Uncovering Social Semantics From Textual Traces: A Theory-Driven Approach and Evidence From Public Statements of U.S. Members of Congress

Yu-Ru Lin
School of Information Sciences, University of Pittsburgh, 710 IS Building, University of Pittsburgh 135 North Bellefield Avenue, Pittsburgh, PA 15260. E-mail: yurulin@pitt.edu

Drew Margolin
Department of Communication, Cornell University, Ithaca, NY 14580. E-mail: dm658@cornell.edu

David Lazer
Department of Political Science and College of Computer and Information Science, Northeastern University, Boston, MA 02115; John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138. E-mail: d.lazer@neu.edu

The increasing abundance of digital textual archives provides an opportunity for understanding human social systems. Yet the literature has not adequately considered the disparate social processes by which texts are produced. Drawing on communication theory, we identify three common processes by which documents might be detectably similar in their textual features—authors sharing subject matter, sharing goals, and sharing sources. We hypothesize that these processes produce distinct, detectable relationships between authors in different kinds of textual overlap. We develop a novel n-gram extraction technique to capture such signatures based on n-grams of different lengths. We test the hypothesis on a corpus where the author attributes are observable: the public statements of the members of the U.S. Congress. This article presents the first empirical finding that shows different social relationships are detectable through the structure of overlapping textual features. Our study has important implications for designing text modeling techniques to make sense of social phenomena from aggregate digital traces.

Introduction

The quantity of human behavior that is stored and observable as raw, unstructured text is rapidly increasing. One of the most promising methods for analyzing text is the use of discursive isomorphism—the identification of texts with similar features or structures—to infer relationships between the authors of the texts (Bramsen, Escobar-Molano, Patel, & Alonso, 2011; Bryden et al., 2013; DiMaggio & Powell, 1983). In this paper we propose a technique for using n-gram similarity to infer the kind of relationship that is likely to exist between authors; specifically, whether the authors are addressing the same subject matter, share rhetorical goals, or consult similar sources of information. We argue that while these relationships are not mutually exclusive each is produced by a distinct theoretical process that can vary and be influenced independently. We then argue that the presence and intensity of these distinct relationships is most detectable in similarities between authors at different lengths of phrases or n-grams. We then test this hypothesis by using the public statements of the members of the U.S. Congress from January 2011 through August, 2012.

Textual data can indicate “invisible networks”—underlying interpersonal, ideological, structural, or other relationships between people that would otherwise only be observable through close observation of and interaction with them (Lin, Margolin, & Lazer, 2014). A number of techniques have been developed to infer these social properties such as the extent to which individuals share an organizational culture (Barnett et al., 1981; Danowski, 1982), members of teams share mental models (Carley, 1997), ideological and attitudinal similarity between individuals (Adamic, Lento, Adar, & Ng, 2014; Danowski, 2011),
common topics and ways of speaking (Griffiths, Steyvers, Blei, & Tenenbaum, 2005; McCallum, Wang, & Corrada-Emmanuel, 2007; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), or group identities (Tamburrini, Cinnirella, Jansen, & Bryden, 2015).

Yet despite the proliferation of techniques there has been little work asking how or why observed isomorphism is achieved. The result is that much research uses similar techniques to infer different relationships and properties. For example, in some research, individuals who use the same words are inferred to share topics of discussion (Rosen-Zvi et al., 2004), whereas in others they are inferred to identify with and conform to similar group norms (Tamburrini et al., 2015). Authors may also produce similar texts because there are incentives for them to imitate one another or a common third party directly (Adamic et al., 2014), suggesting they share access to common sources of information and incentives to save cost or demonstrate a unified front (DiMaggio & Powell, 1983).

The distinction between these relationships is often important. Understanding which process produced the observable relationship between authors can help researchers to identify the relevant theories and boundary conditions for understanding the authors’ social behavior. In political discussions, for example, textual isomorphism due to conformity is likely to indicate group-based polarization (Adamic & Glance, 2005), whereas textual isomorphism due to topic sharing can indicate that individuals are engaging with the needs and demands of the public (Boydston, Glazier, & Pietryka, 2013; Gentzkow & Shapiro, 2010). Isomorphism due to imitation or mimetic processes can indicate a lack of research resources (Gandy, 1982; Grimmer, 2009) or deference to authority (Glazier & Boydston, 2012) or norms that encourage copying (Cordell, in press). Similarly, within an organization isomorphism can occur because of a faction of like-minded individuals (Carley, 1997) or a group that participates on topics with highly regulated, restricted, or ritualized discourse (Hart, Childers, & Lind, 2013; Meyer & Rowan, 1991)?

The meaning and expected endurance of the observed relation depend on the social process that is assumed to produce the textual similarity. Clusters based on ideologically polarized are unlikely to evolve in the same way as clusters produced by imitation driven by resource shortages, nor can they be influenced or corrected with the same remedies. A collective convergence on a common subject of attention may be difficult to sustain, whereas a collective effort to hold an ideological position may be difficult to overcome.

Current techniques do not adequately distinguish these relationships either theoretically nor empirically. We address this gap by articulating distinct social processes: authors sharing subject matter, sharing rhetorical goals, and copying from shared sources; each of which should lead to discursive isomorphism across authors and each of which has distinct implications for the way in which this isomorphism is manifest. Specifically, we argue that each process will produce similarity between authors most distinctly at a different length of n-gram. We then analyze texts produced by members of the U.S. Congress by comparing their observable discursive isomorphism to other, nondiscursive relationships between them. Our work thus takes a first step in computationally distinguishing different kinds of textual similarity. The main contributions of this work are:

1. We provide theoretical justification for distinguishing three processes by which similar texts are authored.
2. We develop a randomized n-gram extraction algorithm to capture text similarity based on n-grams of different lengths that effectively tackles the computational challenge of counting vast numbers of arbitrarily long n-grams based on a document-pair sampling strategy.
3. We conduct a study on a text corpus produced by members of the U.S. Congress, demonstrating that the distribution of n-gram similarity between the members’ statements is tri-modal. We then show that each mode is most closely associated with different relationships known to exist between members of Congress: sharing topics, sharing rhetorical goals, and sharing sources.

The rest of the paper is organized as follows. After a review of related work we identify and define processes by which similar texts can be produced and describe the empirical implications of each process’ operation. We then describe the randomized n-gram extraction procedure that allows the comparison of document texts on n-grams of arbitrary length. We present our empirical results for the public statements of U.S. Congress members. Finally, we provide a discussion and conclusion.

Related Work

Previous research has emphasized only a single textual production process at a time for a couple of reasons. First, analysis techniques have often been developed to be applied to specific kinds of textual data where one process is already known or assumed to be dominant. Second, most techniques tend to simplify the relationships between authors’ thoughts and beliefs and the thoughts and beliefs of their audiences, thus underemphasizing both strategic processes (Atteveldt, Kleinnijenhuis, & Ruigrok, 2008; Eisenberg, 1984), where words are chosen for their effect, and mimetic processes, where they are chosen for their convenience (DiMaggio & Powell, 1983).

For several decades communication researchers have analyzed structural features of documents using semantic networks (Danowski, 1982; Doerfel & Barnett, 1999). A semantic network is derived by considering words as nodes and co-occurrences or meaning relations between the words as links (Carley & Kaufer, 1993). Early semantic network analysis was conducted on texts that were authored in response to surveys where subjects were asked by researchers to describe meaningful relationships between concepts (e.g., Barnett et al., 1981; Carley, 1997; Rice & Danowski, 1993). In this context, where subjects are asked to describe
their cognitions to a researcher in a confidential survey, researchers reasonably assumed that if two individuals produced similar documents the similarity was due to these individuals sharing cognitions, beliefs, or worldviews, as summarized by Danowski (2011, p. 226): “individuals with more similar semantic networks based on their encoding of messages are likely to perceive the world more similarly and behave more similarly.” Similarly, Doerfel and Barnett (1999, p. 589) state that semantic network analysis “focuses on the structure of a system based on shared meaning.” Consistent with this assumption, this technique is generally used to measure how groups perceive the world, such as the “collective consciousness” (Barnett et al., 1981) within an organization or “mental models” used by team members (Carley, 1997).

More recently semantic network analysis has been applied to the observation of natural discourse (e.g., Doerfel & Connaughton, 2009; Rice, 2005; Yuan, Feng, & Danowski, 2013). In these natural contexts the assumption remains, however, that when authors create texts they are trying to clearly indicate meaningful relationships between words, that is, that their communication behavior is a “conduit” of their thoughts (Rogers & Kincaid, 1981). While this assumption is reasonable in some contexts, it is problematic in contexts where there are incentives for strategic communication. In these strategic contexts individuals have an incentive to construct their statements informed by the meanings and cognitions of their audience precisely because it is different from their own. For example, Danowski (2011) describes the creation of “optimal messages” that “steer” the audience “away from certain words or towards them” (p. 229). Authors may also use the fact that some meanings are not shared widely in the audience to create strategically ambiguous messages that mean one thing to one subgroup in the audience and another thing to another (Eisenberg, 1984). In these cases two authors may use the same combinations of words not because these combinations naturally reflect their own (shared) cognitions but because they each predict they will have a similar effect on their audience.

Work in information science and linguistics has also tended to treat texts as neutral “conduits” of conceptual meaning. Researchers have developed a variety of computational techniques for extracting the relationship between “concepts” and “word meanings” such as latent semantic analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997) and probabilistic models such as probabilistic latent semantic analysis (pLSA) (Hofmann, 2001) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). By design, these approaches seek stable, context-invariant relationships between concepts. Strategic uses of language, where these typical relations are exploited to lead audiences to atypical conclusions (Carley & Kaufer, 1993), are largely not considered (one exception is O’Connor [2012]).

The LSA approach has more recently been extended to the topic model technique (e.g., Blei, 2012; Blei et al., 2003; Hofmann, 2001; Wallach, 2006). The LDA (Blei et al., 2003) topic model introduces the idea that documents are mixtures of topics (like “concepts” in LSA or “aspects” in pLSA). Because the aim of topic modeling is the reliable classification of documents, the technique does not make assumptions about why words, or larger n-grams (Wang, McCallum, & Wei, 2007), co-occur (e.g., because they have meaningful relationship) but focuses more exclusively on the statistical fact of this co-occurrence.

Although topic models do not make assumptions about the relationship between word use and an author’s cognition, they are vulnerable to mischaracterizing similarities due to mimetic processes. When a set of words (or longer n-grams) frequently co-occur it is possible topic models assume this is because of a systematic statistical tendency for these individual words to be drawn conditional on the presence of the “topic.” But it is also possible that their co-occurrence is explained by the wholesale repetition of a chunk of text in which they are co-present. This mimetic isomorphism in text has been observed in numerous cases, in the blogosphere (Leskovec, Backstrom, & Kleinberg, 2009; Simmons, Adamic, & Adar, 2011), social media (Adamic et al., 2014), political news reporting (Grimmer, 2009), discussions on Wikipedia and the Supreme Court (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012), as well as within the U.S. Congress (Hassan et al., 2008).

An additional area of research focuses strictly on the detection of mimetic processes in the sense of “plagiarism” (Barrón-Cedeño & Rosso, 2009; Grozea, Gehr, & Popescu, 2009; Lyon, Barrett, & Malcolm, 2004; Lyon, Malcolm, & Dickerson, 2001). These techniques usually rely on the 2- to 7-gram feature analysis to identify candidate documents as the source of the copy based on the low statistical likelihood that independently written texts will have a large number of matches in long n-grams (Lyon et al., 2004) (or longer n-grams). Quinn, Monroe, Colaresi, Crespin, and Radev (2006) applied this technique to the U.S. Congressional Speech corpus and studied the importance of speakers based on how their speeches were similar to other speeches. These techniques presume that the repetition of features are the result of copying. There is evidence, however, that in some social contexts highly restrictive social norms can lead to texts that are highly repetitive even without the presence of copying (e.g., corporate annual reports [Hart et al., 2013]).

**Distinguishing Processes**

**Conceptual Distinction**

As described above, previous analysis techniques have tended to assume that similarities are the result of one particular process even though others can plausibly explain their results. One reason for this may be that a set of candidate semantic organizing processes, the processes by which a set of authors can produce similar texts, have not been conceptually distinguished. Here we distinguish these processes by focusing on the variables that most consistently...
trigger or influence the intensity of each process independent of the others. We expect that in real-world settings these processes often operate at the same time; however, these distinct triggers can lead them to be dominant or subordinate in different contexts. To illustrate these theoretical concepts we provide examples from the corpus we use in our analysis.

The first organizing process can be described as constrained searching (March & Simon, 1993; Margolin & Monge, 2013). In constrained searching, authors look for and consider only words, concepts, phrases, or longer n-grams from a restricted population. Examples of restricted populations include vocabularies that are appropriate to specific subject matter. A vocabulary is a preexisting set of terms and phrases with clear, stable definitions (Loewenstein, Ocasio, & Jones, 2012). Linguistic or cultural norms may also restrict authors to draw from particular sets of terms and phrases. In the context of a discussion of certain ideas, particular concepts may be taboo (Hart et al., 2013; Meyer & Rowan, 1991). Thus, in some settings the fact that a set of terms has been used in one statement may constrain those responding to that statement to use similar terms (Danescu-Niculescu-Mizil et al., 2012). For example, on June 28, 2012, the Supreme Court released its ruling that the Affordable Care Act (colloquially known as “Obamacare”) does not violate the Constitution. On this day, 369 members of Congress (161 Democrats and 208 Republicans) released public statements that addressed this topic. In discussing the Supreme Court’s decision regarding the ACA, members who both supported and opposed the law used healthcare-related terms such as “insurance,” “doctor,” and “affordable.”

Searching processes should be distinguished from selection processes (Kauffman, 1993). In searching processes, authors seek to create distinct statements, but are constrained by external forces to construct those statements from common materials, leading them to be similar. In selection processes, authors are attempting to construct statements that are similar in terms of their meaning or effect. They construct statements and evaluate them in terms of their fit with their goals or intentions, editing the statements to produce those that most closely meet their criteria. Thus, similarity in selection reflects similarity in underlying attributes of the individuals, such as mental models or beliefs, whereas similarity in searching reflects similarity in a situation, context, or predication that authors face.

In the case of the Supreme Court decision, we observe a clear partisan split in how the bill is referred to. Of the 161 Democrats who spoke about the Supreme Court’s decision and thus referred to the law in some way, 102 (63%) explicitly named the law “the Affordable Care Act” (ACA) whereas only 18 (9%) of the Republicans used the official name. This difference is not due to a constraint on searching. Republicans did not eschew the vocabulary terms that comprise the name of the law: 52% say “affordable”; 100% say “care”; and 72% say “act” at least once. They simply avoid combining them to name the law. Instead, many Republicans (47, 23%) referred to the law by a novel construction, “the president’s health care law.” No Democrats used this term, despite the fact that they regularly constructed their statements from its component parts: president (25%); health (100%); care (100%); law (74%); again suggesting an intentional avoidance that belies their underlying position toward the bill.

A third way that authors can produce similar texts is through mimetic processes. In mimetic processes search and selection are short-circuited as authors use preassembled strategies they observe others using (DiMaggio & Powell, 1983). Authors may rely on mimetics because they lack the resources to independently research and construct their own statements (DiMaggio & Powell, 1983; Gandy, 1982) or because there are benefits to appearing to be in exact coordination with other authors (Benford & Snow, 2000; Lammers & Barbour, 2006). Although authors who imitate others are likely to share some constraints and attributes with those whom they mimic, the exactness with which they replicate others’ statements can be misleading. The rationale for imitation is the achievement of a satisfactory but suboptimal solution to a complex search or coordination problem at low cost (DiMaggio & Powell, 1983; March & Simon, 1993). Thus, authors may often imitate others with whom they only partially agree or share constraints with, producing discourse that is only loosely similar to what they would have said had they spent the time and resources to produce their own independent statement. The strength of similarity produced by mimetic processes is thus more of an indicator of resource constraints (Grimmer, 2009), an authority relationship between the authors (Danescu-Niculescu-Mizil et al., 2012), or a highly demanding and complex context.

For example, in response to the Supreme Court decision the phrase “[Health centers in home state have received SX million to] create new health center sites in medically underserved areas, enable health centers to increase the number of patients served, expand preventive and primary health care services” was used by five different Democratic members of Congress. The last 26 words of this phrase can also be found in a document produced by the White House promoting the benefits of the ACA.1 The high semantic similarity across these five individuals could be due to independent searches for and selection of high-quality arguments that independently arrived at identical 26-word constructions, but it appears more likely that this similarity indicates the absence of such processes, as these members were able to produce adequately credible statements that signaled loyalty to the White House with limited effort.

These processes may often be co-present in the construction of texts; however, their distinct antecedents allows them to vary independently. In a presidential debate two candidates may search within the same vocabulary to select phrases with opposite implications (Boydstun et al., 2013), while in local election campaigns candidates addressing different districts but from the same party may select similar

1http://www.whitehouse.gov/sites/default/files/docs/the_aca_helps_rural_america.pdf
constructions while addressing different topics of local interest. Within an organization, departments with excess resources may produce documents with which others only loosely agree with but nonetheless heavily imitate and reuse (Lammers & Barbour, 2006; Monge & Poole, 2008), while departments working with highly technical and disciplined material (e.g., engineering, accounting) may spend substantial resources to produce independently similar statements (Hart et al., 2013). We thus suggest it is both possible and necessary to model a tendency for authors to produce similar texts as a function of the strength of these different processes. Figure 1 presents the challenge we address: distinguishing the appropriate relationship between authors based on an observation that they produce similar texts.

**Empirical Distinction via N-Gram Length**

In this section we explore the possibility of distinguishing these processes by calculating semantic similarity between authors at different lengths of $n$-grams—contiguous sequences of $n$ words from a given sequence of text. The similarity between two documents can be described by the distribution of similarity at all $n$-gram lengths from 1 to $Nd$ where $Nd$ is the number of words in either document. Pairs of documents can show different distributions in their similarity over $n$-grams of different lengths. For example, one pair of documents might use many of the same words but few of the same longer constructions. At the other extreme, a pair of documents might be almost identical except for one or two edits or typos. Figure 2 shows a a space on which to plot the hypothetical similarity distribution of document pairs for the above scenario. In both extreme cases, the pair of documents share a large number of uni-grams. In the former case, the $n$-gram overlap declines precipitously as $n$ increases, whereas in the latter case it remains close to constant as $n$ increases over certain lengths.

The role of the semantic organizing processes described above is to pull the similarity between authors at a particular $n$-gram length “up,” creating a “bump” in the similarity distribution at that length. By examining semantic similarity at multiple $n$-gram lengths it may be possible to observe the operation of different processes, that is, the multimodal similarity distribution displayed by the red dashed curve.

Below we identify three specific semantic organizing processes, each of which represents one of our families of processes: authors sharing subject matter (search process), sharing rhetorical goals (selection process), and copying from shared sources (mimetic process). We hypothesize that each of these different processes will have its strongest implications at a particular level of $n$-gram length.

**Shared subject matter.** Some sets of documents are similar because their authors search among a common pool of ideas to construct their statements. We define such sets of similar documents as sharing *subject matter*. The key demands of subject matter are comprehension and competence. Given a set of ideas to refer to authors must use terms that are recognizable and combinations that make sense (Clark & Wilkes-Gibbs, 1986; Krauss & Fussell, 1991). These terms are most commonly found in the vocabulary of a subject, a preexisting set of terms and phrases with clear, stable definitions that refer to aspects of the subject matter (Loewenstein et al., 2012).

Although public figures often try “set agendas” by focusing on subject matter advantageous to them (McCombs & Shaw, 1972), audiences frequently demand that they address topics of public concern (Boydston et al., 2013; Jerit, 2008). There is evidence that audiences detect and reward those who use terms central to the discussion of these topics (Boydston et al., 2013; Doerfel & Connaughton, 2009). In these cases, inappropriate use of subject-related concepts will cast doubt on the competence or intentions of the author. We argue that vocabularies will tend to be comprised of short $n$-grams and, therefore, the majority of the material found by authors via searches in a shared vocabulary will be of the short $n$-gram variety. Studies of the creation of vocabulary terms indicates that they emerge in relation to ideas and features that frequently co-occur in the external world (Loewenstein et al., 2012; Rosch, 1978). For example, a bird is an animal with “wings” and a “beak”
Vocabulary terms have stable meanings with regard to these ideas and entities, and this need for frequency and stability is more easily achieved by short n-grams. Specifically, as a construction grows in n-gram length each new word added to it is likely to make it either more specific, reducing the likelihood that its referent occurs frequently in the external world, or more ambiguous, reduce the clarity of the consensus around its meaning (Aerts & Gabora, 2005; Carley & Kaufer, 1993; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Experts in scientific fields tend to have vocabulary for specific entities that they encounter frequently that lay individuals lack (Lynch, Coley, & Medin, 2000), while at a certain n-gram length, constructions move beyond noun phrases to include full clauses that have contextually contingent meaning (Corman et al., 2002).

Thus, we expect that when several authors intend to competently discuss a subject, they will tend to search within a common vocabulary and this vocabulary will be largely comprised of short n-grams. From this we derive:

**H1 (Shared subject matter):** To the extent to which two authors share subject matter, they will be more similar in their use of short n-grams than in their use of longer n-grams.

**Shared rhetorical goals.** Some sets of documents are similar because the author(s) of the documents would like their audience(s) to come to a particular conclusion. We describe such documents as sharing rhetoric, that is, strategic or persuasive goals.

When authors share rhetorical goals they will select the phrases and constructions or “optimal messages” (Danowski, 2011) that make their point and eschew those that undermine it (Chong & Druckman, 2007; Entman, 1993). These selections may be conscious, such as in the case of a heavily edited speech from a political campaign (Chapel, 1976), or unconscious, as when individuals find certain expressions offensive and others pleasing as a result of being socialized into a particular institutional view (Lammers & Barbour, 2006).

These constructions may rely on preexisting vocabulary for a subject, but clarity of communication and a demonstration of competence is not the goal. In fact, ambiguity in meaning can be an advantage when pursuing a rhetorical goal (Eisenberg, 1984). Instead, authors have incentives to influence their audience by manipulating the flow of the audience’s attention through the sequence of constructions they use (Carley & Kaufer, 1993; Danowski, 2011). For example, authors may find ways to deploy “slogans,” sequences of words that are relevant to many topics but which dull associations with additional ideas (Carley & Kaufer, 1993). By guiding the audience to think of a slogan, the audience then finds it hard to reason about the problem at hand. For example, on July 29, 2011, 78 Democrats and 123 Republicans released statements on the issue of whether Congress would raise the “debt ceiling” limit on government spending. On that day, 23 Republicans (19%) used the phrase “cut, cap, and balance” in promoting their alternative plan for the budget. This phrase, used by only two (3%) of Democrats, does not appear to refer to any concrete plan described in these statements.

The ability to emphasize particular aspects of an idea can also depend on the way individual words are juxtaposed (Lakoff, 2004). In their reference to the ACA as “the president’s health care law,” Republicans introduce a concept, “the president,” that likely activated negative associations in the mind of many of those who might be persuaded to agree with them about the law (Hardisty, Johnson, & Weber, 2010). By placing this term first in the sequence, it influences their interpretation of the combination of words as a whole (Wisniewski, 1996).

Because slogans and juxtapositions depend on sequence, but do not need to be widely recognized as stable ideas with agreed meaning, they can be deployed at substantially longer n-gram lengths than vocabulary terms. Furthermore, because n-gram sharing across authors becomes generically less likely as n-grams become longer, the fact that rhetorically motivated sequences can produce similarity in longer n-grams should make these shared textual sequences more apparent at longer n-gram lengths. For example, the phrase “I will continue to” is used by 33 (19%) of the Republicans who explicitly state they will attempt to “repeal” the ACA, and used by only two (5%) of the Republicans who do not make this commitment. Another n-gram, “and replace it with,” is also common among those advocating repeal. It turns out that more members say the 1-gram “repeal” than either of these longer phrases. However, these 4-gram sequences are more salient in relative terms. Each is in the top 10 of most shared 4-grams (5th and 8th, respectively) while “repeal” is only the 15th most shared 1-gram. “Repeal” is harder to detect because it is buried among other short n-grams, such as “health” and “care” that are chosen as subject matter words. In the analysis of longer n-grams, there are fewer vocabulary terms and the rhetorical constructions are more salient.

The length of the sequences for which an exact match should be expected to emerge from shared rhetorical goals is also limited, however. Not all differences in word sequence will make a material impact on interpretation. Two individuals can convey the same frame with similar, but nonidentical sequences where key, framing concepts are in the same sequence but less important placeholders and stop-words are not. Although 49 Republicans say “the president’s health care” and 47 say “president’s health care law,” only 16 say “of the president’s health care law.” At some point the rhetorical motivations are not strong enough to constrain discursive choices of such distinction. From this we derive:

**H2 (Shared rhetorical goals):** To the extent to which two authors share a rhetorical goal, they will be more similar in their use of moderate n-grams than in their use of short n-grams or long n-grams.

**Shared sources.** In mimetic processes, where authors copy one another or a shared third party, there is no constraint on n-gram length. However, copies may fit contexts less frequently than shorter stock vocabulary and framing
constructions, so their absolute frequency across texts may be likely to be limited when only short n-grams are taken into account. We thus anticipate that absolute sharing of very long n-grams will be low, but even small numbers of long shared n-grams will be evidence of sharing sources, as it is very difficult to explain the highly improbable convergence of authors around very long constructions in the absence of a common source (Barrón-Cedeño & Rosso, 2009).

**H3 (Shared texts):** To the extent to which two authors share a textual source, they will be more similar in their use of long n-grams than in their use of shorter n-grams.

**Method**

**Data**

We gathered 0.4 million documents from the public statements of members of the U.S. Congress from the Vote Smart Project website.² According to Vote Smart, these public statements include any press releases, statements, newspaper articles, interviews, blog entries, newsletters, legislative committee websites, campaign websites, and cable news show websites (“Meet the Press”, “This Week”, etc.) that contain direct quotes from the member. We include all of these in our data and focus on the statements made by the members of the 112th Senate and House, during the period between January 2011 through August 2012.

The individual attributes of the members of Congress, such as party and district, were retrieved using Sunlight Congress API.³ We use the DW-NOMINATE scores for the U.S. Congress (Carroll, Lewis, Lo, Poole, & Rosenthal, 2009) as measures of legislators’ ideological locations.⁴ The estimates are based on the history of roll call votes by the members of Congress and have been widely used in political science studies and related fields. Based on their method, each member’s ideological point is estimated along two dimensions. Previous research has shown that the first dimension reveals standard left–right or economic cleavages and the second dimension reflects social and sectional divisions.

**Randomized N-Gram Extraction**

**Computational challenge: n-gram explosion.** A naive method to observe authors using overlapping n-grams would be to scan all documents and count each author’s use of every n-gram that appears in the corpus. The problem with this approach is that the number of unique n-grams increases drastically as n increases.

Figure 3 shows the growth of n-grams in our data set. The total number of unique n-grams increases much faster for longer n-grams. On average, members of the U.S. Congress introduced 43 K new words, 6 M new trigrams, and 19.5 M 5-grams per year. The growth curves of n-grams are roughly linear, but curves of longer n-grams have larger slopes. Table 1 shows the number of total and unique n-grams in our data set for n ∈ {2, . . . , 7}. The total number of unique n-grams is roughly the same for different n’s, but the total number of unique n-grams for larger n’s is much higher. The fraction of

---

²http://www.votesmart.org
³http://services.sunlightlabs.com/docs/Sunlight_Congress_API/
⁴Estimates for the 112th Congress are available at: http://www.voteview.com/housevotes.asp
unique $n$-grams quickly approaches 1 when $n$ increases. Moreover, as shown in Table 2, the average fraction of unique 1-grams (words) in a document is around 50%, the fractions of unique 3-grams and 5-grams per document are around 95% and 98–99% (similar statistics appear in documents in other years). This indicates that the fraction of unique $n$-grams in a document quickly approaches 1 when $n$ increases, and is greater than 95% when $n > 3$. As a consequence, the total memory needed for counting the $n$-gram frequency approaches the length of corpus when $n$ increases, and most of the counters are used to store unique $n$-grams.

**Extraction technique.** We developed a randomized dictionary algorithm to tackle this computational issue. The idea is to iteratively sample (with replacement) a pair of documents from the corpus and only count $n$-grams that appear in both documents in each iteration. The algorithm will always find $n$-grams that appear in more than one document, and thus no memory is wasted storing unique $n$-grams. The performance improvement is significant—for example, in the case of 7-grams, the nonunique fraction is less than 18%.

The question remaining for using this randomized dictionary algorithm is how many samples are needed to find all sufficiently frequent $n$-grams. The number of samples can be derived as follows. Suppose our goal is to find all $n$-grams that appear in at least $p$ fraction of documents. Let $w$ be one of such $n$-grams. In our algorithm, the probability of missing $w$ in one iteration is $1 - p^2$. After $k$ iterations, the probability of missing $w$ is $(1 - p^2)^k$. Thus, the probability of finding $w$ in $k$ samples is $1 - (1 - p^2)^k$. Based on the inequality $1 - x \leq e^{-x}$, we have:

$$Pr[\text{find } w \text{ in } k \text{ samples}] = 1 - (1 - p^2)^k \leq 1 - e^{-kp^2}$$

By making $k$ larger, the probability of successfully finding $w$ can be made arbitrarily close to 1. Hence, the number of samples needed to find $w$ with success rate $r$ is

$$k > \left(\frac{1}{r^2}\right) \ln\left(\frac{1}{1 - r}\right)$$

The time complexity for sampling is $O(1/p^2)$, which means the sampling cost depends only on the document fraction of targeting $n$-grams, and is independent of the size of the corpus or the length of $n$-grams. For example, to find all $n$-grams that appear in 1% of total documents, approximately 20,000 samples are required in order to guarantee a 90% success rate.

We summarize the randomized dictionary algorithm in Table 3. For comparing documents from different authors, Step 3 can be further restricted as: randomly draw a document pair $(Da, Db)$ from $D$ with the constraint $\text{author}(Da) \neq \text{author}(Db)$. Step 8 is used to remove $n$-grams without sufficient frequency, but for the purpose of our study this step can be omitted.

The problem formulation falls within the setting of the frequency items problem—of which the goal is to find all items or sets of items whose frequency exceeds a specified fraction of the total amount of input data (Cormode & Hadjieleftheriou, 2009). In particular, our algorithm has a connection to the two-pass randomized sampling-based approach (Toivonen, 1996). The important difference between our method and the prior work is that we sample a document pair and add the candidate $n$-grams in each sampling iteration, rather than sample a document per sampling iteration and determine the candidates after the sampling pass. This is based on the key observation that the number of $n$-grams is enormous but most of them are unique. Our setting considers each $n$-gram as a single item, not a set of unigrams, which is also different from the frequent itemsets setting that usually breaks the relationship between items in the extracted itemsets.

**Validation.** To evaluate the performance of the randomized dictionary algorithm we use the naive method to extract all 3-grams and 5-grams in the corpus to create ground truth from which to measure the success rate of the algorithm. Then we evaluate the success rate of the randomized dictionary with respect to different targeting document fractions ($p$). Figure 4 shows the results of this computational experiment. The gray (solid and dashed) curves are theoretical lower bounds in terms of number of samples. The colored curves are experimental results where a point $(x, y)$ indicates success rates $y$ for $n$-grams with document fraction $p = x$ (i.e., those that appear in at least $x$ fraction of documents). The computational experiments are repeated 30 times for each $k$ and the means and standard deviations are marked with error bars. The figure shows that, for all $k = 200, \ldots, 200,000$ samples, the empirical success rates are bounded below by the

**TABLE 1.** Fraction of unique $n$-grams.

| $n$ | Total       | Unique    | Fraction |
|-----|-------------|-----------|----------|
| 1   | 236,071,504 | 463,101   | 0.002    |
| 2   | 235,704,167 | 12,349,893| 0.052    |
| 3   | 235,336,830 | 57,379,878| 0.244    |
| 4   | 234,969,494 | 116,653,788| 0.496   |
| 5   | 234,602,158 | 159,103,419| 0.678   |
| 6   | 234,234,830 | 181,644,125| 0.775   |
| 7   | 233,867,504 | 192,138,833| 0.822   |

**TABLE 2.** Average fraction of unique $n$-gram per document.

| Year | Max. doc size | Avg. doc size | Avg. 1-gram | Avg. 3-gram | Avg. 5-gram |
|------|---------------|---------------|-------------|-------------|-------------|
| 2008 | 11,512        | 680.841       | 0.505       | 0.959       | 0.989       |
| 2009 | 11,819        | 790.346       | 0.506       | 0.959       | 0.990       |
| 2010 | 11,538        | 560.536       | 0.530       | 0.963       | 0.990       |
corresponding theoretical curves (in gray). This indicates that the algorithm performs well, as expected.

**Author-Author Similarity Networks**

Our hypotheses are at the level of relationships between authors. Below we describe the author–author networks used in this study.

**N-gram similarity.** N-gram similarity was calculated for each pair of members of Congress for all documents in the data set that were not exact copies or “joint press releases” where multiple authors issued completely identical documents. Each member received an n-gram similarity score with each other member at each of the six lengths of n-gram for which matches were tracked. The number for this pair represents the extent to which these two members issued statements that tended to share n-grams of that length. This score was calculated as the total number of n-gram overlaps normalized by the number of statements made by the member. Specifically, for each author i we construct an n-gram vector $v_i$ where each entry $v_{ij}$ represents the number of occurrences of an n-gram $j$ in the documents authored by $i$. The normalized vector is computed by $w_i = \frac{v_i}{\sum_{j} v_{ij}}$, and at each n-gram length, the pairwise n-gram similarity is computed as the cosine similarity of the pair of normalized n-gram vectors. The results of these calculations are six directed $534 \times 534$ networks reflecting author similarity, one for each length of n-gram analyzed. The aggregated text similarities scores showed a highly skewed distribution, and thus scores were log-transformed to create a more appropriate distribution for a correlation analysis.

**Topic similarity.** Hypothesis 1 proposes that when authors share subject matter their statements will overlap more in short n-grams than in long n-grams. To test this hypothesis we use standard topic models (LDA; Blei et al., 2003) to assign topical distribution to a randomly sampled 20% of documents in our corpus. The topic model was based on bag-of-unigram features. For each document pair we compute similarity based on topic distribution. In the topic

![FIG. 4. Sampling success rates. The success rates (y-axis) over n-gram document frequencies (x-axis), by using different number of samples (k = 200, 2000, 20000, 200000). The gray curves are theoretical lower bounds for $1 - e^{-kx^2}$ different samples $k$. The colored lines are experimental results on success rates $y$ for (a) 3-grams and (b) 5-grams with document frequency $x$. The empirical success rates for samples are bounded below by the theoretical lower bound for $k$. The experiments were conducted 30 times for each $k$ and the means and standard deviations are reported. Error bars show the standard deviations. The horizontal dashed line indicates a 90% success rate can be guaranteed at all levels of document frequency. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
modeling, the number of topics $K$ can be determined by predictive likelihood (i.e., perplexity; Blei et al., 2003) based on cross-validation. We explored the topics with $K = 50$, 100, and 150 and found $K = 100$ has the highest perplexity. Note that in our experiment we are not interested in finding the exact optimal number of topics because the topics are used to build a length-$K$ vector for the posterior topic proportions of each document. The topical similarity is defined as the cosine similarity between the topic proportion vectors of the pairs of documents. These similarities of document pairs are then converted into edge weights in the author-to-author networks based on the document authorship.

**DW-Nominate similarity.** Hypothesis 2 stated that when the authors of documents shared a rhetorical goal their documents would share moderate length $n$-grams. It is difficult to obtain ground truth for an author’s intentions on a particular document. Nonetheless, it is possible to measure the overall rhetorical goals of U.S. Congress members through the policy preferences expressed in their votes as captured by the first dimension of their DW-Nominate scores (Carroll et al., 2009). That is, the more similar two members are in their voting record, the more similar we assume are the persuasive goals of their discourse. The maximum score for DW-Nominate is 1, indicating most conservative, and the minimum score is −1, indicating most liberal. For each pair of members the difference of their DW-Nominate scores was subtracted from the maximum difference (2) to transform it into a similarity score. The DW-Nominate similarity network is thus a $534 \times 534$ undirected network where scores range from 2 (most similar) to 0 (least similar).

**Chamber similarity.** Hypothesis 3 stated that when the authors rely on common sources from which they may copy text this similarity will be most visible long $n$-gram lengths. To test Hypothesis 3 we require a relation that does not map directly onto ideology, such that individuals could share ideology (but be less likely to share the source) or share the source (but be less likely to share ideology). We elected to use the chamber structure of the Congress. Both the House and the Senate contain a range of member DW-Nominate scores. There is also an existing mechanism through which members from the same chamber can share sources. A common way for members to solicit sponsors for legislation they propose is through “Dear Colleague” letters (Krust, 2005; Petersen, 2005), in which they write several paragraphs justifying why their bill should become law. Although these letters can be circulated to all members, they are often sent only within one chamber (Petersen, 2005), and, since most bills never make it to the next chamber for consideration, most sponsors are gathered from within a member’s own chamber only (Krust, 2005).

**Analysis**

Correlations between $n$-gram networks and other author–author networks were tested using the quadratic assignment procedure (QAP) (Krackhardt, 1987). The QAP uses the Pearson correlation coefficient between the dyadic scores in each network but calculates the significance based on permutations to control for network interdependency. Since both DW-Nominate similarity and chamber membership are compared to the same networks using the same technique, their partial correlation with $n$-gram networks was also tested using network regression based on the QAP (Dekker, Krackhardt, & Snijders, 2007). To assess the significance of the differences in the $n$-gram correlations, we generate bootstrap samples with 1,000 replicates from each of the $n$-gram networks and calculate the pairwise Tukey’s Honest Significant Differences test.

![Graph of n-gram length distribution](image1.png)

![Graph of n-gram frequency distribution](image2.png)

**FIG. 5.** $N$-grams found by randomized dictionary. The $n$-grams found by randomized dictionary exhibit heavy-tailed length and frequency distributions. (A) The length distribution indicates the existence of very long $n$-grams. (B) The frequency distribution for $n$-grams of different lengths. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
Results

Descriptive Analysis

Figure 5 shows the length and frequency distributions of the n-grams found by a randomized dictionary algorithm. The length distribution is heavy-tailed, indicating the existence of very long n-grams. The frequency distribution for n-grams of different lengths is also heavy-tailed. This means there are a few n-grams occurring very frequently, and many n-grams occurring rarely. In uni-grams, the most frequent ones such as “and” and “the” are often referred to as “stop-words” and removed before more sophisticated text analysis as they do not carry useful semantic information. Figure 5 indicates that, even in the set of very long n-grams, there exist “stop-n-grams” that occur extremely frequently compared to the rest of the set. Inspection of these “stop-n-grams” showed that they tend to be procedural language used within the congressional forum, such as “I yield back the remainder of my time,” that appear in a very large number of documents. Like stopwords, these stop-n-grams need to be removed before further analysis. We describe the procedure to filter such stop-n-grams in the Appendix.

Table 4 shows additional examples of sampled short (bi-grams), moderate (4-grams) and long n-grams (for n \( \geq 16 \)) in three broad categories: health, taxes, and jobs.

To examine the evidence for the theorized “bumps” in the distribution of observed n-gram similarity we randomly sample 20% of the documents from the corpus and then compute Jaccard similarity of the n-gram sets for n \( \in [1, 2^{16}] \) in each pair of documents (excluding those authored by the same individual). Figure 5 shows the distribution of similarities by n-gram length. A multimodal distribution appears in both (a) similarity calculated based on the Jaccard similarity measure of the n-gram sets, and (b) number of matched n-grams in the pairs of documents by distinct authors. Consistent with expectations, we observe three distinct regions of similarity, one for short n-grams, one for moderate n-grams, and one for long n-grams. This provides evidence that there are distinct similarity generating processes happening at different n-gram lengths and suggests that any analysis that looks at only one length of n-gram ignores important information.

Hypothesis Tests

Hypothesis 1 stated that authors who shared subject matter would share short n-grams but not long n-grams. Figure 7 and Table 5 show the first-order dyadic correlations between the author–author similarity at different n-gram lengths and Topic similarity, DW-Nominate similarity and Chamber similarity, respectively. The three types of similarities over different n-gram lengths are shown as three curves in Figure 7. Table 6a–c shows the results of our statistical tests for different lengths of n-gram correlations with Topic similarity, DW-Nominate, and Same Chamber, respectively. Each table corresponds to a curve shown in Figure 7. In each of the three tables (a–c), each row indicates which pair of correlations is being compared. For example, the first row in Table 6a is the (QAP) correlation of (3-gram, topic) minus the correlation of (2-gram, topic), and the columns show the difference in the observed means of the

| Length | N-gram examples |
|--------|-----------------|
| Health | Health care (10,334); health insurance (2,933); public health (1,934); health service (1,054); affordable health (881); mental health (868); health reform (750); health benefit (696); women’s health (597) |
| 4 | Health and human services (1,796); the health care law (1,169); our health care system (668); access to health care (620); the president’s health care (587) |
| 16 | Who have health insurance through their employer or the market for private insurance eliminating health care tax credits for up to (11); bans insurance companies from imposing lifetime dollar limits on health benefits freeing cancer patients and individuals suffering from other chronic (10); on health benefits freeing cancer patients and individuals suffering from other chronic diseases from having to worry about going without (9); health care coverage to 47 million Americans including 39 million seniors and 8 million people under 65 (9); to community health centers doctors and other healthcare providers in the district to improve the community’s health (9) |
| 2 | Tax cut (3,208); taxpayer dollar (2,757); tax increase (2,381); american tax (1,994); tax reform (1,611); income tax (1,224); tax relief (1,052); tax hike (1,014); corporate tax (802); tax loophole (782) |
| 4 | The payroll tax cut (783); the bush tax cuts (448); to raise taxes on (297); the middle class tax (285); the payroll tax holiday (257) |
| 16 | Tax credit of up to $5 600 for hiring veterans who have been looking for a job for more than six months (20); tax credit of up to $9 600 for hiring veterans with service-connected disabilities who have been looking for a job for more than six (17); than 500 employees to take a tax deduction equal to 20% of their active business income (13); included in the payroll tax cut extension bill passed in december didn’t give him enough time to review the project in fact (12); expands much needed lending to millions of small businesses and offers tax incentives to help small businesses grow hire and fuel (12) |
| Job | Create job (8,554); american job (4,063); job creator (3,761); new job (3,222); job growth (2,830); their jobs (1,851); more job (1,802); jobs bill (1,726); million job (1,679); jobs act (1,608) |
| 4 | To create jobs and (1,430); of thousands of jobs (757); economic growth and job (668); jobs here at home (417); that will create jobs (385) |
| 16 | $5,600 for hiring veterans who have been looking for a job for more than six months (28); $9,600 for hiring veterans with service-connected disabilities who have been looking for a job for more than (24); to begin the federal employment process prior to separation in order to facilitate a truly seamless transition from the military to jobs (23); been looking for a job for more than six months as well as a $2 400 credit (23) |
pairwise correlations, the lower and upper end points of the interval, and the \( p \)-value after adjustment for the multiple comparisons. The difference is significant if the \( p \)-value is small. The results indicate that (a) the highest correlation between topic and \( n \)-gram similarities appears at \( n = 2 \), and (b) the correlations drop gradually at \( n = 3, 4, 8 \) (no significant difference after \( n > 8 \)). This is consistent with Hypothesis 1. Authors who share topics are similar in shorter \( n \)-grams but are not particularly likely to be similar in longer \( n \)-grams.

Hypothesis 2 stated that when the authors shared a rhetorical goal, their documents would be similar even at short \( n \)-gram lengths. Table 5 shows the results of the QAP tests. The relationship between shared rhetorical goals and semantic similarity is statistically significant at each length of \( n \)-gram. Table 6b shows the highest correlation between DW and \( n \)-gram similarities appears at \( n = 8 \), and gradually tapers off on both ends.

Hypothesis 3 stated that when the authors construct their documents from common sources, their documents can be similar even at long \( n \)-gram lengths. Table 5 shows the results of the QAP tests. The relationship between working in the same chamber and text similarity is statistically significant at each length of \( n \)-gram. However, at short and moderate \( n \)-gram lengths (\( n \leq 8 \)) the correlation is negative. At \( n = 16 \) and \( n = 32 \), the correlation is positive. The highest correlation between same-chamber and \( n \)-gram similarities appears at \( n = 16 \) and gradually tapers off on both ends. In other words, the test results provide statistical significance for the peaks in each of the three curves in Figure 7.

To compare the relative association between DW-Nominate similarity and chamber membership and \( n \)-gram similarity at different levels of \( n \), we run a network regression that yields the partial correlation coefficient for each of these networks relative to the \( n \)-gram similarity network (Table 7). The results are consistent with the first-order correlation analyses. The DW-similarity coefficient is significant and positive in each regression. The shared chamber membership is negative for \( n = 2 \) through 8, and is not significant for \( n = 3 \) or \( n = 8 \). At \( n = 16 \), however, the coefficient is positive and significant. At \( n = 32 \), the coefficient is positive, significant, and larger than the coefficient for DW-nominate similarity (DW-nominate similarity is on a 0–2 scale, but \( 2 \times \) this coefficient is less than the coefficient for shared chamber at \( n = 32 \)). In other words, at the long \( n \)-gram lengths of 16 and 32, the mimetic processes begins to compete with and then overtake shared rhetorical goals as the process producing similarity.

![Graph](image-url)
In Figure 8, we visualize the congressional social networks based on the n-gram similarity of different lengths. In the networks, each node represents a member of Congress and each edge represents n-gram similarity based on length n. For ease of visual comparison, all three networks are constructed to have approximately the same number of edges (no edges are removed in the statistical analyses). Specifically, we keep removing low-similarity edges in each network until the final edge density reaches 0.1 (this threshold is empirically determined in order to balance between data filtering and visual clarity). The figure shows that ideological differences are apparent at moderate n-gram similarity but the chamber structure of the Congress begins to emerge at longer n-gram similarity.

Discussion

Applied and Theoretical Implications

We provide evidence that different shared processes of authorship, reflecting different kinds of social relationships between authors, can be detected by examining authors’ semantic similarity at different n-gram lengths. We first show that although most existing approaches assume only one authorship process and assume that they have chosen an appropriate n-gram length, there is theory (Figure 2) and evidence (Figure 6) that the full spectrum of text similarities contains systematic information about authors and their relationships. We also show that similarities at specific n-gram lengths have detectable correspondence to specific nontextual attributes and contexts such as voting ideology and the social context (chamber) in which authors work.

These findings suggest a number of applications for designing information systems to support social network analysis and related social science investigations (Karanasios et al., 2013). One application of particular importance is in the parsing of strategic language use in text. Strategic (rhetorical) language is largely ignored by, and therefore confounds, many existing techniques. Our analysis suggests that n-gram-based similarity measures can be used to identify and potentially “control out” the influence of this process, returning discourse that meets the assumptions of other techniques.

Distinguishing subject sharing and goal sharing is particularly important to studies in media and politics (e.g., Glazier & Boydstun, 2012) and organizational change (e.g., Rice, 2005) where changes in semantic similarity are used to measure changes in consensus. The results of our study suggest that consensus in texts can be both conceptualized and measured in two distinct ways: as consensus about what to talk about versus consensus about how to think about what is being discussed. These might be measured independently and tracked over time.

TABLE 5. First-order correlation (QAP) with n-gram networks. (*p < .01, **p < .01, ***p < .001, all significance estimates calculated using QAP.)

| N-gram length | 2      | 3      | 4      | 8      | 16     | 32     |
|---------------|--------|--------|--------|--------|--------|--------|
| Topic         | 0.27***| .05*** | 0.005  | 0.000  | 0.000  | 0.000  |
| DW            | 0.17***| 0.22***| 0.22***| 0.21***| 0.17***| 0.06***|
| Chamber       | −0.13***| −0.12***| −0.09***| −0.05***| 0.21***| 0.1*** |

FIG. 8. Congressional social networks constructed based on co-mentioning (a) 2-grams, (b) 4-grams, and (c) 32-grams in their public statements. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
TABLE 6. Pairwise comparisons using Tukey’s Honest Significant Differences test.

(a) Topic

|   | Diff  | Lower | Upper | Adj p-value |
|---|-------|-------|-------|-------------|
| 3-2 | -0.281 | -0.283 | -0.280 | 0.000 |
| 4-2 | -0.349 | -0.350 | -0.348 | 0.000 |
| 8-2 | -0.358 | -0.359 | -0.356 | 0.000 |
| 16-2 | -0.358 | -0.359 | -0.356 | 0.000 |
| 32-2 | -0.358 | -0.359 | -0.356 | 0.000 |
| 4-3 | -0.068 | -0.069 | -0.066 | 0.000 |
| 8-3 | -0.076 | -0.078 | -0.075 | 0.000 |
| 16-3 | -0.076 | -0.078 | -0.075 | 0.000 |
| 32-3 | -0.076 | -0.077 | -0.075 | 0.000 |
| 8-4 | -0.009 | -0.010 | -0.007 | 0.000 |
| 16-4 | -0.009 | -0.010 | -0.007 | 0.000 |
| 32-4 | -0.009 | -0.010 | -0.007 | 0.000 |
| 16-8 | 0.000 | -0.001 | 0.001 | 1.000 |
| 32-8 | 0.000 | -0.001 | 0.001 | 1.000 |
| 32-16 | 0.000 | -0.001 | 0.001 | 1.000 |

(b) DW

|   | Diff  | Lower | Upper | Adj p-value |
|---|-------|-------|-------|-------------|
| 3-2 | 0.033 | 0.028 | 0.038 | 0.000 |
| 4-2 | 0.043 | 0.038 | 0.048 | 0.000 |
| 8-2 | 0.048 | 0.043 | 0.053 | 0.000 |
| 16-2 | -0.014 | -0.019 | -0.009 | 0.000 |
| 32-2 | -0.098 | -0.103 | -0.093 | 0.000 |
| 4-3 | 0.010 | 0.005 | 0.014 | 0.000 |
| 8-3 | 0.015 | 0.010 | 0.020 | 0.000 |
| 16-3 | -0.047 | -0.052 | -0.042 | 0.000 |
| 32-3 | -0.131 | -0.136 | -0.126 | 0.000 |
| 8-4 | 0.005 | 0.000 | 0.010 | 0.030 |
| 16-4 | -0.057 | -0.062 | -0.052 | 0.000 |
| 32-4 | -0.141 | -0.146 | -0.136 | 0.000 |
| 16-8 | -0.062 | -0.067 | -0.057 | 0.000 |
| 32-8 | -0.146 | -0.151 | -0.141 | 0.000 |
| 32-16 | -0.084 | -0.089 | -0.079 | 0.000 |

(c) Same chamber

|   | Diff  | Lower | Upper | Adj p-value |
|---|-------|-------|-------|-------------|
| 3-2 | 0.008 | 0.002 | 0.013 | 0.002 |
| 4-2 | 0.043 | 0.037 | 0.048 | 0.000 |
| 8-2 | 0.088 | 0.082 | 0.093 | 0.000 |
| 16-2 | 0.328 | 0.322 | 0.333 | 0.000 |
| 32-2 | 0.261 | 0.256 | 0.267 | 0.000 |
| 4-3 | 0.035 | 0.030 | 0.041 | 0.000 |
| 8-3 | 0.080 | 0.075 | 0.086 | 0.000 |
| 16-3 | 0.320 | 0.315 | 0.326 | 0.000 |
| 32-3 | 0.254 | 0.248 | 0.259 | 0.000 |
| 8-4 | 0.045 | 0.040 | 0.051 | 0.000 |
| 16-4 | 0.285 | 0.279 | 0.290 | 0.000 |
| 32-4 | 0.219 | 0.213 | 0.224 | 0.000 |
| 16-8 | 0.240 | 0.233 | 0.245 | 0.000 |
| 32-8 | 0.174 | 0.168 | 0.179 | 0.000 |
| 32-16 | -0.066 | -0.072 | -0.061 | 0.000 |

Our emphasis on and evidence for the importance of distinguishing mimetic processes in such systems is also important. To date, work on mimetic processes has largely taken place in a parallel but separate track from other studies of textual similarity. It may be possible to use n-gram similarity to answer basic questions about mimetic processes such as whether individuals tend to copy others who use similar, but not copied, words and phrases.

More broadly, our results suggest new lines of theoretical and methodological investigation. Theoretically, more attention is needed to distinguish different authorship processes. For example, shared subject matter and shared goals appear to be a process of structural equivalence, in which authors independently choose similar constructions, whereas shared sources appear to be a relationship of social influence. Our study also suggests the value of testing generic text analysis techniques on corpuses produced by authors with known attributes. Our technique focuses on features—shared n-grams—that can be calculated across any set of documents without knowledge of the native language. Rather than relying on dictionaries or the interpretation of coders, we test our hypotheses through the use of other author behavioral data. Future work may consider adapting this approach to other settings where texts are difficult to interpret and social relationships of interest are difficult to observe, such as discovering terrorist networks (Danowski, 2011).

Limitations

Our analysis is a first attempt to distinguish disparate social relationships from text archives. In general, the ground truth for connecting texts with possible social processes is difficult to obtain. In our case, because our analysis relies on the known attributes of political actors, there might be concerns about generalizability to nonpolitical contexts. One way to address this is to test these hypotheses on corpuses from other domains. Sports reporting in local newspapers may provide a good case, as sports writing describes events for which there is a well-known vocabulary but does so from a particular point of view. Sportswriters also share sources in the form of press conference interviews. Essays written by undergraduates for different assignments and under different group conditions might also be a useful place to run tests.

We also do not intend to suggest that text similarity as a function of n-gram is the only dimension for distinguishing social processes. Further investigation of possible dimensions may prove fruitful for areas including document retrieval, natural language processing, author attribution, and social network analysis.

Conclusion and Future Work

In this paper we present an integrated approach for analyzing text similarity in which multiple processes are considered. Drawing on communication theory we identify three broad processes by which two documents might be detectably similar in their textual features based on the length of the n-grams in which document texts overlap. We employ a novel technique for detecting n-gram overlaps of

---

4We are grateful to the anonymous reviewer for this suggestion.
TABLE 7. Partial correlation (estimated using random sample of 100 members; significance level: \(*p < .01, **p < .01, ***p < .001\)).

| n-gram length | 2    | 3    | 4    | 8    | 16   | 32   |
|---------------|------|------|------|------|------|------|
| DW            | 0.555*** | 0.864*** | 0.859*** | 1.306*** | 2.668*** | 0.359 |
| Same chamber  | 0.428*** | 0.122  | 0.679*** | 0.198  | 3.145*** | 0.010*** |
| R-squared     | 0.057  | 0.057  | 0.062  | 0.068  | 0.073  | 0.06  |
| Coefficient ratio (Abs) | 1.295  | 7.058  | 1.264  | 6.580  | 0.848  | 0.445  |

arbitrary length to show that there exist distinct text signatures reflecting the operation of three different social processes within a corpus where information about the authors’ subject interests, goals, and exposure to one another is available. This work takes the first step in computationally distinguishing textual similarity due to known theoretical processes and has important implications for modeling text and underlying social processes. Our future work includes probing the generalizability of our hypotheses in additional corporuses and investigating different features that may also distinguish social processes.

Acknowledgments

The authors gratefully acknowledge the support of MURI grant #504026, DTRA grant #509475, and ARO #504033. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors alone.

References

Adamic, L.A., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. ACM.

Adamic, L.A., Lento, T.M., Adar, E., & Ng, P.C. (2014). Information evolution in social networks. arXiv Preprint arXiv:1402.6792.

Aerts, D., & Gabora, L. (2005). A theory of concepts and their combinations. Journal of Learning Analytics, 3(1/2), 151–175.

Atteveoldt, W.V., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing semantic networks and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. Political Analysis, 16(4), 428–446.

Barnett, G.A., Wigand, R.T., Harrison, R.P., Woelfel, J., & Cohen, A.A. (1986). Framing theory. Annual Review of Political Science, 10(1), 103–126.

Barrón-Cedeño, A., & Rosso, P. (2009). On automatic plagiarism detection. Journal of Organizational Behavior, 31(1), 533–558.

Bartels, M., Deeg, M., & Kats, J. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.

Boydstun, A.E., Glazier, R.A., & Pietyka, M.T. (2013). Playing to the crowd: Agenda control in presidential debates. Political Communication, 30(2), 254–277. doi: 10.1080/10584609.2012.737423.

Bramsen, P., Escobar-Molano, M., Patel, A., & Alonso, R. (2011). Extracting social power relationships from natural language. Proceedings of the 2nd International Conference on Information and Communication Technology (ICT), 773–782.

Bryden, J., Funk, S., & Jansen, V.A. (2013). Word usage mirrors community structure in the online social network Twitter. EPJ Data Science, 2(1), 1–9.

Carley, K., & Kaufer, D. (1993). Semantic connectivity: An approach for analyzing symbols in semantic networks. Communication Theory, 3, 183–213.

Carley, K.M. (1997). Extracting team mental models through textual analysis. Journal of Organizational Behavior, 18, 533–558.

Carroll, R., Lewis, J., Lo, J., Poole, K., & Rosenthal, H. (2009). Measuring bias and uncertainty in DW-nominate ideal point estimates via the parametric bootstrap. Political Analysis, 17(3), 261–275.

Chapel, G. (1976). Speechwriting in the Nixon administration. Journal of Communication, 26(2), 65–72.

Chong, D., & Druckman, J.N. (2007). Framing theory. Annual Review of Political Science, 10(1), 103–126.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1–39.

Corman, S.R., Kuhn, T., McPhee, R.D., & Dooley, K.J. (2012). Studying complex discursive systems. Human Communication Research, 28(2), 157–206.

Cormode, G., & Hadjieleftheriou, M. (2009). Finding the frequent items in streams of data. Communications of the ACM, 52(10), 97–105.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In Proceedings of the 21st International Conference on World Wide Web (pp. 699–708).

Danowski, J.A. (2001). Computer-mediated communication: A network-based content analysis using a CBBS conference. In M. Burgoon (Ed.), Communication Yearbook (Vol. 6, pp. 874–904). Beverly Hills, CA: Sage.

Danowski, J.A. (2011). Counterterrorism mining for individuals semantically-similar to watchlist members. In U.K. Wiil (Ed.), Counterterrorism and open source intelligence (Vol. 2, pp. 223–247). New York: Spring-Verlag/Wien.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

DiMaggio, P.J., & Powell, W.W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review, 48, 147–160.

Doerfler, M.L., & Barnett, G.A. (1999). A semantic network analysis of the 1992 US election: Election year winners and losers in U.S. televised presidential debates, 1960–2004. Journal of the American Society for Information Science and Technology, 60(2), 201–218.

Eisenberg, E.M. (1984). Ambiguity as strategy in organizational communication. Communication Monographs, 51(3), 227–242.
Appendix: Stop $N$-grams

FIG. A1. $N$-gram active phase. The figure shows six examples of $n$-grams with active phase (A,C,E) and without active phase (B,D,F). Gray, blue, and red lines indicate number of documents mentioning the $n$-gram per days, the smoothed numbers, and the mean of the smoothed curves, respectively. The active fraction is computed based on the portion of above-mean in the smoothed curves. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
The idea is that such stop-n-grams are used for reasons independent of the document context, and hence can be differentiated by how they are used in the corpus over time. Figure A1 shows the time series of six example n-grams. For each n-gram, we show the number of documents mentioning the n-gram over days in a gray curve. We can see the n-grams such as “the united states” and “my colleagues on the other side of the aisle” are used almost all the time and are not necessarily determined by the context of the documents, while the n-grams such as “the debt ceiling” and “the full faith and credit of the united states” are used in relating to particular political issues. Note that the 3-gram “the united states” occurs frequently but the 9-gram “the full faith and credit of the united states” does not.

We differentiate those stop-n-grams by computing the active fraction of each n-gram. The active fraction $\alpha$ is defined as the above-mean ratio in the smoothed n-gram mentioning time series. Here we empirically chose a 21-day sliding window to smooth the time series. As can be seen in Figure A1, those with larger active fractions are less relevant to specific document context.