The Structure of Reasoning: Inferring Conceptual Networks from Short Text

Sarah Shugars[1]

[1]Network Science Institute, Northeastern University

Abstract

Public opinion is often considered as an aggregation of preferences, but the field has the potential to be much richer. For decades, scholars have argued for the value of going beyond measuring political preferences and examining how individuals reason about and justify those preferences. However, this task has only recently become tractable with the emergence of modern computational methods. In this paper, I present a text-based approach for inferring characteristics of individuals' political reasoning. This method identifies the key concepts a person raises and examines the implicit connections between those concepts – what ideas are connected to which other ideas? This structural approach is theoretically justified in both the cognitive and linguistic literatures, which repeatedly suggest that humans store, retrieve, and interpret information through network structures. I show that this approach provides insight into the quality of a person's reasoning and reveals meaningful individual variation which is correlated with known behavioral traits. The ability to measure and interrogate individuals' expressions of political reasoning holds the potential to shed new light on the dynamics of public opinion and political behavior. Questions of persuasion, ideological fracturing, and conversation quality all rely upon understanding individual styles of political expression. These dynamics are driven not just by what someone says but by how they say it.

1 Introduction

The individually distinctive ways in which people express their preferences holds the potential to reveal broader variations in political behavior. A population's agreement on a given policy position, for example, may elide deeper divisions in the motivation behind that position. Similarly, discussants with opposing preferences may, under some circumstances, be able to find ways to engage productively across their differences. While the bulk of public opinion literature has rightfully focused on the output of what people believe – i.e., individuals' discrete policy preferences – a robust understanding of the full deliberative system additionally requires analysis of how people express these preferences and interpret the preferences of others. That is, while political preference models are invaluable for capturing trends in public opinion and predicting policy outcomes, they are

not designed to analyze *interactions* between individuals' preferences. What arguments can lead to opinion change and under what circumstances? What factors drive a political conversation to be productive or divisive? How can a society function democratically in the face of increasing levels of affective polarization? If we hope to answer such critical questions of public opinion, we need individual-level models of political reasoning and expression.

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

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

In pursuit of these joint goals, this paper presents a computational, text-based approach for inferring conceptual networks from text. Through the application of this method to two distinct datasets, the paper then demonstrates the potential for these network structures to generate behavioral insights. The method proposed here, described in Section 4, leverages the grammatical structure of a text in order to infer the implicitly encoded connections between mentioned concepts. Furthermore, the

method uses embeddings (Mikolov et al., 2013) to identify words within a text which point to the same concept.

This method is applied to two datasets, which are described in detail in Section 3. The first is an original dataset of 100 subjects recruited through Amazon's Mechanical Turk. Subjects responded to two political prompts and completed an extensive survey battery, estimating Big 5 personality traits (John and Srivastava, 1999), Moral Foundations tendencies (Haidt and Joseph, 2008), and other politically-relevant measures (Center, 2017; Knobloch et al., 2013; Carpini and Keeter, 1993). These personality traits are known to be correlated with political ideology (McCrae and Costa Jr, 1999; Haidt, 2012) and therefore bolster external validation for the method while simultaneously providing behavioral insights of their own. These measures and their hypothesized correlations are discussed further in Section 3. The second dataset is a sample of nearly 1,000 respondents recruited through YouGov for a survey designed by Daniel Hopkins (University of Pennsylvania) and Hans Noel (Georgetown University)[1]. In an "ideological Turing test," subjects generated texts for both the "liberal" and "conservative" positions on a single issue. This dataset therefore allows for a disaggregated examination of ideological and individual correlates beyond what is possible with the Mechanical Turk data. Furthermore, since nearly half of respondents (44%) provided ironic answers which did not meaningfully reflect a given ideological position, this dataset additionally allows for a rough examination of argument quality, e.g., of the extent to which a text reflects a reasonable representation of a given view.

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

[1]I would especially like to thank Drs Hopkins and Noel for generously sharing their data for this analysis.

2 Related Work

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

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

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

new insight to behavioral understandings.

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

These psychological theories have been applied in a range of settings. Semantic network libraries such as BabelNet (Navigli and Ponzetto, 2012), ConceptNet (Speer and Havasi, 2012), and SNePS (Shapiro and Rapaport, 1987) rely on the core psychological intuition that a concept, encoded as a word, can be best described though its associated concepts, which themselves are encoded as words. Education scholars have similarly leveraged psychological theories to argue that knowledge itself has a network structure and that "learning" can therefore be considered as a process of developing the right knowledge structures. In other words, the skill of applying existing knowledge to new situations relies upon developing an understanding of how relevant information is interconnected (Dorsey et al., 1999; Hong et al., 2004; Shaffer et al., 2009; Shavelson, 1974). Social scientists have further argued that conceptual networks can be used to examine how individuals reason and make choices between alternatives (Axelrod, 1976; Carley, 1993). In weighing possible outcomes, a person evaluates connected concepts and consequences; exploring paths within their conceptual network in order to determine the optimal choice. Political deliberation provides a natural venue to extend such models, as participants may enter conversation with differing views and must therefore attempt to share structured knowledge before reaching a decision.

Notably, the exchange of knowledge is most frequently done through language; leading to a separate stream of work engaging the structure of language as a proxy for the structure of knowledge. Perhaps the most well developed such models trace their roots back to Aristotelian efforts to define the structure of argumentation (Toulmin, 1958). Such structures may be relatively simple: a major premise

connected to a minor premise leads inevitably to a logical conclusion; or it may be significantly more complex, such as in the two dozen schemes described by Walton (1996) or the Context Free Grammar introduced by Mochales and Moens (2011). But while theorists have differed in the specifics of the models they put forth, their approaches all begin with implicit acceptance of the network structure of arguments: the soundness of a conclusion rests not only upon the ideas supporting it, but on the ways in which those ideas are connected. In other words, arguments fundamentally have a coherent structure expressed through linguistic structure and defined by evidence relationships (Cohen, 1987). The search for these structures has given rise to a rich body of research known as argument mining, in which supervised and semi-supervised computational methods automate the search for the sorts of argument structures articulated by Aristotle or Toulmin (Mochales and Moens, 2011). The conceptual networks inferred via these methods tend to be more structured and hierarchical than those inferred from open-ended psychological approaches, but the basic structure of nodes and edges representing ideas and their interconnections remains.

While psychological and linguistic approaches aim to infer and examine conceptual network structure, an important line of work in philosophy has developed normative theories regarding the properties of these networks. These theories rely primarily on principles of coherence, considering a moral position valid insofar as it is coherent with other views (Christen and Ott, 2013; Dorsey, 2006; Rawls, 1993). What constitutes "coherence," however, differs between philosophers, leading to differing topological interpretations. In Henry Sidgwick's influential version of utilitarianism, for instance, "the current moral rules" such as "do not lie" are used to generate most of our actual judgments (Sidgwick, 1907), leading to topologies in which some ideas serve as central gatekeepers. In particularist moral theories, by contrast, each moral judgment is only linked to others by loose and local analogies (Dancy, 1993), implying that no ideas should enjoy disproportionate centrality in a person's network of moral ideas. McNaughton and Rawling (2000) argue that this is the flaw of particularism, because some concepts really are "central" to morality. This suggests a hybrid approach in which core ideas are central but do not dominate the reasoning structure. These varied definitions of "coherence" share an understanding that consistency between individual pairs of beliefs is too low of a standard for judging the validity of a moral position. On the one hand, individual beliefs may be consistent but unrelated, while on the other hand, expecting all pairs of beliefs to be directly connected is too stringent a standard since moral views range over a wide variety of

topics. Several scholars have therefore explicitly argued for whole network approaches to coherence. Thagard (1998) proposes a theory involving literal network relations, though he overlooks many of the relevant formal features of networks. Berker (2015) posits that an individual's beliefs should be modeled as a network to reveal its degree of coherence and begins to explore the variety of forms that a network of moral values can take.

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

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

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

3 Data

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

3.1 Amazon's Mechanical Turk

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

By measuring both conceptual network structure and individuals' personality traits, this dataset provides external validation for the method and illustrates the potential for substantive behavioral insights. That is, if we believe that the structure of a subject's response is connected in some way to their personal style, we would expect to see a number of correlations between individuals' inferred

network structure and measured personality traits. Further, if we believe that such correlations are more than spurious, we would expect those correlations to cluster in meaningful ways – e.g., traits which have been found to be similar to each other in other settings (McCrae and Costa Jr, 1999; Haidt, 2012) should be suggestive of similar network structures here. While the specific measures of network structure are described in detail in Section 4.3, their expected correlations with the specific personality traits measured in this experiment are described below.

At the highest level, I would expect to see clustering between personality traits associated with liberal ideology and traits associated with conservative ideology. Moral foundations theory (Haidt and Joseph, 2008; Haidt, 2012) argues that political opinions are driven by an individual's orientation along at least five moral dimensions, and suggests that political divides can be traced back to fundamental differences in the weighting of these moral dimensions. This would suggest that there is a "conservative" way of thinking and, separately, a "liberal" way of thinking, each of which should be reflected in inferred network topology. That is, I would expect to see distinctive conservative network structures that meaningfully differ from liberal network structures. Specifically, I would expect that respondents who score high on the moral foundation measures of purity and authority – which are associated with conservative thought (Haidt, 2012) – would produce network structures similar to those who are ideologically conservative (Center, 2017). I would further expect this conservative thought to be reflected in heterogeneous networks with dissassortative connectivity. These characteristics would indicate more hub-and-spoke like networks, where concepts differ significantly in importance (as measured by degree) and high-degree nodes tend to connect to low-degree nodes. Such networks are representative of the utilitarian view (Sidgwick, 1907), where a few core rules dictate judgments. On the other hand, I would expect respondents who score high on the moral foundations composite score of progressivism – characterized by the traditionally liberal traits of aversion for harm and concerns about fairness – to have networks whose structure is more in line with particularist moral theories (Dancy, 1993). Such structures would be heavily interconnected, allowing for flexible, context-aware moral reasoning. This would be reflected by networks with higher average degree and a single connected component. These networks may be more homogenous, indicating the lack of any dominating, central ideas, or they could have densely connected cores surrounded by a sparse periphery – in which case, we would still see high standard deviation and dissasortativity. These specific measures and their implications for different conceptions of moral

"coherence" are discussed further in Section 4.3.

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

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

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

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

This could result in more connected networks as well as more content on average.

While this paper is largely exploratory, given the relative paucity of modern work in this area, the personality measures used here were specifically chosen in order to evaluate the potential of conceptual networks to provide meaningful behavioral insights. Many of our priors are conflicting; for example, it is unclear whether we would expect someone with more ideas (nodes) to have denser networks (more edges) or sparser networks (relatively fewer edges). Thus, in evaluating the results of these data – which will be discussed in Section 5 – I will look primarily for overall trends and clusters of network characteristics which appear to behave similarly.

3.2 YouGov Ideological Turing Test

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

Based on the evaluation of human coders, roughly half (56%) of respondents participated in good faith and tried to genuinely argue both sides of their assigned issue. Interestingly, the other half of respondents did not generally submit pure linguistic nonsense, but rather made inauthentic arguments which were merely caricatures of opposing views. For example: "'Bomb everybody who disagrees with us," or "It is okay to murder a fetus, as long as a gun is not involved!" In other words, many of these non-compliant participants did technically submit both liberal and conservative arguments, though some of their arguments – particularly those which didn't align with their own ideology – were of low quality. This presents a particularly challenging but interesting NLP problem: from a purely linguistic point of view, there is nothing wrong with these texts; they make

perfect grammatical sense. However, they are poor arguments in a more meaningful sense – they offer no evidence or justification, and may not have a coherent premise.

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

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

4 Methods

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

In this section, I will theoretically motivate the node and edge representations in a conceptual

network, and describe my method for inferring these constructs. Then, I will describe the challenges of network measurement and present a number of tools to measure and compare inferred network structures.

4.1 Inferring concepts

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

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

Fortunately, this is exactly the motivation behind word embeddings (Mikolov et al., 2013). Trained on vast corpora of data, words can be embedded in high-dimensional space representing the contexts in which those words appear. Words which are similar – or, more precisely, which occur in similar contexts (Firth, 1957) – will then appear close together in this space and can be clustered into concepts. Because training such models requires enormous quantities of data, I use a publically available dataset of pre-trained embeddings which have been found to be appropriate for most tasks (Spirling and Rodríguez, 2019). Specifically, I use embeddings pre-trained on 100 billion words from the Google News corpus (Mikolov et al., 2013). This dataset embeds words in 300-dimensional space

by maximizing the average log probability

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

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

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

4.2 Inferring connections

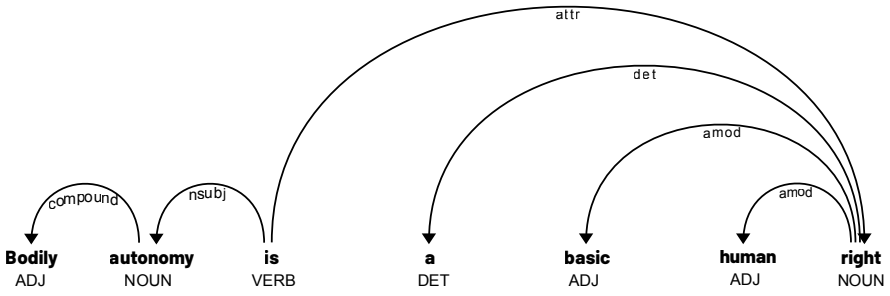


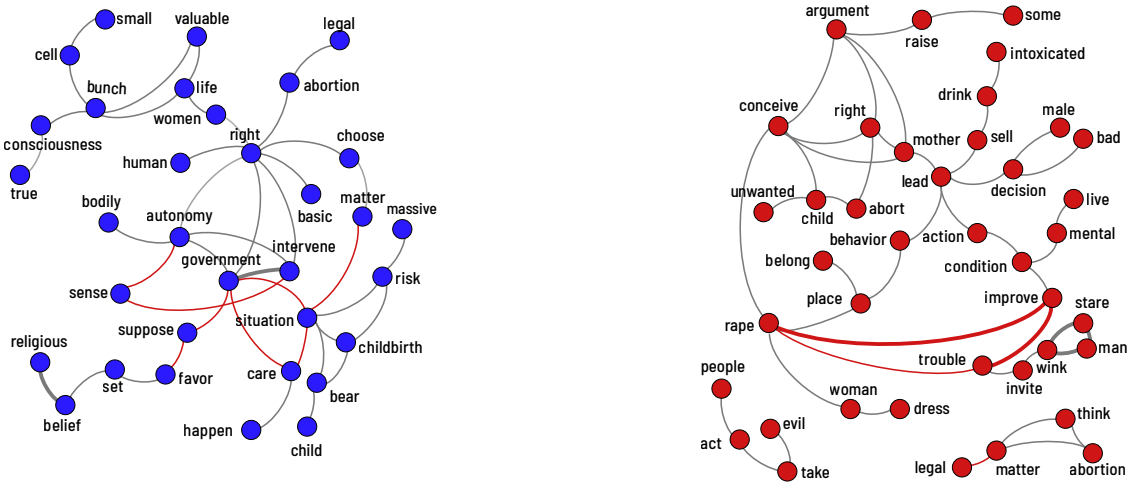
Figure 1: Example of the grammatical parse of a sentence

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

must be syntactically near in order to be related.

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

I therefore propose an approach which leverages grammatical structure in order to determine conceptual relations. Specifically, I determine the grammatical parse of a text by identifying each word's part of speech and their syntactic dependency relations. In future work, I plan to integrate semantic as well as syntactic parsing. This syntactic parse identifies the grammatical relations between words, linking, for example, adjectives to the nouns they modify and subjects to their related objects. An example grammatical parse can be seen in Figure 1. Importantly, these grammatical connections can be meaningfully interpreted – indeed, the very purpose of these grammatical rules to serve as a tool to help humans encode and decode linguistic communication.



(a) Example liberal position on abortion

(b) Example of conservative position on abortion

Figure 2

While the grammatical parse serves as the network's foundation, this structure is modified through the process of inferring concepts described above. When terms, such as stops words, are removed

from the network, any remaining parent and child nodes are connected in their place. Any concept which occurs multiple times – either through the repetition of a word, use of a referent word, or through conceptually similar words – are taken to be the same node, with all their external links shared. Additionally, negative words (such as "not") are removed and replaced with a negative tie between grandparent and child terms. These steps result in a weighted, signed network of conceptual interrelations. Example networks using these methods to infer concepts and their relations from text can be seen in figure 2.

4.3 Network measures

There are many methods of network comparison, but these frequently rely upon networks having the same content (eg, nodes and node labels), and measure network distance as differing patterns of interconnection. Here, the networks generated across individuals do not necessarily share any of the same nodes and we are primarily interested in how structure varies, independent of variations in content.

Portrait divergence Bagrow and Bollt (2019) provides one antidote to this, allowing for pairwise comparison of arbitrary networks. This approach defines a graph's portrait as an asymmetric matrix, B, in which the B_{kl} entry captures the number of nodes k which have path length l. This portrait contains both high and low dimensional information about the structure of the network. The network's degree distribution, for example, is captured by the first row ($B_{1,k}$), while the shortest path distribution is encoded as $\frac{1}{2}\sum_{k=0}^{N} kB_{l,k}$. This matrix is normalized to the row-wise cumulative distribution of B, and the similarity between two networks is calculated as the Kolmogorov-Smirnov test statistic K_l, e.g., the maximum distance between the two matrices.

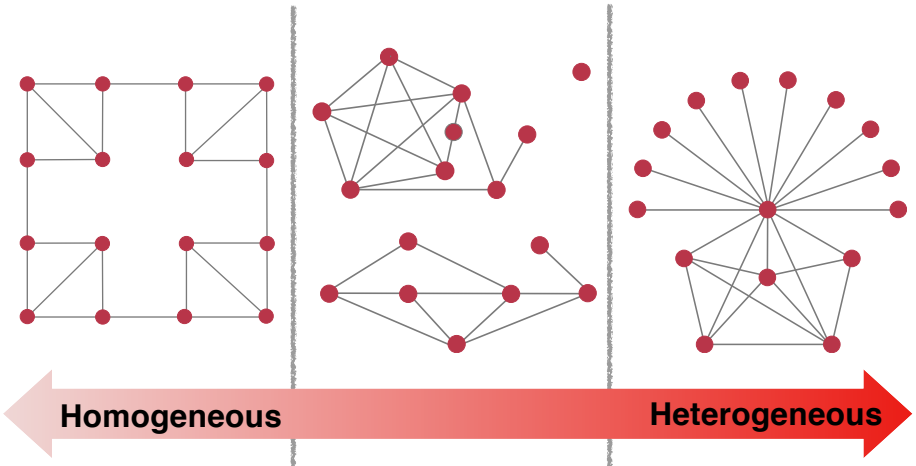
While I will use portrait divergence to measure pair-wise similarly between networks in Section 5.2, this approach does not allow for a fine-grained understanding of the ways in which dissimilar networks are structured. I therefore also describe the topology of inferred networks through seven network measures that can compare resulting structures across several key dimensions. Building off the moral philosophy literature, I engage measures which capture connectivity, complexity, and hierarchy – each representing different understandings of what good moral "coherence" should look like.

Connectivity serves as the baseline for coherence, and is measured here through the percent of nodes which are in the giant component. Complexity is suggestive of the particularist view (Dancy, 1993) of coherence which advocates for richly interconnected and resilient networks. This approach is captured through the measures of average degree, clustering, and density. Finally, hierarchy reflects the view utilitarianism as advanced by Sidgwick (1907) and covers measures of entropy, disassortativity, and standard deviation of degree. These measures are described in detail in Table 1.

Connectivity (Baseline)	
Giant component percent	The percent of nodes in the largest component of the network, N_G/N. This measure indicates how cohesive the network is. A value of 1 indicates the network has a single component (e.g., a path exists between any two nodes), while lower values indicate that the network has multiple, disconnected components.
Complexity (Dancy, 1993)	
Average Degree (k avg)	The average degree across all nodes in the network. Higher values indicate that nodes have more connections on average.
Clustering	A measure of how locally-connected a network is. High values indicate triadic closure (Saramäki et al., 2007), while low values indicate a network that is locally tree-like.
Density	The ratio of existing edges to the total possible edges, $2E/(N(N-1))$. This is a measures the overall interconnectivity of a network with a value of 1 indicating that every idea is connected to every other idea and a value of 0 indicated that no concepts (nodes) are connected.
Hierarchy (Sidgwick, 1907)	
Entropy	Estimates the amount of information contained in the network's normalized degree distribution (p_k) (Shannon, 1948). This measure is dependent on both the length of the distribution (eg, N) and the heterogeneity of the distribution. Measured as $-\sum(p_k \times \log(p_k))$
Disassortativity	Measured as the inverse of the Pearson correlation coefficient, -r, disassortativity captures the degree homophilly of the network (Newman, 2003). Values range from -1 to 1, with a value of 1 indicating that high degree nodes tend to connect to low-degree nodes, as in a star network. Note that for this study, we use disassortativity in order to have the same dimension and valence as standard deviation.
Standard deviation of degree (k std)	The standard deviation of the network's degree distribution. Lower numbers indicate that nodes are more homogeneous in their degree whereas larger values indicate greater difference between the lowest and highest degree nodes in the network. For the purposes of this study, k_{std} is normalized against a hub-and-spoke network with the same number of nodes N.

Table 1: Measures of network structure.

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



giant component	1.0	0.4	1.0
k avg	3.0	3.0	3.0
clustering	0.2	0.4	0.3
density	0.2	0.2	0.2
entropy	0.0	1.5	1.0
disassortativity	-1.0	-0.1	0.7
k std	0.0	3.0	3.5

Figure 3: Comparison of network statistics across three stylized example networks. Each network has $N = 16$ and $E = 24$

To provide a more intuitive sense of what these measures indicate, Figure 3 compares these statistics across three stylized example networks. Each network has a fixed number of nodes ($N = 16$) and edges ($E = 24$) – resulting in equal density ($d = 0.2$) – and are constructed to have equal average degree ($k = 3$). However, these networks display strikingly different topological properties,

which are conveyed through our additional network statistics. In particular, we see that higher standard deviation and higher disassortivity are both indicative of heterogeneous, hub-and-spoke like structures. Entropy provides a weakly opposite indicator with homogeneous networks having slightly higher entropy than heterogeneous networks. Given that entropy is calculated as $-\sum(p_k \times \log p_k)$, the minimal effect of variations in degree distribution suggest that higher entropy is more likely to be indicative of higher node count. The final two network measures, giant component percent and clustering, each provide unique topological insight not captured by the other network measures. Specifically, the giant component percent indicates whether a network is connected or fractured into multiple components, while clustering indicates the presence of triangles – eg, the tendency of nodes which share a neighbor to themselves be connected.

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

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

5 Results

This paper presents a method for inferring the latent network structure of concepts within textual documents. While the literature clearly supports the theoretic motivation for the existence of such latent network structure, it remains to be seen whether there is value in developing such a method. I therefore demonstrate the value and implications of this approach through three applications. First, using the Mechanical Turk dataset, I demonstrate that the network structure inferred from individual's text is meaningfully correlated with ideology as well as known latent personality traits. This suggests that expressions of political preferences – not just the preferences themselves – are tied

to behavioral traits. Second, I use the YouGov dataset to illustrate that these correlations aren't simply ideological and more reflective of ideological correlates than simple partisan talking points. Finally, I use the same dataset to show that the method presented here can distinguish between ironic responses and texts which authentically reflect an ideological view. This further suggests that there is a meaningful signal within the very structure of these expressed reasons.

5.1 Personality and Reasoning

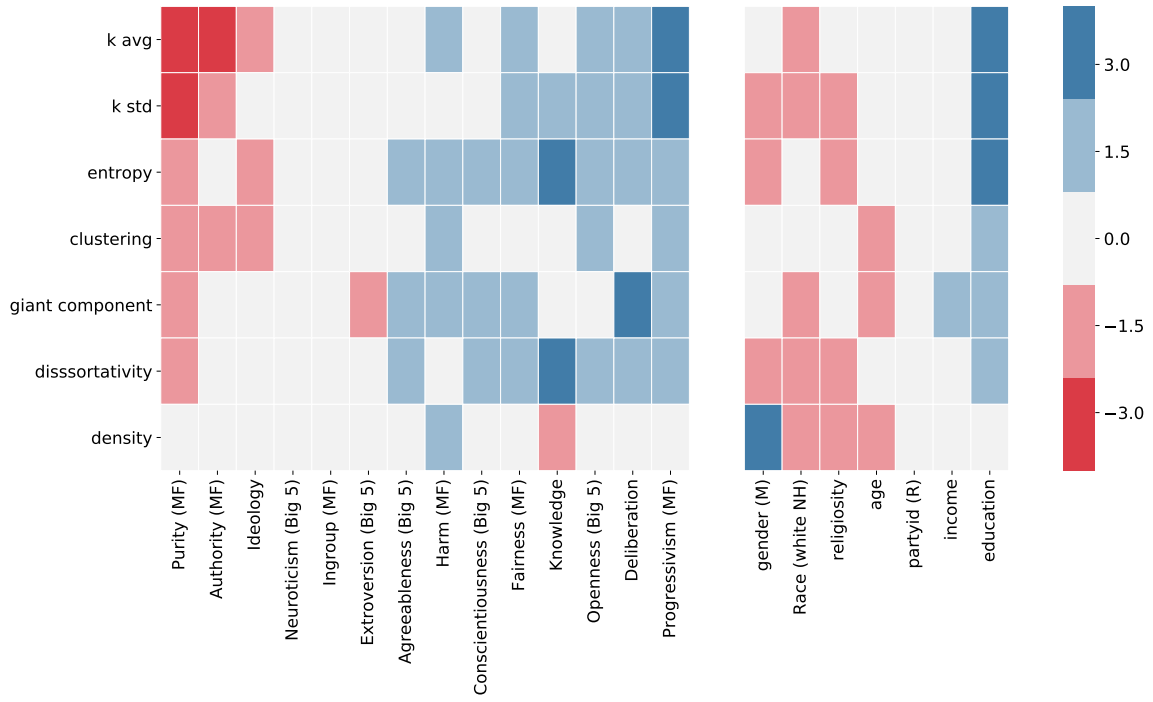


Figure 4: Correlations between network statistics and latent personality and demographic measures

As described in Section 3, subjects in the Mechanical Turk study completed three network elicitation activities for two issue area prompts. Therefore, in order to assess the relationship between network statistic s and personal trait p, I employ a multilevel model which includes topic and method random effects:

$$s = \beta\, p + \alpha_t + \alpha_m + \epsilon \tag{2}$$

The resulting correlations between inferred structure and personality measures are shown in Figure 4.

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

Specifically, we see that "progressive" subjects tend to create networks which have higher standard deviation of degree (k std), entropy, average degree (k avg), clustering, and disassortativity while "conservative" subjects tend to be lower on each of these dimensions. As illustrated by the example networks in Figure 2, higher values of k std and disasortativity suggest more heterogeneous, hub-and-spoke like networks. On the other hand, higher values of k avg and clustering suggest more interconnected networks, while high entropy suggests either more homogeneous structure or more content (nodes). Taken together, this combination of network statistics suggests that progressive subjects tend to form networks with a core-periphery structure – that is, networks with an interconnected core of central ideas surrounded by a periphery of loosely connected auxiliary ideas. The weak signal sent by the density metric is further suggestive of this, as a network with a dense core and sparse periphery would have a non-remarkable density on average. While the the higher values of k std, entropy, and dissasortativity are suggestive of hub-and-spoke type networks and may be reflective of Sidgwick (1907)'s utilitarian view of coherence, the higher values of k avg and giant component percent suggest these networks have a richer, more interconnected structure which may be more in line with the particularist philosophy (Dancy, 1993).

Conservative subjects, on the other hand, produce networks with lower k std, k avg, entropy, and

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

While this analysis suggests meaningful correlations between personality and expressed structure, it cannot disambiguate between possible effects. These same personality traits are believed to lead to ideological positions Haidt and Joseph (2008), making it unclear whether variations in structure are indeed an indication of personality or more generally a refection of a given position's talking points.

5.2 Reasoning Fingerprint

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

In the YouGov dataset, each respondent provides two, ideologically opposed reasons which have been judged by a human coder to be authentic attempts to represent those points of view. We can therefore ask whether individuals tend to produce networks similar to themselves or similar to others within the same category. That is, will the structure of C_i, the conservative essay produced by respondent i, be more similar to that same user's liberal structure, L_i, or to the structure C_j: user j's conservative essay on the same topic? If C_i and C_j are more similar, it suggests that any structural features are driven in someway by the content; e.g., that conservative arguments are have

similar structure regardless of who is doing the arguing. If C_i and L_i are more similar to each other, it suggests that there is some individual argument style – that i will produce similar structures across dissimilar topics. Finally, we may find no patterns in similarity, suggesting that their is neither an individual nor ideological signal within the inferred structure of text.

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

This creates a distribution of self-comparison scores as well as a distribution of each subject's comparison to the rest of the subject pool. These distributions are shown in Figure 5.

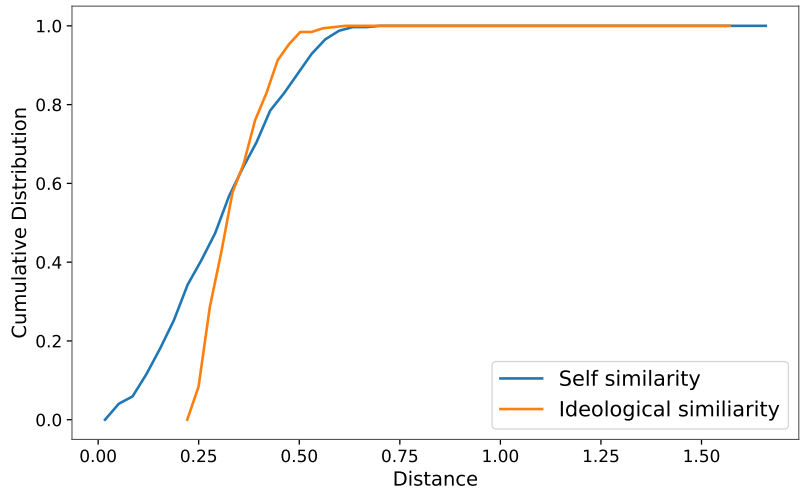
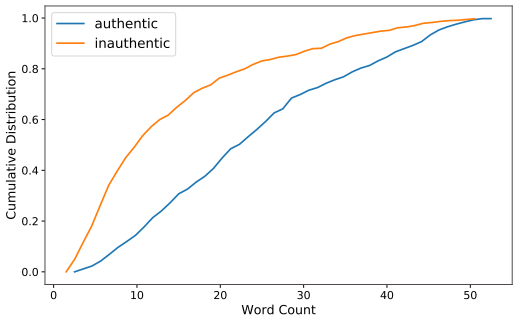
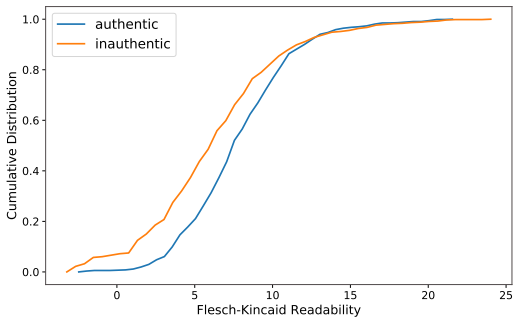


Figure 5: Distribution of network similarity for authentic respondents. "Self" captures similarities between a single respondent's liberal and conservative text, while "Other Respondents" captures within-topic similarity. A similarity of 0 indicates that networks have identical portraits.

As we can see, networks inferred from a single individual are slightly more likely to be more similar than networks inferred from different individuals. A t-test shows that this difference is significant ($p < 0.05$). While there may be other individual correlates, such as education, driving this result,



(a) Distribution of word count in respondent text



(b) Distribution of Flesch-Kincaid Score in respondent text

Figure 6: Distributions of common quality metrics between authentic and inauthentic respondents

this finding does suggest that respondent essays do not merely reflect ideological talking points. Particularly on subjects like abortion, when ideological views are very established and well known, it is entirely possible that conceptual network structure would have been largely driven by ideology. This finding therefore suggests that the inferred network structure is capturing something about individual style or preference. Put differently, the structure of reasoning appears to be an individual characteristic rather than a topical or ideological one.

5.3 Authenticity and Irony

As described in Section 3, nearly half of respondents in the YouGov dataset provided inauthentic, but linguistically meaningful answers. We can therefore aim to separate authentic from inauthentic answers using our inferred structural features. If such a classification can be done, it suggests that the inferred networks are indeed meaningfully encoding the latent structure of the text. I compare two approaches for this task.

First, as a baseline measure, I consider two common measures of text sophistication: word count and Flesch-Kincaid readability. A text's Flesch-Kincaid score is calculated based off the number of words, sentences, and syllables in a text, with higher scores indicating more complicated texts. Figure 6 shows the cumulative distribution of these measures within both the authentic and inauthentic responses. Here we see that inauthentic responses do tend to have few words, but not dramatically so. Texts in both samples have nearly identical Flesch-Kincaid scores, suggesting this will not be a

helpful feature for separating these categories.

My second model considers the inferred network structure of the text, using the network measures described in Section 4.3, and a third model includes all features from Models 1 and 2.

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

Looking at the effects of each feature, we see that word count is indeed driving the performance of Model 1, with the Flesch Kincaid score producing a small and insignificant result. In Model 2, we see that several network features appear to encode a signal supporting the classification. Specifically, the average degree and density both help indicate whether or not a response is authentic. In Model 3, we see that these effects continue when all features are considered, with word count, average degree, and density all containing information.

Specifically, as we would expect, texts with more words are more likely to be authentic. From the network measures, we further see that higher average degree is more likely to indicate authentic texts while networks with higher density are less likely to indicate authentic texts. Noting that word count is highly correlated with the number of nodes in a network, this suggests that authentic networks are characterized not only by more content, but by content with meaningful interconnections. That is, nodes on average have higher degree, but, due to the higher number of nodes, authentic networks are overall less dense. Ironic responses, on the other hand, are characterized by denser networks with overall less content. While highly interconnected, these networks are limited in how many other nodes can be connected to, resulting in a lower average degree than the authentic responses.

One way to interpret these results is to think of these network features illustrating the strength of an argument in the Aristolelian sense. An argument structured as a major premise followed by a minor premise (Toulmin, 1958), would have network characteristics similar to what we see in the authentic arguments: multiple points which are strongly, but not overly densely, connected to each

[2]With an 80% in-sample, 20% out-sample split.

other. This could suggest that ironic responses are lacking any such minor premise and rather state a simple and densely connected claim without providing any justification.

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

6 Discussion

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

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

Table 2: Results of logistic regressions.

	(1)	(2)	(3)
word count	3.122***		1.189**
	(0.295)		(0.475)
Flesch Kincaid	0.480		-0.357
	(.450)		(.509)
giant component		-0.543	-0.260
		(0.394)	(0.409)
k avg		3.011***	2.672**
		(1.100)	(1.129)
clustering		0.610	0.705
		(0.655)	(0.659)
density		-2.612***	-2.273***
		(0.494)	(0.518)
entropy		-0.111	0.170
		(0.614)	(0.624)
disassortativity		0.058	0.198
		(0.390)	(0.396)
k std		-0.876	-1.711
		(0.918)	(0.986)
Notes:			*p<0.1; **p<0.05; ***p<0.01
			Model (1): 70% predictive accuracy
			Model (2): 74% predictive accuracy
			Model (3): 74% predictive accuracy

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

Specifically, this paper has presented a text-based method for inferring conceptual network structure.

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

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

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

too surprised that individual covariates play an important role here, this serves as an important reminder that there is more than mere ideology at play.

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

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

References

Axelrod, R. (1976). *Structure of decision: The cognitive maps of political elites*. Princeton university press.

Bagrow, J. P. and Bollt, E. M. (2019). An information-theoretic, all-scales approach to comparing networks. *Applied Network Science*, 4(1):45.

Berker, S. (2015). Coherentism via graphs. *Philosophical Issues*, 25(1):322–352.

Campbell, Angus; Converse, P. E. M. W. E. S. D. E. (1960). *The American Voter*. University of Chicago Press.

Carley, K. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology*, 23:75–126.

Carpini, M. X. D. and Keeter, S. (1993). Measuring political knowledge: Putting first things first. *American Journal of Political Science*, pages 1179–1206.

Center, P. R. (2017). Are telephone polls understating support for trump? Report, Pew Research Center.

Christen, M. and Ott, T. (2013). Quantified coherence of moral beliefs as predictive factor for moral agency. In *What Makes Us Moral? On the capacities and conditions for being moral*, pages 73–96. Springer.

Cohen, R. (1987). Analyzing the structure of argumentative discourse. *Computational linguistics*, 13(1-2):11–24.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Dancy, J. (1993). *Moral reasons*.

Dorsey, D. (2006). A coherence theory of truth in ethics. *Philosophical studies*, 127(3):493–523.

Dorsey, D. W., Campbell, G. E., Foster, L. L., and Miles, D. E. (1999). Assessing knowledge structures: Relations with experience and posttraining performance. *Human Performance*, 12(1):31–57.

Firth, J. R. (1957). *Studies in linguistic analysis*. Wiley-Blackwell.

Gastil, J., Knobloch, K., and Kelly, M. (2012). *Evaluating Deliberative Public Events and Projects*, book section 10. Oxford University Press, Oxford; New York.

Grice, H. P. (1975). Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haidt, J. and Joseph, C. (2008). *The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules*, volume 3. Oxford University Press.

Hong, L., Page, S. E., and Baumol, W. J. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 101, pages 16385–16389.

John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

Knobloch, K. R., Gastil, J., Reedy, J., and Cramer Walsh, K. (2013). Did they deliberate? applying an evaluative model of democratic deliberation to the oregon citizens' initiative review. *Journal of Applied Communication Research*, 41(2):105–125.

Lane, R. E. (1962). *Political ideology: why the American common man believes what he does*. Free Press of Glencoe.

McCrae, R. R. and Costa Jr, P. T. (1999). A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(1999):139–153.

McNaughton, D. and Rawling, P. (2000). Unprincipled ethics. *Hooker and Little*, 2000:256–275.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Systems Research and Behavioral Science*, 12(5):410–430.

Rawls, J. (1993). *Political Liberalism*. John Dewey essays in philosophy. Columbia University Press.

Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K., and Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shapiro, S. C. and Rapaport, W. J. (1987). Sneps considered as a fully intensional propositional semantic network. In *The knowledge frontier*, pages 262–315. Springer.

Shavelson, R. J. (1974). Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching*, 11(3):231–249.

Shugars, S., Beauchamp, N., and Levine, P. (2019). Mapping conceptual networks. Working paper.

Sidgwick, H. (1907). *The methods of ethics*. Hackett Publishing.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Spirling, A. and Rodríguez, P. L. (2019). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. working paper.

Thagard, P. (1998). Ethical coherence. *Philosophical Psychology*, 11(4):405–422.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Walton, D. N. (1996). *Argumentation Schemes for Presumptive Reasoning.* L. Erlbaum Associates.

Zaller, J. R. et al. (1992). *The nature and origins of mass opinion.* Cambridge university press.