order to characterize and predict behavior.

Neither approach offers a complete model of core deliberation, nor do these ideas comprise the only possible core. Nor is the first model readily applicable to domains without well-measured

persuasive outcomes, or the second readily applicable to the sorts of brief, more diffuse conversations more common on Facebook, Twitter, etc., let alone face-to-face conversation. But both suggest an approach that might allow us to focus on some core notion of deliberation that is fairly general yet conceptually rich, and which is neither a collection of superficial stylistic correlates, nor a vast collection of heuristics for capturing every aspect of deliberation hypothesized by three decades of theorists. But whatever the specific approach, a model that focuses on responsive conversation in the service of conceptual improvement, or some other such core, allows us to both discern fundamentally deliberative environments or moments in the wild, and take some action to tweaking those environments in beneficial ways without requiring that we construct complete online deliberative utopias in order to achieve the civil and democratic purposes for which deliberative theory was originally developed.

7 Conclusion

Deliberation theory began to grapple with online conversation almost as soon as internet communication became prevalent. While most comparisons have been unfavorable whether or not the online environment was found or constructed (Dubrovsky et al., 1991; Bordia, 1997; Hobman et al., 2002; Min, 2007), we have also seen that online modes offer many potential advantages, including anonymity, diversity, engagement, cost, and more flexible open-ended network structures. While deliberative polls and even purpose-build online deliberative forums are prohibitively costly to deploy at the scales common for social media, social media itself has many deliberative characteristics and subcommunities. The challenge is discerning these more

deliberative spaces and moments without getting swamped either by the immense proliferation of deliberative criteria, or the immense scale of measuring terabytes of textual content and other online behaviors. While NLP and argument mining methods present new approaches to measure deliberative behavior in the wild at scale, these domains also tend to run wild with numerous criteria and measures for gauging argument and discussion quality, many of which measures are also disappointingly superficial when examined in detail.

While it is unlikely that this large, multi-disciplinary community will converge upon a single core notion of deliberation any time soon, it has been argued here that it is still worth seeking out a core concept of deliberation. This theory should ideally be distinct from existing political theories, including many aspects of environment or input (such as equality or representativeness), much of the output (such as information gain), and even many procedural criteria (such as civility or respect) which are all potentially well-established norms in existing theories of democracy or civil society. Arguably, this deliberative core is, as the term deliberation suggests, less about adding information or behaving well, and more about how existing arguments, facts and ideas are evaluated by the thinking individual or conversing group – how the better arguments and ideas are sifted from the worse, and how the connections of support and contradiction among them are worked through to construct more consistent conceptual networks and courses of action.

To measure these abstract qualities, however, requires not just superficial measures of textual features, individual opinions, or environmental structures, but rather more substantial procedural models of the quality and interrelations among arguments. Separate models of these two things were presented here, but ultimately we will need a deeper model combining the

strength of ideas, the network of support and contradiction linking them, and their behavioral consequences for the individual and a group trying to come to a collective decision. While no model will truly be able to measure the content and quality of ideas or the immensely complex logical, conceptual, and cultural connections between them, more substantive models of deliberation will allow us to better discern not just brief moments of deliberative quality online, but also to discover what sorts of network topologies, content moderation, individual characteristics, and content topics are best suited to productive conversation, and guide whatever interventions we can manage into the fast-evolving jungle of social media.

References

Ackerman, B. and J. S. Fishkin (2002). Deliberation day. *Journal of Political Philosophy 10*(2), 129–152.

Adamic, L. A. and N. Glance (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43. ACM.

Albrecht, S. (2006). Whose voice is heard in online deliberation?: A study of participation and representation in political debates on the internet. *Information, Community and Society 9*(1), 62–82.

Arrow, K. J. (1987). Arrow's theorem. *The new palgrave: a dictionary of economics 1*, 124–126.

Bächtiger, A., M. Spörndli, M. R. Steenbergen, and J. Steiner (2005). The deliberative dimensions of legislatures. *Acta Politica 40*(2), 225–238.

Beauchamp, N. (2016). Someone is wrong on the internet: Modeling argument and persuasion via an exchange of ideas. *Working paper*.

Beigman, Y. S. M. H. B. and K. P. Deane (2014). Applying argumentation schemes for essay scoring. *ACL 2014*, 69.

Berger, J. and K. L. Milkman (2012). What makes online content viral? *Journal of marketing research 49*(2), 192–205.

Biran, O. and O. Rambow (2011). Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing 5*(04), 363–381.

Bohman, J. (2000). *Public deliberation: Pluralism, complexity, and democracy*. MIT press.

Bordia, P. (1997). Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. *Journal of Business Communication 34*(1), 99–118.

Boulianne, S. (2009). Does internet use affect engagement? a meta-analysis of research. *Political communication 26*(2), 193–211.

Brundidge, J. (2010). Encountering "difference" in the contemporary public sphere: The contribution of the internet to the heterogeneity of political discussion networks. *Journal of Communication 60*(4), 680–700.

Choi, S. (2014). Flow, diversity, form, and influence of political talk in social-media-based public forums. *Human Communication Research 40*(2), 209–237.

Cohen, J. (1989). Deliberation and democratic legitimacy. *1997*, 67–92.

Dahlberg, L. (2001). The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society 4*(4), 615–633.

Douglas, W. (1996). Argumentation schemes for presumptive reasoning.

Dryzek, J. S. (1994). *Discursive democracy: Politics, policy, and political science*. Cambridge University Press.

Dryzek, J. S. (2009). Democratization as deliberative capacity building. *Comparative political studies 42*(11), 1379–1402.

Dubrovsky, V. J., S. Kiesler, and B. N. Sethna (1991). The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups. *Human-computer interaction 6*(2), 119–146.

Eckle-Kohler, J., R. Kluge, and I. Gurevych (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics*. Citeseer.

Eveland Jr, W. P. and S. B. Kleinman (2013). Comparing general and political discussion networks within voluntary organizations using social network analysis. *Political Behavior 35*(1), 65–87.

Feng, V. W. and G. Hirst (2011). Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 987–996. Association for Computational Linguistics.

Fishkin, J. (2013). Deliberation by the people themselves: Entry points for the public voice. *Election Law Journal 12*(4), 490–507.

Friess, D. and C. Eilders (2015). A systematic review of online deliberation research. *Policy & Internet 7*(3), 319–339.

Garrett, R. K. (2009). Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of Communication 59*(4), 676–699.

Gastil, J. and L. Black (2007). Public deliberation as the organizing principle of political communication research. *Journal of Public Deliberation 4*(1).

Gold, V., M. El-Assady, T. Bögel, C. Rohrdantz, M. Butt, K. Holzinger, and D. Keim (2015). Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, fqv033.

González-Bailón, S., R. E. Banchs, and A. Kaltenbrunner (2012). Emotions, public opinion, and us presidential approval rates: A 5-year analysis of online political discussions. *Human Communication Research 38*(2), 121–143.

Gonzalez-Bailon, S., A. Kaltenbrunner, and R. E. Banchs (2010). The structure of political discussion networks: a model for the analysis of online deliberation. *Journal of Information Technology 25*(2), 230–243.

Gruber, A., Y. Weiss, and M. Rosen-Zvi (2007). Hidden topic markov models. In *AISTATS*, Volume 7, pp. 163–170.

Gutmann, A. and D. Thompson (2009). *Democracy and disagreement*. Harvard University Press.

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT press.

Habermas, J., J. Habermas, and T. McCarthy (1985). *The theory of communicative action*, Volume 2. Beacon press.

Hargittai, E., J. Gallo, and M. Kane (2008). Cross-ideological discussions among conservative and liberal bloggers. *Public Choice 134*(1-2), 67–86.

Himelboim, I. (2008). Reply distribution in online discussions: A comparative network analysis of political and health newsgroups. *Journal of Computer-Mediated Communication 14*(1), 156–177.

Himelboim, I. (2011). Civil society and online political discourse the network structure of unrestricted discussions. *Communication Research 38*(5), 634–659.

Himelboim, I., E. Gleave, and M. Smith (2009). Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication 14*(4), 771–789.

Hobman, E. V., P. Bordia, B. Irmer, and A. Chang (2002). The expression of conflict in computer-mediated and face-to-face groups. *Small group research 33*(4), 439–465.

Holzinger, K. (2004). Bargaining through arguing: an empirical analysis based on speech act theory. *Political Communication, 21*(2), 195–222.

Janssen, D. and R. Kies (2005). Online forums and deliberative democracy. *Acta política 40*(3), 317–335.

Jones, L. M., K. J. Mitchell, and D. Finkelhor (2013). Online harassment in context: Trends from three youth internet safety surveys (2000, 2005, 2010). *Psychology of Violence 3*(1), 53.

Kayany, J. M. (1998). Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet. *Journal of the American Society for Information Science 49*(12), 1135–1141.

Kiesler, S., J. Siegel, and T. W. McGuire (1984). Social psychological aspects of computer-mediated communication. *American psychologist 39*(10), 1123.

Kriplean, T., J. Morgan, D. Freelon, A. Borning, and L. Bennett (2012). Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 265–274. ACM.

Landwehr, C. and K. Holzinger (2010). Institutional determinants of deliberative interaction. *European Political Science Review 2*(03), 373–400.

Lodge, M. and C. S. Taber (2013). *The rationalizing voter*. Cambridge University Press.

Manin, B. (2005). Democratic deliberation: Why we should promote debate rather than discussion. In *Paper delivered at the program in ethics and public affairs seminar, Princeton University*, Volume 13.

Mansbridge, J. (2015). A minimalist definition of deliberation. *Deliberation and Development: Rethinking the Role of Voice and Collective Action in Unequal Societies*, 27–50.

Mansbridge, J., J. Bohman, S. Chambers, D. Estlund, A. Føllesdal, A. Fung, C. Lafont, B. Manin, et al. (2010). The place of self-interest and the role of power in deliberative democracy. *Journal of political philosophy 18*(1), 64–100.

Min, S.-J. (2007). Online vs. face-to-face deliberation: Effects on civic engagement. *Journal of Computer-Mediated Communication 12*(4), 1369–1387.

Mochales, R. and M.-F. Moens (2011). Argumentation mining. *Artificial Intelligence and Law 19*(1), 1–22.

Moens, M.-F., E. Boiy, R. M. Palau, and C. Reed (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pp. 225–230. ACM.

Mutz, D. C. (2006). *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

Nabatchi, T. (2012). Putting the "public" back in public values research: Designing participation to identify and respond to values. *Public Administration Review 72*(5), 699–708.

Nelimarkka, M., B. Nonnecke, S. Krishnan, T. Aitamurto, D. Catterson, C. Crittenden, C. Garland, C. Gregory, C.-C. A. Huang, G. Newsom, et al. (2014). Comparing three online civic engagement platforms using the "spectrum of public participation" framework. In *Proceedings of the Oxford Internet, Policy, and Politics Conference (IPP)*, pp. 25–26.

Niemann, A. (2006). Beyond problem-solving and bargaining: genuine debate in eu external trade negotiations. *International Negotiation 11*(3), 467–497.

Oraby, S., L. Reed, R. Compton, E. Riloff, M. Walker, and S. Whittaker (2015). And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. *NAACL HLT 2015*, 116.

Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.

Palau, R. M. and M.-F. Moens (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pp. 98–107. ACM.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society 6*(2), 259–283.

Papacharissi, Z. (2009). The virtual sphere 2.0: The internet, the public sphere, and beyond. *Routledge handbook of internet politics*, 230–245.

Park, J. and C. Cardie (2014). Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pp. 29–38.

Park, J., A. Katiyar, and B. Yang (2015). Conditional random fields for identifying appropriate types of support for propositions in online user comments. *NAACL HLT 2015*, 39.

Peldszus, A. (2014). Towards segment-based recognition of argumentation structure in short texts. *ACL 2014*, 88.

Peldszus, A. and M. Stede (2015). Joint prediction in mst-style discourse parsing for argumentation mining. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 938–948.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates 71*, 2001.

Price, V. and J. N. Cappella (2002). Online deliberation and its influence: The electronic dialogue project in campaign 2000. *IT & Society 1*(1), 303–329.

Reinig, B. A., R. O. Briggs, and J. F. Nunamaker Jr (1997). Flaming in the electronic classroom. *Journal of Management Information Systems 14*(3), 45–59.

Rinott, R., L. Dankin, C. Alzate, M. M. Khapra, E. Aharoni, and N. Slonim (2015). Show me your evidence–an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in NLP (EMNLP), Lisbon, Portugal*, pp. 17–21.

Schneider, S. M. (1997). *Expanding the Public Sphere through Computer-Mediated Communication: Political Discussion about Abortion in*. Ph. D. thesis, Massachusetts Institute of Technology.

Shaffer, D. W., D. Hatfield, G. N. Svarovsky, P. Nash, A. Nulty, E. Bagley, K. Frank, A. A. Rupp, and R. Mislevy (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning.

Shermis, M. D. and J. Burstein (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Stab, C. and I. Gurevych (2014a). Annotating argument components and relations in persuasive essays. In *COLING*, pp. 1501–1510.

Stab, C. and I. Gurevych (2014b). Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pp. 46–56.

Steenbergen, M. R., A. Bächtiger, M. Spörndli, and J. Steiner (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics 1*(1), 21–48.

Steiner, J. (2004). *Deliberative politics in action: Analyzing parliamentary discourse*. Cambridge University Press.

Sunstein, C. R. (2009). *Republic. com 2.0*. Princeton University Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Thompson, D. F. (2008). Deliberative democratic theory and empirical political science. *Annu. Rev. Polit. Sci. 11*, 497–520.

Toulmin, S. (1974). The uses of argument. 1958. *Cambridge: Cambridge UP*.

Trénel, M. (2004). Measuring the quality of online deliberation. coding scheme 2.4. *Unpublished paper 18*, 2004.

Vogel, R., E. Moulder, and M. Huggins (2014). The extent of public participation. *International City/County Management Association (ICMA)*.

Wachsmuth, H., M. Trenkmann, B. Stein, and G. Engels (2014). Modeling review argumentation for robust sentiment analysis. In *COLING*, pp. 553–564.

Walton, D., C. Reed, and F. Macagno (2008). *Argumentation Schemes*. Cambridge University Press.

Wang, L., N. Beauchamp, S. Shugars, and K. Qin (2016). Here's where you're wrong: Joint effects of content and style on debate outcomes.

Wang, L. and C. Cardie (2014). A piece of my mind: A sentiment analysis approach for online dispute detection. In *ACL (2)*, pp. 693–699.

Wojcieszak, M. E. and D. C. Mutz (2009). Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication 59*(1), 40–56.

Wyner, A., R. Mochales-Palau, M.-F. Moens, and D. Milward (2010). *Approaches to text mining arguments from legal cases*. Springer.

Yanase, T., T. Miyoshi, K. Yanai, M. Sato, M. Iwayama, Y. Niwa, P. Reisert, and K. Inui (2015). Learning sentence ordering for opinion generation of debate. *NAACL HLT 2015*, 94.

Yu, C.-N. J. and T. Joachims (2009). Learning structural svms with latent variables. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1169–1176. ACM.